ICSTI | STIC

International Council for Scientific and Technical Information Annual Conference

2011 ICSTI Public Conference

June 7-8, 2011 Beijing, China

http://www.icsti2011.org

Scientific Research Publishing, USA
美国科研出版社
2011

ISBN: 978-1-935068-76-1

Printed copies of this publication are available from:

Scientific Research Publishing, Inc.
5005 Paseo Segovia, Irvine,
CA 92603-3334, USA.
E-mail: service@scirp.org

Produced by Scientific Research Publishing

For information on producing a conference proceedings and receiving an estimate, contact service@scirp.org
http://www.scirp.org

Preface

Welcome to the International Council for Scientific and Technical Information Annual Conference 2011, also known as "ICSTI'2011." ICSTI'2011 is organized by the International Council for Scientific and Technical Information (ICSTI) and hosted by the Institute of Scientific & Technical Information of China (ISTIC).

ICSTI is a broad-based, international, not-for-profit membership organization. It was established by International Council of Scientific Unions (ICSU), consisting of the national scientific and technical information agencies, libraries and information resource management agencies, publishers, technology companies, and other entities that are related to ICSTI's mission. ICSTI offers a unique forum for interaction among organizations that create, disseminate and use scientific and technical information. ICSTI's mission cuts across scientific and technical disciplines, as well as international borders, to give member organizations the benefit of a truly global community. ICSTI fosters cooperation among all stakeholders engaged in the scientific communication process with the aim of improving the effectiveness of scientific research. It fully exploits its unique position at the intersection of scientific and technical knowledge creation, organization, dissemination and use, to identify and act upon key challenges, without a politicized or commercially driven agenda.

ISTIC was created in 1956 as a national scientific and technical information service institute under the auspices of the Ministry of Science and Technology of China. Its areas of activity include: the collection and provision of scientific and technical documents; S&T policy research based on factual databases; theoretical and methodological research on information science; sharing and management of scientific and technical information; development and application of knowledge technology and language technology; construction of large S&T information service platforms; and the training and further education of information professionals. In addition, ISTIC is responsible for the maintenance and operation of the National Engineering and Technology Library and the National Centre for Integrated Utilization and Public Services of Science and Technology Information Resources.

We are witnessing a global trend transforming sci-tech information-based services to knowledge-based services. Therefore, under the new digital environment, one of the most important challenges is to provide high-quality knowledge services to meet the knowledge required by a broad spectrum of entities to innovate. ICSTI'2011 provides a unique forum for engagement around the conference theme "Upgrading Information to Knowledge," as well as other common interests of the participants, who may have different frames of reference in which to deploy what they learn at the conference.

There are myriad ways to transform information into knowledge. This year we are pleased that ICSTI'2011 has attracted a number of international professionals from libraries, information agencies, academia, and industry who come to share their experiences, benchmark, learn best practices, and find out about emerging trends. We have selected 76 papers from all the submissions and have classified the conference research topics into three general tracks or areas of research. Each track contains several topics as follows.

Development of Information Resources and Services

This area concerns the following topics:

- Construction and management of information resources, including management and construction of

technical information, policy and theoretical methods of information resources construction in the digital age.

- Integration and sharing of information resources, including theory and appreciation method of information integration and sharing, access of science and technology information resources, and the construction of sharing system.
- Information service and development, including progress and practice of information service and user research, digital information service in a networked environment, the evolution and development of knowledge services, business information service, and the development of specialized application to meet user specifications.

Knowledge Organization and Knowledge Discovery

The area is concerned with:

- All aspects of identifying, acquiring, modeling, and managing knowledge, and its role in the construction of knowledge-intensive systems, services and products for the digital library, semantic web, knowledge management, e-government, natural language processing, intelligent information integration, etc.
- Association rules, classification, clustering, parallel/ distributed data mining, KDD Process, and human interaction, knowledge visualization, high dimensional data, scientific databases, semi-structured/unstructured data, web and the internet, etc.

Engaging with the Information Environment

This area concerns the following aspects:

- Digital justice and social development in a networked environment, including digital equity, the digital divide, information literacy, etc.
- Interaction of information services, including the relationship between people and technology in a digital society.
- Issues on intellectual property in a digital environment, including intellectual property and information services, access, method of information service, and delivery.
- Information assurance, authenticity, and accuracy.

The members of ICSTI'2011 Program Committee and Organizing Committee have worked hard to organize peer review over submitted papers and to provide feedback to the authors. All submissions have been reviewed by at least two reviewers. The members of the International Advisory Committee, not only participated in the review process, but also provided other valuable input during the conference planning. We would like to thank all of them for their effort, time and expertise. We also express our gratitude to all the authors for their contributions.

General Co-Chairs:
Roberta I. SHAFFER, President of ICSTI
Defang HE, President of ISTIC
June, 2011

International Advisory Committee

Co-Chairs: Roberta I. Shaffer, President of ICSTI
Defang He, President of ISTIC

Members:

Pam Bjornson, Director General of the CISTI-NRC Canada

Jan Brase, German National Library of Science and Technology, Vice President of ICSTI

Chaomei Chen, Associate Professor, Drexel University, Philadelphia, PA, USA

Jiachang Chen, Vice President of ISTIC, China

Bernard Dumouchel, Special Advisor of ICSTI

Huiling Feng, Vice President of Renmin University of China

Eleanor Frierson, Deputy Director of National Agricultural Library, USA

Brain Hitson, OSTI-DOE, USA, Chairman of the ICSTI TACC

Guohua Jiang, Member of Advisory Committee of Experts for Beijing Jili University, Standing Vice Director of Education Institute, China

Ping Ke, Director of Department of Information Management, Nankai University, China

Maosheng Lai, Vice Director of Department of Information Management, Peking University, China

Loet Leydesdorff, Professor, University of Amsterdam, Holland

Zhanping Liang, Research Fellow, ISTIC, China

Tony Llewellyn, Executive Director of ICSTI

Feicheng Ma, Director of Information Resources Research Center, Wuhan University, China

Jie Peng, Director of Resources Sharing & Promotion Center, ISTIC, China

Kirsi Tuominen, Head of Knowledge Solutions, VTT Technical Research Centre of Finland

Yuguang Wang, Professor of Department of Information Management, Beijing University, China

Zhongtuo Wang, Academician, Director of Center for Knowledge Science and Technology, Dalian University of Technology, China

Walter Warnick, Director of the U.S. Department of Energy Office of Scientific and Technical Information (OSTI)

Yishan Wu, Chief Engineer of ISTIC, China

Jiyuan Ye, Professor of Nanjing University, China

Haibo Yuan, Director of NSTL, China

Xiaolin Zhang, Deputy Director of Digital Library Professional Committee of China Society for Library Science, China

Zhiyun Zhao, Vice President of ISTIC, China

Qinghua Zhu, Professor of Department of Information Management, Nanjing University, China

Program Committee

Co-Chairs: Tony Llewellyn, Executive Director of ICSTI
Jie Peng, Director of Resources Sharing & Promotion Center, ISTIC, China

Members:

Brian Hitson, Chairman of the ICSTI TACC

Roberta I. Shaffer, President of ICSTI

Organizing Committee

Co-Chairs: Defang He, President of ISTIC
Tony Llewellyn, Executive Director of ICSTI

Members:

Jiachang Chen, Vice President of ISTIC, China

Linhao Chen, Deputy Director-General, Depart of International Cooperation, MOST, China

Yu Chen, Director of Economic Science Laboratory of Renmin University of China

Maitre-Allain Elisabeth, Executive Secretary of ICSTI

Brain Hitson, Chairman of the ICSTI TACC

Yongqing Jiang, Vice President of ISTIC, China

Maosheng Lai, Vice Director of Department of Information Management, Peking University, China

Ping Li, Director-General, Department of Personnel, MOST, China

Tony Llewellyn, Executive Director of ICSTI

Weiwei Lv, Vice President of ISTIC, China

Feicheng Ma, Director of Information Resources Research Center, Wuhan University, China

Roberta I. Shaffer, President of ICSTI

Guchao Shen, Professor of Department of Information Management, Nanjing University, China

Zhijin Wang, Professor of Department of Information Management, Nankai University, China

Xueti Wu, Deputy Director-General, Department of Facilities and Financial Support, MOST, China

Yishan Wu, Chief Engineer of ISTIC, China

Xianli Xing, Secretary of the CPC, ISTIC, China

Haibo Yuan, Director of NSTL, China

Mannian Zhang, Vice President of ISTIC, China

Xiaolin Zhang, Deputy Director of Digital Library Professional Committee of China Society for Library Science, China

Zhiyun Zhao, Vice President of ISTIC, China

Contents

Session 1: Development of Information Resources and Services

Session 2: Knowledge Organization and Knowledge Discovery

Session 3: Engaging with the Information Environment

目　录

分论坛 1　信息资源建设与服务

分论坛 2　知识组织与知识发现

分论坛 3　信息环境构建

Session 1

Development of Information Resources and Services

分论坛 1

信息资源建设与服务

Scientific Research

Knowledge-Base of Engineering Design and Its Application in Design Archives Management System

Tao CHEN, Zhanli FENG, Dunru REN, Liqiu CHANG, Qi YUAN

Design and Research Institute of Anshan Iron and Steel Group Corporation

Anshan City, Liaoning Province, China

Email: qq5547661@yahoo.cn

Abstract: In expert systems, knowledge-base is the core of the design. This paper discusses the architecture, the technology characteristic and the design methods of Knowledge-Base of engineering design. It is described that an application example of design archives management system for knowledge-base.

Keywords: Knowledge-Base; engineering design; design archives; intelligent search

1 Introduction

With the rapid development of network technology, database technology and artificial intelligence, artificial intelligence has penetrated into all spheres of society. It is an important research direction that the expert system of artificial intelligence applied to the management and engineering design. In expert systems, knowledge-base is the core of the design. This paper discusses the architecture, the technology characteristic and the design methods of knowledge-base. It is described that an application example of design Archives management system for knowledge-base.

2 The Structure of Knowledge-Base and Knowledge Representation

The structure of engineering design knowledge-base is shown in Fig. 1. Knowledge-base consists of five parts. That is knowledge-base body, knowledge acquisition part, external interface part, knowledge maintenance part and the results expression part.

2.1 Knowledge Representation

Human knowledge is expressed as a natural language. However, due to the ambiguity of natural language, grammatical and semantic description is difficult. In the knowledge-based systems, common methods of knowledge representation are predicate, production rule, meta-knowledge, frameworks and other expressions. Engineering design knowledge includes the following. Participation in professional, organizational design, high-stage design, new technology and optimize the design, construction drawings, technical specifications, project investment hot spots. Depending on the application project, knowledge representation use the object-oriented frameworks methods. Shown in Fig. 2.

2.2 Knowledge Acquisition

There are three ways that establishment of knowledge in the engineering design knowledge-base. First, knowledge is collected from multimedia database by means of data mining workbench. The second is to obtain from the data standards. Third, project chief designer input knowledge. There are hundreds of thousands of sets of archival documents in the archives department of design and research institute of Anshan iron and steel complex. Including research reports, budget estimate, project plan, preliminary design, calculations, contract documents, construction drawings and so on. Knowledge of engineering design knowledge-base is extracted by means of data mining workbench from the design database and borrowing database. Projects are organized by the knowledge content. knowledge is stored by project examples.

2.3 External Interface

After customers enter the design requirements, the inference-machine is started in the external interface part. The corresponding operations are processed.

2.4 Knowledge Maintenance

Knowledge maintenance part completes to increase, modifies or deletes knowledge in the engineering design knowledge-base.

2.5 Results Expression

According to user needs, the appropriate knowledge of knowledge-base is called. The results expression is in the form of report.

3 The Knowledge-Base Design

The design of knowledge-base is a process that knowledge representation method is realized in the computer system. Owing to the relational database structure is simple and easy to maintain, it is used widely in database management system. The knowledge representation will translate into a corresponding relationship model in the engineering design knowledge-base

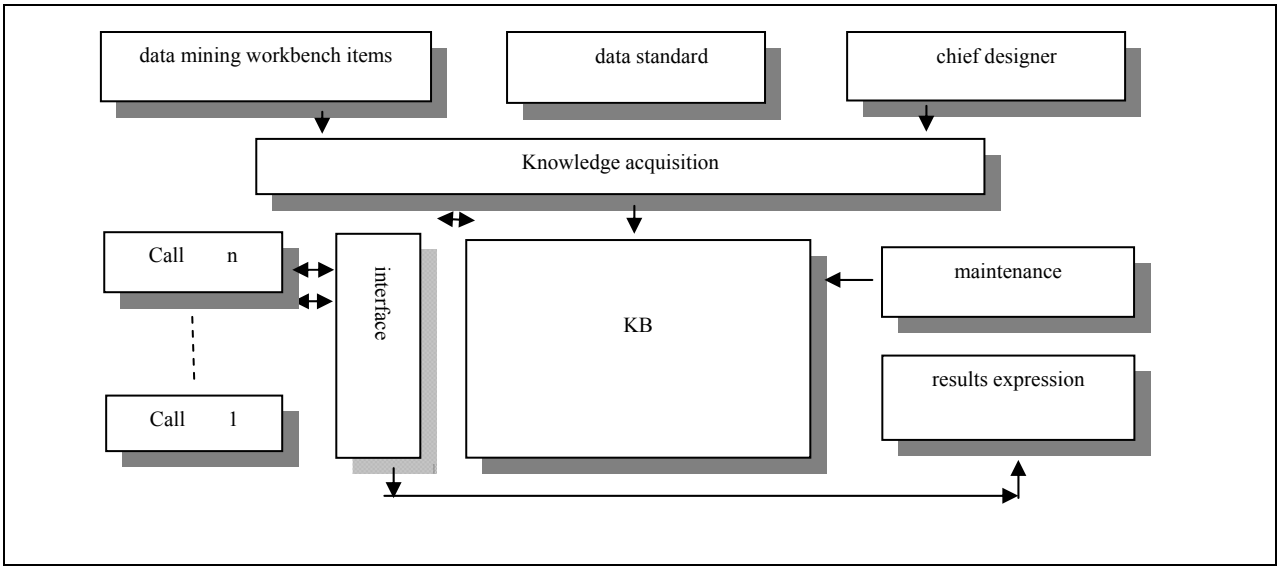


Figure 1. The structure of engineering design knowledge-base

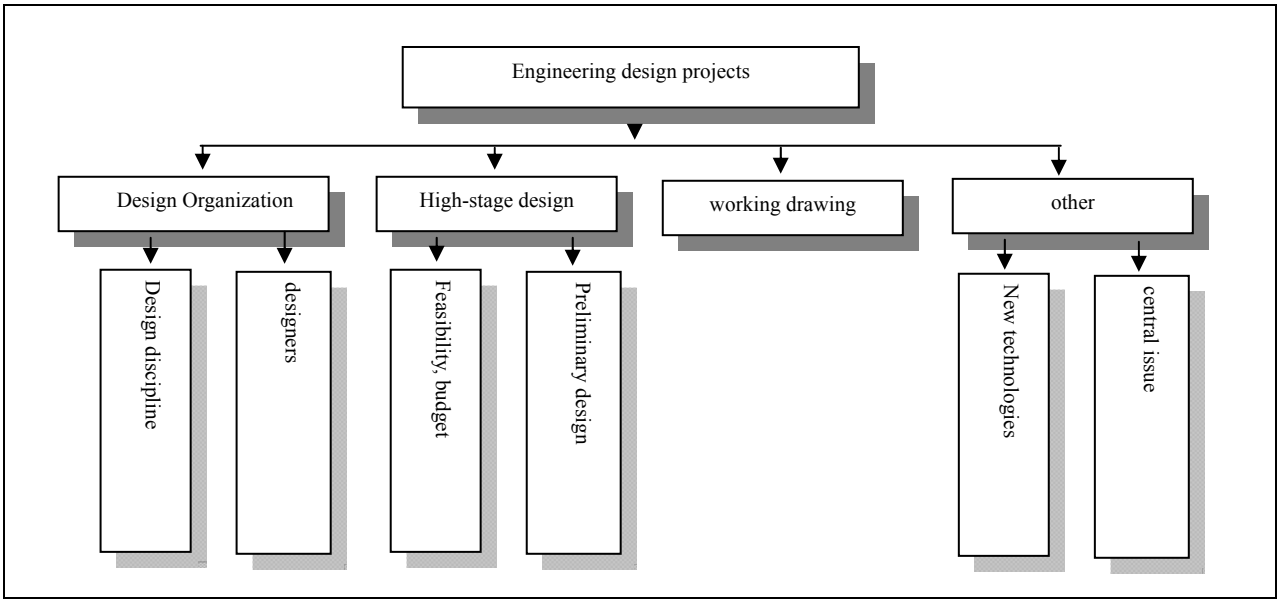


Figure 2. Engineering design projects

by means of the platform of relational database management system. In a relational database, data table is the basic unit of storage. Real world knowledge in the relational database is stored in the form of a table. The relationship between the knowledge is connected through the related field of the table. The management of knowledge is realized by means of the database management system.

4 An Application Example of Design Archives Management System

Engineering design archives management computer network system is an important part of the design and research enterprises and an operations center of the engineering design department. Traditional design archives management system based on database management system can only provide the certain data search and cannot effectively use the design archives information of management information systems that exist in a lot of potential, useful, large knowledge in the enterprise database system. Mining useful knowledge in the database, carrying out engineering design project is an important field of artificial intelligence research. We describe an embedded engineering design expert system in design

archives management system.

4.1 System Access

In Figure 3, select "Engineering Design Expert System" entry into the main control screen of engineering design intelligence system.

4.2 Entry Engineering Requirements

Figure 3. Main control screen

Figure 4. Engineering design expert system

In Figure 4, select the first item, enter the screen of engineering design requirement information as shown in figure 5. In the screen, owner, section, subsection, specialty, project stage and date is imported. Such as: injecting pulverized coal to blastfurnace, flow, major equipment, high stage and so on.

4.3 Output

Form of a report output:

Related projects: the new blastfurnace 1 #, BF 10 #

Related subject: iron, electric power, automation, airiness, gas, heat, telecommunications, engineering economics, civil engineering

Personnel organization: (select from the staff database)

Relevant documents: 22 KE 28, 3.18 KE 1

New Technology: injecting Coke Oven Gas to blastfurnace

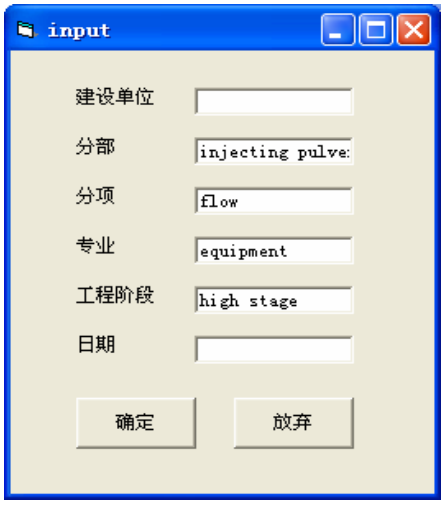


Figure 5. Engineering design requirement information

5 Conclusion

The example of design Archives management system of Design and Research Institute of Anshan Iron and Steel Group Corporation show that to embed the knowledge-base into the design Archives management system, to mine useful knowledge in the database, to achieve engineering design expert system is a useful exploration.

References

[1] Wang Wan-sen, “Principle and Application of Artificial Intelligence”, Electronic Industry Press, 2007.

[2] Ben can-ong, Zhang Yan-Duo, “Artificial Intelligence” Tsinghua University Press, 2006.

[3] Liu Tianhui, “Visual Basic Programming Guide”, Tsinghua University Press, 2006.

978-1-935068-76-1

Advances in Comparable Corpus Construction

Sa LIU, Chengzhi ZHANG
Department of Information Management, Nanjing University of Science and Technology, Nanjing, China, 210094
Email: liusa321@gmail.com, zhangchz@istic.ac.cn

Abstract: Parallel corpus is one kind of information resources and plays an important role in multilingual information services, such as machine translation, cross-lingual information retrieval. However, parallel corpus is scarce and difficult to be collected for some under-resourced languages or special domains. Compared to parallel corpus, comparable corpus is easier to be obtained by finding multilingual text collections with similar topics rather than find collections that are translations of each other. The study on building and using comparable corpus will enrich the field of information resources management. In this survey we first introduces characteristics of comparable corpus, explains some applications in multilingual information processing. Next, we describe advances in comparable corpus construction and the key technology related to it. Then, we point out some existing problems and introduce related work in our research group. Finally, we conclude the paper with a discussion of some applications of comparable corpus in cross-language information services.

Keywords: Comparable Corpus; Parallel Corpus; Multilingual Information Processing; Special Domain Corpus Construction; Multilingual Information Resource

可比语料构建研究进展

刘飒，章成志
南京理工大学信息管理系，南京，中国，210094
Email: liusa321@gmail.com, zhangchz@istic.ac.cn

摘　要：作为一种重要的多语言信息资源，平行语料在诸如机器翻译、跨语言信息检索等多语言信息服务中具有重要作用。但是，在某些语种或者专业领域内，平行语料仍然是稀缺资源，存在着资源获取瓶颈问题。相对于平行语料而言，可比语料不受源语言与目标语言互为翻译的约束，且更易于获取。可比语料构建与应用研究，在一定程度上可丰富信息资源处理与管理领域的研究内容。本文首先分析可比语料研究进展、描述可比语料构建中的关键技术，然后指出当前存在的问题，简要介绍本课题组的相关研究情况，最后讨论可比语料的应用。

关键词：可比语料；平行语料；多语言信息处理；专业领域语料库构建；多语言信息资源

1 引言

平行语料(Parallel Corpus)是由源语言文本及其翻译文本构成的语料[1]。平行语料在诸如机器翻译、跨语言信息检索等多语言信息服务中具有重要作用。然而，平行语料属于稀缺资源，获取代价高，对于一些语言资源比较匮乏的语种或领域，采集到大规模的平行语料是比较困难的；另一方面在机器翻译、跨语言信息检索中，领域语料比通用语料有更好的领域适应性，但已有平行语料基本局限在有限的领域内。因此，平行语料在资源获取、规模以及领域覆盖等方面都存在难以克服的缺陷，不能有效应对多语言环境下的信息处理需求。相比而言，可比语料(Comparable Corpus)由于没有受到源语言与目标语言互为翻译的约束，易于获取。互联网存在着大量的多语言资源，如双语网站、在线维基百科，易于从此类资源中获取大规模的可比语料。因此，可比语料构建与应用研究可以减少因平行语料匮乏造成的应用瓶颈，值得关注。

作为一种重要的多语言信息资源，可比语料的研究在一定程度上可丰富信息资源处理与管理领域的研究内容。本文分析可比语料研究进展、描述可比语料构建中的关键技术，然后指出当前存在的问题，简要介绍本课题组的相关研究情况，最后讨论可比语料的应用。

资助信息：本文系国家自然科学基金项目“基于可比语料的多语言文本聚类研究”(项目编号：70903032)的研究成果之一。

2 可比语料库的特点与应用

2.1 可比语料的特点

到目前为止，可比语料库的定义在学术界尚未有一致结论。Baker 于 1995 年最早提出可比语料库的概念：“由相关语言中译自特定源语的译本组成的语料库”与“由该语言中的非翻译源文本组成的语料库”共同构成可比语料库[2]。该定义强调不同的语料集合不具有翻译关系，且均为源语言文本。Bowker & Pearson 也认为不同语料的文档集合没有翻译关系，同时指出它们之间具有一些共同的特征[3]。季妲将语料之间共有的特征细化为主题[4]。Skadina 等人认为可比语料是根据一系列标准采集的相似文档集合[5]。

从可比语料的各种定义看出，可比语料至少具有三个特点：①源语言和目标语言文档均独立产生，不具有严格的翻译关系[1-3]；②源语言和目标语言文本均是在该语种文化环境下产生的自然语言文本[6]；③源语言和目标语言文档具有某种程度的相似，这种相似可能体现在采集原则、类型、领域、规模、主题、语料形式和功能等[1,4-6]。

相比于平行语料，可比语料还具有以下特点：①易于获取[4-6]，采集非平行的数据比平行数据更容易；②时效性强，大量的潜在可比语料资源更新速度快，如新闻报道，几乎每天更新[5]；③可比语料的对齐程度一般不如平行语料。其中前两个特点使可比语料的应用领域广为拓展，第三个特点表明可比语料的利用比平行语料更有难度。

2.2 可比语料在多语言信息处理中的应用

可比语料在多语言信息处理中的应用可分为两种情形：直接应用和二次开发后再应用。

可比语料在多语言信息处理中的直接应用是指利用可比语料中的文档可比关系，进行多语言映射转换，继而实现跨语言的信息处理，如 2009 年，Yogatama & Tanaka-Ishii 利用可比语料中的语义信息，进行跨语言的语义融合，实现以相似度传播算法为基础的多语言文本聚类[7]；2010 年，Zagibalov 等人则利用英-俄书评可比语料进行情感分析的研究[8]。

基于可比语料的二次开发，指的是通过可比语料抽取双语术语[9]、命名实体[10]、平行句子[11]等翻译知识，再利用这些翻译知识进行词典编纂[12]，或将翻译知识应用到跨语言信息检索[13]、统计机器翻译[14]等任务中。

Table 1. Projects on comparable corpus in machine translation
表 1. 可比语料在机器翻译中应用项目

项目名称	项目来源	项目起止时间	语料类型	涉及语种
TTC	欧盟 EU	2010.1.1~2012.12.31	专业领域	英、法、德、西班牙、拉脱维亚、汉、俄
Accurat	欧盟 EU	2010.1.1~2012.12.31	新闻、专业领域、维基百科	克罗地亚、爱沙尼亚、希腊、拉脱维亚、立陶宛、罗马尼亚
CC4EASMT	美国	2010~	新闻、维基百科	英、阿拉伯语

需要特别指出的是，最近国际学术界对可比语料在机器翻译领域的应用研究尤为重视。2010 年欧盟资助的 TTC 项目(Terminology Extraction, Translation Tools and Comparable Corpora)[1]和 Accurat 项目(Analysis and evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation)[2]，美国的 CC4EASMT 项目(Learning from Comparable Corpora for Improved English-Arabic Statistical Machine Translation)[3]都计划利用可比语料以实现机器翻译性能的提升，具体情况如表 1 所示 。

此外，第一届[4]、第二届[5]、第三届 BUCC[6](Building and Using Comparable Corpora)国际会议的陆续召开也表明了学术界对可比语料库的构建和应用的日益重视（第四届 BUCC 会议也将于 2011 年 6 月召开[7]）。从历届 BUCC 会议的征文主题也可以看出：可比语料被用于跨语言信息检索、机器翻译、跨语言文本分类等任务，逐渐引起学术界的广泛关注。

3 可比语料研究进展与构建关键技术

本节首先回顾了可比语料研究的三个阶段，然后

[1] http://www.ttc-project.eu/
[2] http://www.accurat-project.eu/
[3] http://www.qnrf.org/awarded/nprp/3/index.php?ELEMENT_ID=1328
[4] http://www.limsi.fr/~pz/lrec2008-comparable-corpora/
[5] http://comparable2009.ust.hk/
[6] http://www.fb06.uni-mainz.de/lk/bucc2010/
[7] http://www.limsi.fr/~pz/bucc2011-comparable-corpora/

978-1-935068-76-1 © 2011 SciRes.

分别从构建方法与可比度度量两个角度对可比语料构建的关键技术进行分析和比较。

3.1 可比语料发展的三个阶段

根据可比语料研究在语料类型、主要技术、语料用途、语料规模等方面的发展状况，将可比语料划分为三个阶段，如图 1 所示。

在可比语料研究初期即第一阶段（1995~2006 年），其特点是根据不同的应用创建小规模的可比语料。该阶段语料类型主要为新闻报道，语料中相似文档筛选的方法，主要包括基于外部特征和文本描述内容的特征匹配技术，构建的可比语料基本用于术语抽取[15]、词典编制[16]、跨语言信息检索[17]、多语言文本聚类[18]、机器翻译[19]、多语言文本分类[20]、其他粒度的翻译知识抽取如命名实体[10]、平行句子[11]提取等应用，且语料规模较小。

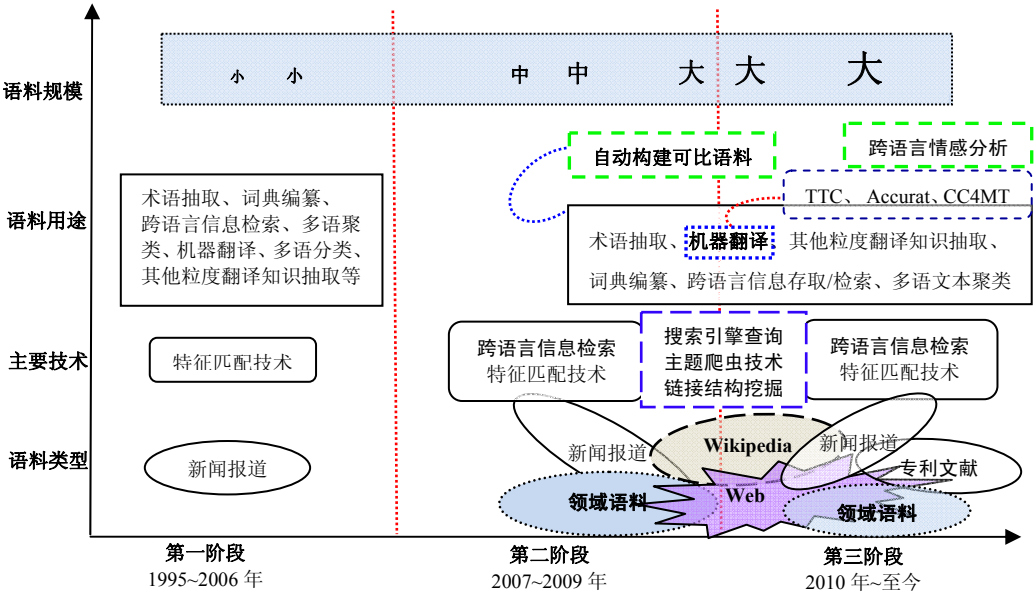


Figure 1. Three phases of development on comparable corpus

图 1. 可比语料发展的三个阶段

在第二阶段（2007~2009 年），用于构建的语料类型丰富起来，除了新闻报道，还包括领域语料和维基百科[8]，并出现利用网络构建新闻领域可比语料的研究[21]。跨语言信息检索被用于新闻可比语料的构建[13]，领域可比语料的获取主要通过主题爬虫(Topic Crawler)[22]和搜索引擎查询[23]技术实现，基于维基百科挖掘可比语料，主要依赖链接结构的挖掘[24]。该阶段可比语料在双语术语抽取[25-26]、机器翻译[27]、多语言文本聚类[7]、其他粒度翻译知识抽取[4]、双语词典编制[24]、跨语言信息存取和检索[28]等应用中的研究得到进一步发展，同时出现大量直接以自动构建可比语料为目的研究成果[21-23,29]。

第三阶段（2010 年~至今），是第二阶段的继承和发展，不但出现专利可比语料的研究[30]，且基于维基百科的可比语料研究受到更多的关注[31-33]，同时利用 Web 进行可比语料的研究引起学术界重视，如拟定于 2011 年 6 月召开的第四届 BUCC 会议，其主题即为“可比语料和网络”(Comparable Corpora and the Web)。该阶段可比语料在机器翻译[30]、其它粒度翻译知识抽取[33]、词典编制[34]等应用中的方法和技术更加成熟，出现利用可比语料进行情感分析的研究[8]，同时涌现大量利用可比语料提升机器翻译性能的研究项目，如前所述的 TTC、Accurat、CC4EASMT 等。

从图 1 不难看出，随着研究深入和应用驱动，可比语料研究的语料类型逐渐多样化，不但持续关注新闻语料研究，而且更加重视专业领域语料研究和利用互联网资源及维基百科挖掘可比语料；语料构建中采用的技术方法越来越智能化，越来越多的研究旨在探索自动构建可比语料及实现语料的对齐；最后可比语料的应用研究逐步完善，语料规模不断增大。

[8] http://www.wikipedia.org/

3.2 三种不同类型可比语料的构建方法分析

按照语料来源类型，可比语料可分为新闻、专业领域、维基百科等。由于不同类型语料的自身特点不同，相应的构建方法也有所区别。本节分别进行说明。

（1）新闻可比语料库构建方法

早期的可比语料采集研究大多是面向新闻领域，采集方法为：直接选取不同新闻机构发布的新闻报道作为生语料，通过报道的外部特征（发布日期）和文档描述内容（主题）的比较，利用特征匹配和过滤技术生成对齐文档，构建可比语料。1996 年 Sheridan & Ballerini[17]，1998 年 Braschler & Schauble[35]采用这种方法构建新闻可比语料。2007 年，Talvensaari & Laurikkala 等利用跨语言信息检索技术进行新闻可比语料构建[13]。2008 年，Gupta 利用网络爬虫采集新闻网站的报道，构建可比语料[21]。该研究与之前研究的最大不同在于：生语料集合来自开放的网络。2009 年，于海涛通过对中英文新闻网站的爬取，获取生语料；再利用跨语言信息检索进行双语映射和对齐，构建可比语料[29]，该研究的主要创新有两点：一是改进增量采集技术，解决可比语料时效性问题；二是深入研究跨语言信息检索中的未登录词识别问题，提高跨语言映射的准确率。2010 年，黄德根和李丽双等人使用跨语言信息检索进行新闻汉英可比语料构建，但重点研究基于最大熵模型的多词短语抽取和基于多项特征过滤的技术对语料采集的影响[36]。

（2）专业领域可比语料的构建方法

专业领域可比语料一般分为两个阶段：单语种语料采集和多语种语料过滤、对齐，其具体步骤为：①在搜索引擎中查询领域词表中的词语；②将返回文档认为是与查询关键词相关的，进行下载、保存，获取单一语言的文本集合；③利用领域过滤技术进行过滤，并通过跨语言映射实现文档对齐，生成可比语料。

专业领域可比语料构建的关键问题是如何获取高质量的领域词表。2008 年，Talvensaari & Pirkola 通过 Google 查询获取每种语言的领域词表，再利用主题爬虫技术采集可比语料[22]。2009 年，Leturia & Vicente 通过搜索引擎技术收集可比语料[23]。该研究中词表获取方法有两种：一是分别收集不同语种的领域语料，从中抽取关键词作为查询词；二是仅搜集一种语言的语料，从中抽取关键词，然后利用词典翻译该语言的关键词为另一种语言，从而获取目标语言的领域词表。不同方法获取的词表重合部分有多少，是否能互为补充进行语料采集，目前尚缺乏相关实证研究。

（3）维基百科可比语料构建方法

现有的研究中，利用链接结构挖掘可比语料一般有两个方法：一是首先从维基百科中下载不同语种的所有维基百科数据，然后使用语间链接进行双语对齐，即 “先下载所有数据，再进行双语对齐”；其二是首先收集词表，获取页面标题中含有词表中词语的单一语种页面，再使用语间链接采集含有该词语的其他语种维基百科页面，即“先下载部分数据，然后利用链接采集剩余数据”。

2009 年，Yu & Tsujii 先下载维基百科中相关语言的全部数据，再进行对齐映射[24]，生成可比语料。2010 年，Otero & L´opez 以维基百科的类别信息作为主题约束，以语言链接进行双语映射，实现基于维基百科的考古领域可比语料构建[31]。该研究中也是通过先下载全部数据再进行映射方法进行领域可比语料的构建。2010 年，Ion & Tufiş 通过“先下载部分数据，然后利用链接采集剩余数据”方法进行可比语料构建[32]。该研究首先从 WordNet 中抽取命名实体，接着下载含有这些命名实体的英语维基百科页面；然后通过挖掘语言链接结构，采集对应的罗马尼亚语和德语维基百科页面，生成英-罗马尼亚-德语可比语料。

表 2 为不同类型语料构建方法的比较。从表中看出：三种类型的语料因各自语料特点不同，所使用的语料构建技术也存在较大差异；新闻可比语料研究因起步较早，技术方法较为成熟，领域语料和维基百科构建研究起步较晚，成果少，仍有很多问题亟待解决。

Table 2. Comparison between different methods of constructing comparable corpus
表 2. 三种不同类型的可比语料构建方法的比较

类型	语料特点	主要技术	局限性
新闻语料	富含事件、日期、地点、人物等命名实体	特征匹配、跨语言信息检索、网络爬虫	特征匹配仅局限于新闻语料，跨语言信息检索中的查询翻译影响双语对齐结果
专业领域	领域术语大量存在	搜索引擎、主题爬虫	领域词表的质量和数量影响最终语料的质量和规模，语料采集受到搜索引擎技术影响
维基百科	页面标题为主题或概念，富含术语、链接结构	链接结构挖掘技术	语料质量和规模依赖对链接结构的挖掘

978-1-935068-76-1 © 2011 SciRes.

3.3 可比语料的可比度度量

与平行语料的定义“互为翻译的文本”不同，可比语料定义中“相似的内容”或“共同特征”等类似的表述本身就比较模糊，从而使得“可比语料”这个概念本身就存在一定的模糊性。可比语料是根据应用而创建的，需要根据具体应用场景对可比语料的“可比”和“相似”给出适当的描述、并规定度量指标。

在可比语料库的可比度和相似度度量方面，国内外研究中至今未形成一致的度量指标和系统的度量体系。2003 年，Belinda Maia 指出：在可比语料库的相似度方面，应该考虑的标准有形式和内容、语料的结构和功能、语料模式等；在可比度方面，应该考虑的指标有语料的规模、构建原则、语种内部和不同语种之间的比较、语料类型、语料领域等[6]。2005 年，Gliozzo & Strapparava 认为可以将两个语料词表中交叉词语所占的比例作为语料可比度度量的一个标准[20]。2006 年，Munteanu 根据可比语料中文档中含有的平行数据的多少，将语料中文本分为完全平行、噪声平行和完全不平行等三个层次[37]。以上的研究仅是从理论层面对一些可比度指标进行界定，没有进行实验分析。

2008 年，Saralegi & Vicente 等融合领域、报道类型、文档主题、发表日期等多个特征，估计新闻的总体可比度[29]。2009 年，Leturia & Vicente 等计算语料关键词的关联值衡量领域语料的可比度[26]。2010 年，Bo Li & Eric Gaussier 基于发现语料中每个词翻译的期望值来定义可比度计算公式[34]，并通过实验证实该公式的有效性。与以上通过语料内容度量可比度的研究不同，Sharoff 通过语料结构即组成要素来分析语料之间的相似和差异[38]。

可比度度量的关键是根据应用选择相应的度量指标，并且可比度的度量指标应该基于不同的应用有相应的侧重：机器翻译中，度量指标应该关注如何体现语料中翻译知识的质量；在跨语言信息检索中，度量指标应重点考察可比语料与查询词之间的主题相似性；对于术语抽取任务，度量指标可侧重反映子领域主题的相似性。需要说明的是，这些问题已经引起部分学者的重视，但在可比度计算研究方面目前仍未有实质性进展。

4 可比语料研究存在的问题

本文针对可比语料构建研究中存在的一些问题进行简单讨论。

（1）专业领域可比语料构建问题

在专业领域可比语料构建方面，主要存在三个问题。第一个问题是：现有的可比语料构建研究基本集中在新闻领域，而新闻可比语料采集方法难以有效应用到专业领域语料构建中。第二个问题是领域语料资源本身存在着严重的语种分布不均衡现象。第三个问题是：目前专业领域可比语料构建基本以种子词作为采集的起点，因此语料的质量受到初始种子词数量和质量的约束。

（2）基于维基百科的可比语料问题

利用维基百科挖掘可比语料的重要条件是维基百科是一个丰富的多语言资料库且富含语间链接。值得注意的是，维基百科中的语间链接是有限的，且不同语种的维基百科页面中也存在严重的信息不对称现象。因此，基于维基百科构建的可比语料规模和质量受到链接结构的影响。需要说明的是，该问题已经引起学术界的关注，如 NTCIR-9 中的跨语言链接测评任务[9]旨在通过对锚文本的挖掘，丰富维基百科的多语言链接结构。此外，维基百科的词条是有多人协作编写，词条质量和权威性不如领域专家编写的百科全书。

（3）可比度计算问题

现有的研究中语料内容的相似比较、语料组成的差异分析、语料规模、语料类型、语料可比度层次等仍然处于探索阶段，仅有少量研究对小规模语料进行实验，缺乏可信的可比度计算方法和指标。

（4）可比语料评估

在可比语料评估研究中，外部评估考察语料在具体应用中的效用，其结果依赖于特定的应用场景；内部评价利用语料内部特征的比较分析的结果，直观但有效性不易验证。鲜有研究将两个方面结合进行考虑。另外，国际上公开的可比语料测试集比较罕见，这也限制了可比语料评估的研究发展。

5 本课题组的相关研究

本节简介本课题组在可比语料构建上开展的研究工作，分别从研究框架和已开展工作两部分说明。

5.1 基本研究框架

针对可比语料研究中存在的问题，本课题组开展相关研究，研究框架包括：多策略的中英文可比语料构建研究、结合语料特点并以文档相似度和主题可比

[9] http://ntcir.nii.ac.jp/CrossLink/

度为基础的可比语料可比度分析、从内部评价和外部评价两方面入手对可比语料质量进行综合评价等。

5.2 已开展的工作

（1）平行语料的采集

利用双语网站在 URL 上的特性，自动获取相关的双语网页，为进一步构建可比语料提供资源基础。

课题组通过双语 URL 的匹配模式，获取高可信度匹配模式。利用高可信度匹配模式，结合网站的链接关系，进行双语语料的增量式获取。

实验结果表明基于增量式方法的双语语料采集结果，抽样结果正确率高达 99%，说明基于迭代方法获取双语语料的高效性。目前，课题组已经采集篇章级的平行网页近 40 万对，下一步将抽取平行句对。

（2）可比语料的采集

在前期研究中，本课题组利用维基百科完成了图书情报领域的可比语料采集。该工作的思路是：借助双语词表生成 URL 列表，再利用 Wget[10]下载维基百科词条页面。其中词表获取有两种方法：一是利用维基百科分类页面“图书资讯科学”和“Library and information science”中的页面标题；二是从学术数据库、网络中抽取。

Table 3. Domain comparable corpus from Wikipedia
表 3. 利用维基百科采集的领域可比语料信息

语种	词表资源	数目	下载页面	大小	去重后
中文	维基页面标题	125 个	119 个	3.82M	321 个 12.8M
	图书情报词条	707 个	206 个	9.30M	
英文	维基页面标题	183 个	178 个	6.8M	674 个 38.2M
	图书情报概念	913 个	510 个	32.0M	

表 3 为图书情报领域维基百科语料基本信息。从中可以看出，直接利用维基百科的页面标题采集到的语料是非常有限的。因此本研究通过文献数据库、网络等收集更多领域词条，采集维基页面。该方法获取的语料规模远远大于直接用页面标题采集的语料，并且重合的页面较少，说明两种方法融合后可以获取更大规模的语料。

此外，课题组还采集百度百科的 160 多万条词条。百度百科资源可用于维基百科和百度百科的语料融合，从而丰富可比语料中的中文语料部分。

[10] http://www.gnu.org/software/wget/

6 可比语料的应用

可比语料构建的兴起原因之一在于：不受翻译限制，比平行数据获取容易。随着可比语料构建方法不断完善，利用已有的技术方法容易获取大规模高质量的可比语料。一方面这些语料可以缓解由于平行语料匮乏而引起的资源瓶颈问题，跨越语言障碍实现信息资源的充分利用，从而有效解决多语言环境下的信息资源整合问题。另一方面，由于可比语料存在主题上相似等特点，今后可比语料在多语言信息环境下的学科领域热点监测、跨语言抄袭监测、多语言专题数据库构建等应用中，将发挥重要作用。

致　谢

非常感谢审稿专家的宝贵意见。

References（参考文献）

[1] McEnery A, Xiao R. Parallel and comparable corpora: What are they up to? [C]. In: Proceedings of Incorporating Corpora: Translation and the Linguist Translating Europe Multilingual Matters, Clevedon , UK. 2007.

[2] Baker M. Corpora in Translation Studies: An Overview and Some Suggestions for Future Research [J].Target, 1995, 7(2): 223-243.

[3] Bowker L, Pearson J. Working with Specialized Language: A Practical Guide to Using Corpora [M]. London/New York, 2002.

[4] Ji H. Mining name translations from comparable corpora by creating bilingual information networks [C]. In: Proceedings of the 2nd Workshop on Building and Using Comparable Corpora: from Parallel to Non-parallel Corpora, Suntec , Singapore. 2009: 34-37.

[5] Skadiņa I, Aker A, Giouli V, et al. A Collection of Comparable Corpora for Under-resourced Languages[C].In: Proceedings of the Fourth International Conference Baltic HLT2010, Riga, Latvia. 2010:161-168.

[6] What are Comparable Corpora? [EB/OL]. [2010-12-28]. http://web.letras.up.pt/bhsmaia/belinda/pubs/CL2003%20workshop.doc

[7] Yogatama D, Tanaka-Ishii K. Multilingual spectral clustering using document similarity propagation [C].In: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing , Singapore. 2009: 871-879.

[8] Zagibalov T, Belyatskaya K, Carroll J. Comparable English-Russian Book Review Corpora for Sentiment Analysis [C]. In: Proceedings of 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis, Lisbon, Portugal, 2010.

[9] Morin E, Daille B, Takeuchi K, Kageura K. Bilingual terminology mining - using brain, not brawn comparable corpora. [C].In: Proceedings of 45th ACL, Prague, Czech Republic.2007: 664-671.

[10] Klementiev A , Roth D. Weakly supervised named entity transliteration and discovery from multilingual comparable corpora [C].In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics (ACL-44), Morristown, NJ, USA.2006: 817-824.

[11] Munteanu D S, Marcu D. Extracting parallel sub-sentential fragments from non-parallel corpora [C]. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, Sydney, Australia. 2006:81-88.

978-1-935068-76-1 © 2011 SciRes.

[12] Fung P. A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-parallel Corpora. [C]. In: Proceedings of the Third Conference of the Association for Machine Translation in the Americas on Machine Translation and the Information Soup (AMTA'98), London, UK. 1998: 1-17.
[13] Talvensaari T, Laurikkala J, Järvelin K, et al. Creating and Exploiting a Comparable Corpus in Cross - Language Information Retrieval [J]. ACM Transactions on Information Systems (TOIS). 2007, 25(1): 322-334.
[14] Munteanu D S, Marcu D. Improving Machine Translation Performance by Exploiting Non-Parallel Corpora [J]. Computational Linguistics, 2005, 31 (4): 477-504.
[15] Rapp R. Identifying word translations in non-parallel texts [C]. In: Proceedings of the 33rd annual meeting on Association for Computational Linguistics, Cambridge, Massachusetts.1995: 320-322.
[16] Fung P. Compiling bilingual lexicon entries from a non-parallel English-Chinese corpus [C]. In: Proceedings of the 3rd Annual Workshop on Very Large Corpora, Boston, Massachusetts. 1995: 173-183.
[17] Sheridan P, Ballerini JP. Experiments in multilingual information retrieval using the SPIDER system [C]. In: Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval, Zurich, Switzerland. 1996:58-65.
[18] Chen H H, Lin C J. A Multilingual News Summarizer [C]. In: Proceedings of the 18th conference on Computational linguistics(COLING '00), Morristown, NJ, USA. 2000: 159-165.
[19] Munteanu D S, Fraser A, Marcu D. Improved machine translation performance via parallel sentence extraction from comparable corpora [C]. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT/NAACL2004), Boston, USA. 2004:265-272.
[20] Gliozzo A, Strapparava C. Cross language text categorization by acquiring multilingual domain models from comparable corpora [C]. In: Proceedings of the ACL Workshop on Building and Using Parallel Texts (ParaText'05), Morristown, NJ, USA. 2005:9-16.
[21] Gupta M. Automatic Acquisition of Bilingual Comparable Corpus [D]. Indian Institute of Technology, 2008.
[22] Talvensaari T, Pirkola A, Järvelin K, et al. Focused web crawling in the acquisition of comparable corpora [J]. Information Retrieval. 2008, 11(5):427-445.
[23] Leturia I, Vicente I S, Saralegi X. Search engine based approaches for collecting domain-specific Basque-English comparable corpora from the Internet [C]. In: Proceedings of the Fifth Web as Corpus Workshop (WAC5), Basque Country, Spain. 2009: 53-61.
[24] Yu K, Tsujii J. Bilingual dictionary extraction from Wikipedia [C]. In: Proceeding of MT Summit XII, Ottawa, Canada, 2009.
[25] Morin E, Daille B, Takeuchi K, Kageura K. Bilingual Terminology Mining- Using Brain, not Brawn Comparable Corpora [C]. In: Proceedings of the 45th Annual Meeting of the ACL, 2007: 664-671.
[26] Saralegi X, Vicente I S, Gurrutxaga A. Automatic Extraction of Bilingual Terms from Comparable Corpora in a Popular Science Domain [C]. In: Proceedings of LREC2008 Workshop on Comparable Corpora, 2008: 27-32.
[27] Hewavitharana S, Vogel S. Enhancing a Statistical Machine Translation System by using an Automatically Extracted Parallel Corpus from Comparable Sources [C]. In: Proceedings of the LREC'2008 Workshop on Building and Using Comparable Corpora, Marrakesh, Morocco. 2008:7-10.
[28] Lee L, Aiti A, Thuy V, et al. MARS: multilingual access and retrieval system with enhanced query translation and document retrieval [C]. In: Proceedings of the ACL-IJCNLP2009 Software Demonstrations, Suntec, Singapore. 2009: 21-24.
[29] Yu Haitao. The Research and Construction of Comparable Corpus[D].Dalian University of Technology,2009(Ch).
于海涛. 可比较语料库的研究与构建[D].大连理工大学,2009.
[30] Lu B, Jiang T, Kapo C. Building a Large English-Chinese Parallel Corpus from Comparable Patents and its Experimental Application to SMT [C]. In Proceedings of BUCC2010, Malta, 2010:42-49.
[31] Otero P G, L´opez I G. Wikipedia as Multilingual Source of Comparable Corpora [C]. In: Proceedings of the 3rd Workshop on Building and Using Comparable Corpora, LREC2010, Malta. 2010: 21-25.
[32] Ion R, Tufiş D, Boroş T, et al. On-Line Compilation of Comparable Corpora and their Evaluation [C]. In: Proceedings of the 7th International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL7), Dubrovnik, Croatia.2010: 29-34.
[33] Smith J R, Quirk C, Toutanova K .Extracting parallel sentences from comparable corpora using document level alignment [C].In: Proceeding of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, Los Angeles, California,US.2010:403-411.
[34] Li B, Gaussier E. Improving Corpus Comparability for Bilingual Lexicon Extraction from Comparable Corpora [C]. In: Proceedings of the 23rd International Conference on Computational Linguistics (Coling2010), Beijing, China. 2010:644-652.
[35] Braschler M, Schäuble P. Multilingual Information Retrieval based on document alignment techniques [C]. In: Proceedings of the Second European Conference on Research and Advanced Technology for Digital Libraries, Greece. 1998: 183-197.
[36] Huang DG, Zhao L, Li L S, Yu H T. Mining Large-scale Comparable Corpora from Chinese-English News Collections [C].In: Proceedings of Coling2010, Beijing, China. 2010:472-480.
[37] Munteanu D S. Exploiting Comparable Corpora [D]. Information Sciences Institute, University of Southern California, USA, 2006.
[38] Sharoff S. Analysing Similarities and Differences between Corpora[C].In: Proceedings of the 7th Conference "Language Technologies", Ljubljana , Slovenia. 2010.

Education Information Technology in China Meteorological Administration

Zhenghong CHEN[1], Guifang YANG[2]

[1]*China Meteorological Administration Research & Development Centre Beijing, 100081*

[2]*School of Earth Sciences and Resources, China University of Geosciences Beijing, 100083*

Email: chenzhengh@cma.gov.cn, yangguifang@cugb.edu.cn

Abstract: This contribution attempts to characterize the China meteorological administration training hosted by China Meteorological Administration Training Centre (CMATC) and relates it to the current international settings. As a higher education & training organization previously, CMATC has been serving as the national base for higher continued education and on-the-job training in the meteorological field. The main task of CMATC is to promote the training for the administrative staff at middle or higher levels in China meteorological departments, such as the novel and high-level technicians, operational backbones and instructors, and compile the meteorological training materials. On the other hand, a remote education & training system typical of the meteorological division has been formed and largely improved. Additionally, we conclude that it should be a long-term development process for Chinese social education in meteorological field to entirely integrate into the current international settings.

Keywords: Chinese education information technology; CMATC; remote; meteorology

1 Introduction

The aim of the study, in conjoint with previous work, is to characterize the current education information technology system in Chinese Meteorological Administration with specific references to the training centers. A short review of development history regarding social education in this department was implemented and present institutions were summarized for better exploring the future orientation, allowing more intuitive access to this specific field. Our results will definitely facilitate the detail study of the development history and sustainable management of social education involved in meteorology in China.

Presently, the major bodies for social education in China are the subordinate administrative institutions in national ministries and local governments. These agencies usually keep inextricable correlations with the government departments, or even initiate considerable responsibility for the industrial management or guidance. They are also in charge of the enhancement the professional abilities of the workers in their fields.

2 Education Information Technology in China Meteorological Field Training System

Following the nowadays trends of science and technology in meteorological field, the Chinese meteorological training system actively extend its outreach application areas, training scale and effects. The vigorous development of China's meteorology also requires a systemic training process for most staff and managers. There are over 50,000 employees and more than 2,000 weather stations. There are a huge number of meteorological employees who are hungry for the job training.

Chinese meteorological social education system is currently comprised of the national, regional and provincial levels, offering great training bases for most of the people who involved in the meteorological work and study. The China Meteorological Administration Training Centre (CMATC), the only one national meteorological social education organization, has made tremendous contribution to the Chinese meteorology by providing various trainings not only for people within meteorological departments but also for the people from civil aviation, agriculture and other industries. Under the guidance of CMATC, a large variety of regional and provincial training centres also enhance the meteorological team-building. CMATC and regional/provincial training centres constantly invests lots of money in training construction and promotes the meteorological staff to grasp the new technologies and associated information. They together have built the meteorological remote training system serving as the platform for all over the country.

3 Study of Education Information Technology in CMATC

The CMATC was originally from the Beijing Meteorological College, the first professional school in meteorology built in 1954. In 1979, just after the Cultural Revolution, the college reenrolled the students in meteorological field. And in 1984, the Beijing Meteorological College was upgraded as Beijing Meteorology University and further renamed as the China Meteorological Administration Training Center—CMATC in 1999.

CMATC chiefly provides meteorological services in

978-1-935068-76-1

continuing education for the middle to high-level management cadre, professional and technical personnel. CMATC also charges for meteorological operations and technical promotion training in meteorology. Especial CMATC emphasizes not only training on forecasting and foresight of weather and climate but also cultivating meteorological managers. Additionally, its training programs usually involves as follows:

(1) Teacher training and Meteorological training materials production;

(2) Meteorological distance education and training;

(3) pre-service training for staff who would study abroad and other relevant foreign language training, as well as business and management training for meteorological personnel in the Asian region and other developing countries;

(4) Construction of meteorological distance education and training system;

(5) Operational guidance for the provincial (autonomous regions and municipalities) Bureau of Meteorology Training Centre and so on.

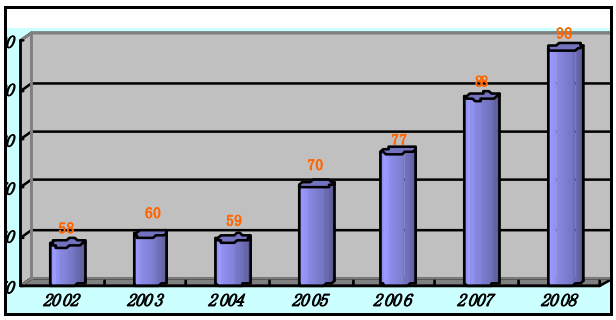


Figure 1. The number of CMATC training sessions in the duration of 2002-2008(Courtesy by CMATC annual report)

After Beijing Meteorology University was renamed as the China Meteorological Administration Training Centre --CMATC in 1999, it carried out lots of social education work (Fig.1).

Up to the end of 2008, CMATC has organized various training courses for more than 605 times, during which a total of 26,425 people were involved. And additional more than 170 thousand persons attended the remote training courses. Especially in 2008, around 98 training sessions was held, almost every person having 10 hours for distance learning. It is, however, the great achievements in the social education of meteorological field in China.

CMATC had been actively expanding the international training programs in recent couples of decades. In 2003, CMATC has become one of the important branches of World Meteorological Organization (WMO), Regional Meteorological Training Centre of Nanjing Division in Beijing. This has remarkably strengthened the collaboration with the Chinese Ministry of Commerce, Ministry of Science and Technology, Food and Agriculture Organization (FAO) and other international departments. These organizations positively carry out various training programs regarding drought and water resources, disaster prevention and mitigation, climate change, weather satellite applications, near-forecasting and other kinds of international training courses. By the July of 2008, CMATC has already trained 352 foreign scholars (including bilateral training). In addition, the CMATC has developed the international field of training and sent outstanding teachers to Vietnam to present 3 international training sessions for nearly 150 meteorological workers.

4 Remote Education Information Technology in CMATC

The complete remote training system of CMATC has also yielded great impact on the meteorological remote training field. In 2007, CMATC had successfully organized CALMET VII international conference[1] which supported by WMO. Over the international conference, China's meteorological distance education (or CMATC's remote training) has received the highest honor from WMO and peer educationists from the United States, Britain, Australia and other countries due primarily to its advanced technology and standardized training process (Fig.2). WMO has examined and approved the CMATC as the excellent virtual laboratory for satellite meteorology.

Significantly, many advances have been made in terms of teaching and scientific research recently. Since 1981, CMATC's scholars have published thousands of papers and more than 30 monographs. A big number of achievements have received the national awards and produced significant social benefits in social education field.

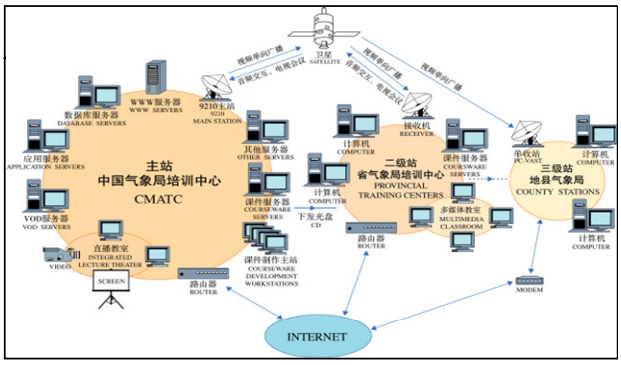


Figure 2. CMATC's remote training system (from professor Shen's report)

5 Future Orientation

CMATC's future aim is to build the advanced international and domestic first-class training base for meteoro-

[1] 2007 CALMet VII international conference, Beijing, China, http://stream1.cma.gov.cn/calmet/calmet.htm.

logical social education and cultivate the high-quality research teams. To arrive at this goal, CMATC can recruit the excellent academic elites both at home and abroad, particularly invite high-level educationists and meteorological training experts, scholars and academic leaders to join the CMATC. CMATC also try to incubate key teachers with sound theoretical background and innovative abilities within the future 3-5 years.

The CMATC should make full use of domestic and international educational and academic resources to establish a more open, interactive training mechanism and strengthen cooperation with universities and institutes to exchange and share resource in terms of meteorological business, service areas of the modernization of science and technology issues. By doing these, CMATC attempt to expand the meteorological operations and cultivate more high-level comprehensive talents.

And CMATC will compile relevant training materials or lectures according to the latest achievements of weather-related disciplines and business. It will also establish standardized training materials for approaching the scientific meteorological education process according to the requirements of meteorological social education and training system.

Additionally, CMATC will further optimize the organizations of library system and strengthen the meteorological science and technology literature information service function. CMATC would maybe change its name to ‘Meteorological Science and Technology Leadership Academy" in the near future.

6 Conclusion

The Chinese social education is an opening system, comprising the China government training institutions, associated universities, social training institutions and international education/training institutions. Three levels of training systems have already been established at the national, regional and provincial levels. The Chinese government has greatly enhanced fewer but key training bases of national and regional levels and tried to form unique training system by entirely combining the educational resources at home and abroad.

In addition, Chinese social educational institutions have paid more attention to the construction of teaching staff in both professional and part-time dimensions. More part-time personnel and fewer full-time employees will be required for numerous remote training tasks as well as a small number of face-to-face training.

Notably, the further expansion of China's opening-up largely requires the personnel with high quality and creative ability due to the intensive internationalization and globalization. In this regard, a large number of professionals can not have access to formal academic education. They might need continued education and social training to update their knowledge and technical skills. On the other hand, after China joined into WTO, Chinese educational service market is gradually opening up and a large number of foreign and overseas training institutions and high-quality education and training resources are swarming into China. It is a new inevitable trend for social education institutions to create an international platform for Chinese education technology development.

References

[1] Chen, B.. 2007. The Social Education in Community and Its Revelations to Modern Education. China Agricultural University Journal of Social Sciences Edition.

[2] Feng, Q.X., Wang, L., 2009. Social Education and the Construction of Civic Virtues. Journal of Shenyang Normal University (2).

[3] Fu, X.L., 2003. Reasons for Emergence and Development of Social Education in Modern China. Journal of Southwest Jiaotong University (Social Sciences) 4(4), 1-7.

[4] Gao X.H., 2007, Meteorological Distance Learning: Theory, Practice and Influence on Culture Construction, 2007 CALMet VII international conference, Beijing, China.

[5] Hou, H.Y, ，2008. Several Issues of Social Educational Research in China. Educational Research 29(12), 39-43.

[6] Kang, Y., 2006. The Theoretical Analysis on the Educational Equity of China. Lanzhou Academic Journal (8), 182-184.

[7] Qu, T.H., Yuan Y., 2006. The social education in the Chinese Soviet areas and its value. Journal of Hebei Normal University(Educational Science Edition) 8(6), .

[8] 8. Shen X. N., 2007, China Meteorological Administration Training Centre, 2007 CALMet VII international conference, Beijing, China.

[9] Zhou, H.M., 2007. Teacher Education Model of Social Education and Its Characteristics in the Time of the Republic of China. Teacher Education Research.

[10] Zhang, X.Q., The central government and social education: a perspective of the science education, Journal of Hebei Normal University (Educational Science Edition) 9(2).

978-1-935068-76-1 © 2011 SciRes.

The Application of Grid Technology in Knowledge Management

Yunhong WANG

Centre for Resource Sharing Promotion, Institute of Science and Technology Information of China, Beijing, China

Email: wangyh@istic.ac.cn

Abstract: Grid technology is an emerging information technology which has been applied to many fields. This paper introduces the concepts, the developments, and the architecture of Grid as well as Knowledge Grid systematically, discusses the application and importance of Knowledge Grid, finally concludes the problems and future of the Knowledge Grid.

Keywords: Grid; Knowledge Grid; Knowledge Management

1 Introduction

In the 1990s, the emergence and increasing popularity of Internet and WWW brings unexpected profound impact to our life, science and technology as well as economy. After almost 20 years' development, WWW technology has been used to serve people in difference areas. More and more scientists begin to think about the question: what is the next generation of information technology? Will www technology be substituted by another advanced technology?

In the era of mega science, there are a lot of demands on resource integration, resource sharing and complex computing. Scientists are trying to integrate the computing resources including CPU, storage devices, databases etc. from different locations in order to provide high-performance service. That is the origin of the Grid invented. The Grid integrates many distributed high-performance computers, large-scale databases, expensive devices (such as electron microscope, array radar, particle accelerator, and astronomical telescope), telecommunication devices, visualization devices and different sensors into a super computer to support the scientific computation.

Just as the expert of the Grid Ian Foster predicted, the Grid technology will be applied more widely. After almost 20 years study on the Grid, we realized gradually that the Grid technology could be applied to different fields form types of grids, such as Computational Grid, Information Grid, Knowledge Grid, and Manufacturing Grid. All these grids have the same idea "resource sharing and collaboration".

The application of the Grid technology in knowledge management formed Knowledge Grid superior to the original knowledge management pattern. With the development of technology, the Knowledge Grid will be further improved, and knowledge sharing will be realized in the "global intelligence Grid" leading to knowledge innovation and more knowledge value.

2 Generation and Implication of the Grid

2.1 The Generation of the Grid

The generation of the Grid could be dated back to the end of 1980s or early 1990s (Figure 1). At that time a batch of European scientists tried to integrate different kinds of computation resources to provide high-performance service. Hence the scientists integrated the computation resources in France and Switzerland via high speed internet. The results of this experiment indicates that the scientific researchers in Switzerland could submit their computing work to the high performance computer cluster located in France instead of working on their own computer system without wasting more money. In the process of using the resources, the experts found using the computer cluster far away was the same convenient as using the local computer without considering where the resources came from. Using the Grid is just like using electricity.

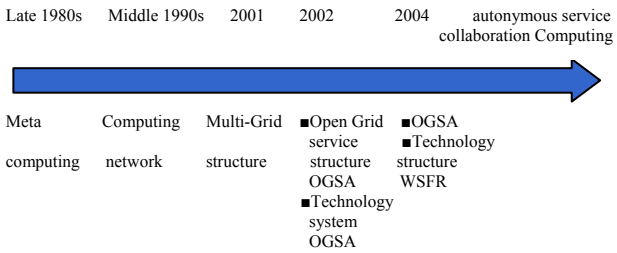


Figure 1. The timeline of the development of Grid

2.2 The Implication of the Grid Technology

As an emerging technology the Grid has already been applied to business area, but it still needs to be developed. The Grid could integrate, configure, manage and coordinate the resources scattered in different physical places via Internet. The technology could connect all different computers into a grid and every computer could be a nod

of it. Then the Grid has super computing and processing capacity.

The pioneer of the Grid area, Ian Foster, the senior scientist of the Argonne National laboratory and the leader of program Globus, published a book titled "The Grid: A New Infrastructure for 21st Century Science". In this book he described the Grid as "a set of emerging technologies, linking sensors, high-speed internet, high performance computers, and large databases, which provides researcher and ordinary people with more resources, functions and interactions. Internet could only provide service like email service, webpage browsing, comparatively the Grid could give people more chances to use distributed computational resources transparently." In 2000 Ian Foster described the Grid as "a technology to share resources and solve problems by synergy between the dynamic changes of many virtual institutions". Distributed institutions could form virtual organization to share different resources on computing, storage, information, software, telecommunication, knowledge and expertise etc. via the Grid. [2]

3 Transformation from Document Management to Knowledge Management

Knowledge is gradually regarded as the focus of the scientific innovation with the transformation from industrial age into knowledge economy.

More and more experts home and abroad queried the literature management when information technology was applied to manage information resources. Then knowledge management emerged and attracts a lot of researcher's interest. The existing knowledge organization methods(such as bibliographical reference, index, abstract, literature database etc.) has two disadvantages: first knowledge organization is to organize the knowledge content instead of the literature——the media of knowledge; secondly the target literature could only indicate the clue where the knowledge is located, but can't reveal the inner relationship between the knowledge which is important to knowledge innovation.

The Grid is superior to Internet in knowledge production and resources sharing. The development of Grid technology provides the possibility of transformation from literature management to "knowledge unit" management. The emergence of Grid technology could bring the revolution for knowledge management, and Knowledge Grid will be a platform of knowledge innovation service.

4 Application of Knowledge Grid in knowledge management

4.1 The Concept and Features of Knowledge Grid

American Information scientist, Fran Berman, chief researcher of program TeraGrid, pointed out that Knowledge Grid is to "extract, synthesize knowledge based on the technology of grid, data mining and reasoning in order to make search engine more intelligent to reason, answer questions and make conclusions." using the Grid, with the support of data mining". From the definition we could know that Knowledge Grid is the combination of the Grid and knowledge management.

Knowledge Grid is a intelligent, connected environment that would help the users to get access, publish, share and manage the knowledge resources, and also provide the knowledge service to support knowledge innovation, cooperation, problem-solving and decision making.

Knowledge Grid is a human-computer connected environment. The existing technology such as internet, Grid, knowledge discovery, information retrieval, Q&A, human intelligence could support the research and application of Knowledge Grid.

Knowledge Grid has 5 characteristics:

1) Getting access to and managing knowledge distributed globally from a single semantic entrance without knowing where the knowledge is located;
2) Pooling the distributed knowledge and providing intelligent knowledge service based on the technology of reasoning and interpretation mechanism;
3) No obstacles in sharing the service based on reflecting, restructuring, abstracting in a single semantic space;
4) Searching knowledge globally in order to provide suitable knowledge for problem solving;
5) Knowledge stored and updated dynamically; Knowledge Grid could configure knowledge dynamically with the grid computing, and in the process knowledge service will be improved.

As a complex system, Knowledge Grid has the characteristics: integrity, comprehensiveness, structured, dynamic, hierarchical, evolving etc.. [5]

4.2 Structure of Knowledge Grid

Knowledge Grid consists of five layers: construction layer, connectivity layer, resource layer, pooling layer, application layer [6]. Figure 2 indicates five layers of Knowledge Grid.

Construction layer controls different kinds of resources such as network, computers, sensors etc., and provides the interface for resources sharing. Connectivity layer regulates core protocol of communication and security in order to improve the safety and simplicity of communication. Resource layer sets the sharing operation protocol on single resource and provides a protocol interface of resource access. Pooling layer coordinates different resources in Knowledge Grid and manages the whole process and cross area inter-operation. Application layer provides different tools to provide new service via program interface and collect knowledge on the grid to complete the task.

978-1-935068-76-1 © 2011 SciRes.

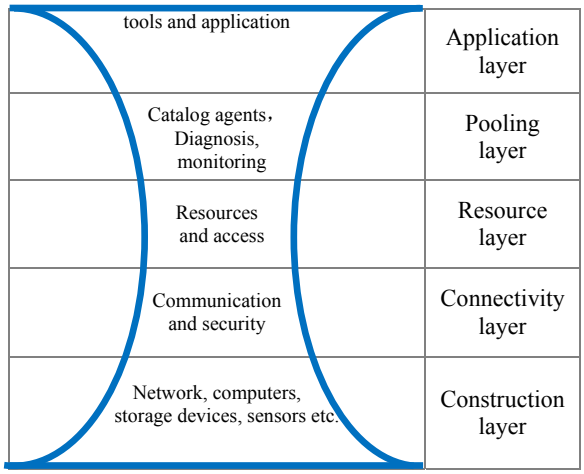


Figure 2. Five-layer structure of Knowledge Grid

4.3 Function of Knowledge Grid in Knowledge Management

Knowledge management is to manage codified and tacit knowledge, transferring proper knowledge to people who need on proper time. The fundamental objective of knowledge management is to facilitate knowledge innovation. Knowledge Grid could support knowledge management in the following ways.

4.3.1 Facilitating Users to Get Access to Knowledge and Innovation

In the Knowledge Grid any people could be the knowledge source or receiver, the grid could help people or the virtual people to acquire or publish their knowledge and support the versatile representation of knowledge with the platform.

Knowledge Grid could connect and integrate knowledge resources of different levels and different areas in order to support analogy inference, problem solving and scientific discovery. Knowledge grid could improve the knowledge exchange, innovation and the value.

4.3.2 Effective Knowledge Organization and Exchange

Knowledge Grid could organize knowledge by the semantic norm in order to ensure the effective searching. Meanwhile Knowledge Grid could wipe off the redundancy, refine the knowledge repository and expand the useful knowledge to improve the knowledge quality.

Knowledge Grid could boost the diffusion and sharing of knowledge in the network which resulting the knowledge exchange and utilization.

4.3.3 Knowledge Management More Intelligent

Intelligent means convenience, interface consistence, active service, less user input and operation, flexible platform, and personalized service.

Knowledge Grid is to use Grid technology to facilitate knowledge sharing and manage knowledge system, which use specific software and provide a united entrance to access to file system, archive system, database system and overcome the heterogeneity. Hierarchical design of metadata bring the possibility of transparent resource access and beneficial to knowledge management.

4.3.4 United Solution and Flexible Service for Knowledge Management

Knowledge Grid provides a united solution for users from knowledge repository to knowledge reusing. To establish a service platform doesn't need reprogramming, only administrator's simple configuration. On this platform information process is distributed, coordinated and intelligent, users could access to all information via a single entrance. Intelligent means convenience, interface, active service, less user input and operation, flexibility and personalized service. The intelligent characteristic of the platform, the superiority of Knowledge Grid, will affect the users' attitudes.

4.3.5 Transferring Explicit Knowledge into Tacit Knowledge

In Knowledge Grid knowledge units of the text documents could be knowledge nods. The Knowledge units are not isolated from each other, but have relationship among them, which could form a knowledge chain. In the knowledge chain, the connections between knowledge units are the important ways to generate knowledge. Different chain

Different connections could transfer tacit knowledge into explicit knowledge from different fields, and finally generates knowledge innovation or creates a new research area.

5 Results

In knowledge society how to learn and acquire knowledge is a very important question and activity that people are focused on. The application of the Grid technology will be a new solution for knowledge management. Knowledge Grid's connecting every individual and knowledge boosts the rapid knowledge innovations. Knowledge Grid will change the ways of knowledge production, diffusion, innovation. [7]

But it is more difficult to set up Knowledge Grid than computing Grid or information Grid because Knowledge Grid is not merely technological problems. To resolve the difficulty needs more collaboration between computer experts, scientists and experts in other fields.

Knowledge Grid would bring bright future for knowledge management, but there are still a lot of problems need to be solved, such as the problem of intellectual capital, security of network resources, and personalized service for different fields in the process of knowledge sharing. Whether numerous data and computational resources could be shared or used across different organizations is an important problem need to be solved. Al-

though sharing resources is very important, but it should have authorities to different users in order to show respect to intellectual property.

In summary the emergence of knowledge grid brings a lot of superiority compared to traditional knowledge management pattern. With the development of Knowledge Grid technology, it will bring the knowledge transform of "Global Intelligence Sharing" and will be more value in the knowledge economy.

References

[1] D. Fang, J. Liu, and G. Wang, "Research on Coordinated Business Information Infrastructures Based On Grid Technology ", China Association for Information Systems, 2007.

[2] YN. Han, SY. Liu, and L. Qin, "Grid Technology and its Application", IT Age, (1). pp.74-78, 2006.

[3] JH. Li, and YY. Jiao, "On the Application of Grid Technology in Information Resources Management", Information Science. Vol 21 (7), pp. 776-780, 2003.

[4] WF. Ma, and XY. Du, "Research on Knowledge Grid", Library and Information Service. Vol 51(10), pp.64-67, 135, 2007.

[5] Zhuge H. The Knowledge Grid. Singapore：World Scientific Publishing Co，2004.

[6] SY. Zhou, "The New Development of Knowledge Management -- Knowledge Grid", Library Tribune. Vol. 26(1), pp. 235-237, 2006.

[7] JW. Chu, SS. Zhou, and CX. Guo, "Knowledge Grid: the new driven force of transform for knowledge management", Science Research Management, Vol.27 (3), pp.55-60, 2006.

Study on Network Interaction of Library Information Services

Junping QIU[1], Feng MA[2]
School of Information Management, Wuhan University, Wuhan, China
Email: jpqiu@whu.edu.cn, mafeng.920@gmail.com

Abstract: Network interaction has increasingly become an important way for communication. In order to meet the needs of users, modern library should improve the idea of service and develop innovative services system and ways. There are important theoretical and practical significances for adopting Web2.0 information services into modern libraries. Backgrounds and realities for interactive information services based on network of Chinese libraries are elaborated in the beginning. Then, we analyze the classic network interactive process of library information service. Finally, we probe into applications of principles and practices of Web 2.0 in interactive network information services in modern domestic libraries.

Keywords: information service; network interaction; Web 2.0 mode; library and user

1 Introduction

The starting point and end-result of library work is to provide good service to users. Library service work essentially is the process of interaction between library and user. With the wide use of network, more and more people exchange information through network. Network environment has exerted great influence on people's traditional behavior of utilizing physical libraries, and also deeply affected library services [1]. Modern libraries have gone beyond physical limits of traditional libraries. Users can locate materials and download large amounts of information easily and fast from modern libraries by many tools. Network-based interaction increasingly becomes an important part of social interaction. It also has gradually become an important way of communication between the library and user. With the change of the way of information exchange between library and user under network environment, the research on network interaction between library and user are also strengthened among libraries. Libraries need to meet the needs of users on information demand through network interaction, in order to make library network services can be carried out more effectively to satisfy the users.

2 Overview of Network Interaction in Library Information Service

Network has produced profound effect on all aspects of the social life. Library information service also has changed a lot under the network environment. Especially, library information services take on a whole new look after the emergence of Web. Interactive environment between library and user has undergone a fundamental change since the emergence of Web 2.0. The original traditional ways of interaction are being broken. For example, one-to-one and face-to-face patterns are replaced by one-to-many, many-to-one and other more complex interactive patterns. Network interaction is not constrained in time and space with unique presentation process. Some main characteristics of network interaction are virtuality, openness, diversity and unconstraint from time and space.

Virtual network space is the foundation of network interaction between library and user, in which everyone communicate with each other as different identities. Network does not exterminate substantive relation between library and the user, but extend this relationship and expand the interaction space. Under network environment, libraries and users can freely convert between actual and virtual interaction environment. Web undoubtedly provides a new and better technology platform for information, emotions and thoughts interaction between library and user. Autonomy and initiative of users are greatly enhanced. Especially with interactive the development and application of Web 2.0 technologies, interactive communications among libraries and users are further improved. Sharing and transmission model of resource between library and user changed over the single supply - demand model in traditional libraries. In short, from the perspective of information services, based resources for library information services transformed from physical collections to entity collections combined with the virtual collections. The services provided by libraries also transformed from literature services to knowledge services gradually. Besides, web-based interactive communication between library and user can maximize economy and social utility of library resources, for example, user satisfaction, economic and social benefit of library. In order to meet the needs of users, modern libraries should improve the idea of service and develop innovative services systems and ways. Network interactions are increasingly become an important interaction among libraries.

Supported by "the Fundamental Research Funds for the Central Universities".

2.1 Network Interaction

It mainly refers to interaction based on network technology. It is essentially a kind of network technology intermediary and indirect interaction. Network interaction differentiates from direct interpersonal interaction. The main interactive ways are as follows. According to timeliness of interaction, there are real-time and delayed interactions. According to the theme of interaction, there are focused and non-focused interactions. According to the demand of object, network interaction can be classified into thought interaction, emotional interaction, information interaction, and so on. According to the result of interaction, there are cooperative and conflict interaction. According to the relationship between objects and subjects, there are interactive mode and declaring-reading mode. According to the initiative of interactions, there are push type, lead-pulled type and push-pulled type. According to the service scope, network interaction can be divided into open, half-open and open interaction, etc [2].

2.2 Network Interaction in Library Information Service

It refers to method and process of dynamic information exchange and behavioral contacts between the library and user. In network interaction, mutual connection, influence and interdependent between library and user are conducted by network technology. Based on this, there are also mutual psychology and behavior change in network interaction. Network interaction between library and user consists of four aspects. First is that the interaction are conducted between library and user. Second is that the content and mechanisms of interaction function are information exchanges and behavioral effects. Third is that the results of interaction including psychology and behavior change of library and users, and so on. The last is that interaction situation is comprehensive, for example, subjective and objective conditions, network environment are included. Consequently, library and user network interaction actually contains both a series of processes and results. Processes refer to the information and behavior exchanges, both of which are ways of communication. Results refer to the psychology and the behavior changes based on communication.

3 Construction of Interactive Network Communication Pattern in Library Information Service

Transformation on communication of modern libraries needs some certain platforms. The development of library information service work is accompanied by the development of information technology. Each transformation of information technology brings new changes on library information service work. Web 2.0 is just a new information service pattern, which is different from any other Internet service patterns ever before. Web 2.0 has flourished in the global after the concept of Web2.0 proposed in 2001. Nowadays, information service mode based on Web 2.0 has been applied to many fields. There are many definitions of Web 2.0, which was not uniform. There is no a unified recognized authority definition for Web 2.0. Usually, Web 2.0 is a generic name for many social applications and technologies, including blog, wiki, tag, etc. Web 2.0 is a new generation mode with Blog, Tag, SNS, RSS, Wiki, and other applications as the core, based on many new theories and technologies, for example, six degrees separate theory, XML, ajax and other technologies. But various Web 2.0 technologies had already existed before the concept of Web 2.0 concept was proposed. The concept of Web2.0 made a summary for these already produced and increasingly matured Web information services [3]. Web 2.0 shows a strong characteristic of user-focus and counter centralization. Traditional information services regard fundamental resources as the center. Libraries serve customers according to the resources. Web2.0 information service pattern focuses on users. It emphasizes role, experiences and feelings of users. Information demand and utilizing habits of users are also focused on. Information resources construction is conducted according to the requirement of users.

The ideas advocated in Web 2.0 are as follows. First is to encourage participation and contribution of user. Secondly, communications between users are advocated to strengthen. Third is that data are organized according to user-centered principle.

Application of Web 2.0 information service pattern to the library services and information services has important theoretical and practical significances. Firstly, Web 2.0 information service pattern improves the library information service efficiency by changing the way of library service and information service. Secondly, it is favorable toward improving customer satisfaction of library information service. Thirdly, it is beneficial for research environment construction for the information service workers, etc.

Network interaction between library and user is based on network environment. Changes of network environment will inevitably have impact on interaction. At present the network society is in the primary stage of transition from Web l.0 to Web 2.0. Modern libraries need to make some changes to keep pace with social development [4].

3.1 Main Process of Network-based Interaction of Library Information Service

There are many social software and technologies included in Web 2.0. No matter which kind of way of interactive communication is adopted library information service, the main processes are as follows [5].

1)Construction of network information resources

With the development of society, requirements on content of library information interaction also get raised all

978-1-935068-76-1 © 2011 SciRes.

the time. Libraries no longer simply provide literature services, but have to provide various kinds of information with rich content, for example, multimedia services of information with different carriers, information services of academic frontiers provided through the network, etc.

2) Knowing reader's requirements

There are many methods to understand the needs of readers. In addition to some traditional methods, many other kinds of methods are based on network. In general, there are mainly three types of method. First are traditional survey methods, for example, questionnaire, field investigation, analysis on materials about reader and information service, sampling survey, comparison method, etc. Questionnaire investigation method may be the most used method. Besides, online readers survey method and readers' behavior information analysis methods are also adopted by many libraries. Readers' information behavior analysis mainly refers to obtain relevant data about readers' information needs by browsing history and bookmarks of Internet readers, or readers visit record in server logs, etc. This method can help to collect and track individual information demand and interest of readers.

3) Reader relationship management

Communication between library and readers is an information interaction process. It is related to both internal organizational structure and costs involved to sustain the interactions. Reader relationship management uses information technology to classify and aggregate the readers' opinions and requirements, final result of which can become the reliable basis for decision-making. Reader relationship management provides interactive communication for organizing and users, such as online conversation, navigate together, E-mail, etc.

The network interaction effect of library information service is mainly embodied in the third process.

3.2 Main Forms of Network-based Interaction of Library Information Service

Throughout most of the form of network-based interaction adopted, network service of library information has experienced Email, BBS and ICQ in the past few years. Now, blog, wiki and other forms of network communication means gradually become the main ways of interaction. Generally, there are four types of Web 2.0 technology adopted in Chinese libraries: (1) Web 2.0 technologies used on the library websites, including personalized home page service, blog, RSS (Really Simple Syndication) information aggregation, wiki, SNS (Social Network Service,) and so on. (2) Web 2.0 technologies used to embedding services into the users' information technology environment. With this kind of Web 2.0 technology, a particular library service could be embedded into the user's computer desktop, network environments or personal cell phone terminal, for example, RSS information push, browser tool bar, phone library, etc. (3) Web 2.0 technologies used in the reference services, such as QQ, MSN, blog, and so on. (4) Web 2.0 technologies used on the subject service [6].

E-mail is the most used service for communication. The message or information transmitted by E-mail can be with rich content and different forms. E-mail also has advantages on fast transition speed, low cost and strong secrecy, etc. Therefore, many libraries often use E-mail for bibliographic booking, inter-library loan, information pushing, reference consultation, etc.

ICQ (I Seek You) is also used popularly. It is a kind of instant message transmission software developed by Mirabilis Company in Israel in 1996. It has functions in transmitting and sending information in words, speech information, files of chatting real-timely, and it allow users to detect networking state of his/her friends. ICQ have stronger immediacy than E-mail. Libraries mainly use ICQ for real-time interaction and reference consultation.

BBS (Bulletin Board system) also had been popular for some time. On BBS, everyone could participate a discussion about issues they interested. It helps people to ask for answers from others by post message or comments, but also it provides real-time communication for people. BBS has been widely used for its real-time characteristic, highly liquidity, openness, and other characteristics, which are consistent with the library's request for developing the interactive reference services, including book reservation, renewing, recommendation, information browsing service, information bulletin board, bulletin board of reader comments, etc.

Personalized service homepages are a special type of page, which are constructed to meet the needs of the user's personalized access by a library with a public Web page as well. Readers can changes, drag and drop, moving the service module randomly according to their own habits on this kind of page. They could arrangement the columns on the page or move the position of column by personal interests. For example, personalized service homepage of Tsinghua University library is including some parts, such as introduction, commonly used system, notice, e-resource dynamics, information searching, electronic resources, services guide, and so on. The user can delete, move, fold, unfold and other operations in order to form their own favorite pages.

Blog has now become more and more popular. There are three main ways of presentation of blog. First is hosting blog. Libraries don't need register domain name, rent space and compiled page. Though this kind of blog is simple and easy to construct, it is hard to satisfy requirements of establishing library blog site and do not facilitate for unified management. In addition, the stability and speed of them cannot meet the requirements usually. Second is affiliated blog, refers to library blog used as a website or part of an address. Third is blog in self-built independent website. This kind of blog has its own do-

main name, space and page style. Under normal circumstances, a university library has its own website, so the university could use self-built independent website to build a blog center, which has its own characteristics, in order to develop the network interactive services with more autonomy. Blog-based network interactive service is now a main way of interactive library information service. There are many personal blogs created by people in the community of Library and Information Science (LIS), however, blogs opened by a whole library and with the name of a library are few [7].

RSS (Really Simple Syndication) is a kind of content delivery and integration Internet technology, based on XML (X Extensible Markup Language). It is used as a new way for information provision, with three major advantages of information aggregation, information pushing and information filtering. RSS-based information aggregation is an important and widely used technology and way among LIS. There are three main processes. First, information sources, i.e., professional academic library web sites, academic research blogs and other, are collected, filtered, sorted, classified and revealed by workers in a library in the way of RSS. Then, a variety of aggregation tools are used to construct browsing academic resources system. Lastly, the library provide special information services in the library information portal to help researchers quickly and efficiently track the latest developments in related fields [8]. For example, Shanghai University library provides a RSS-based news aggregator systems (LIS page), the main contents of which are about LIS field.

Wiki is a Web-based tool supporting for some persons to create and exchange collaboratively. It provides a platform for more than one person to input, upload and publish information. Based on this, a knowledge network system could be built to support the sharing of domain knowledge within a community.

SNS (Social Networking service) is a kind of platform for information publishing, communication and sharing for the users. It is focused on customer relationship management services form [9]. Compared with other Web 2.0 tools, SNS is most characteristic in the virtual reproduction of the true relationship network between people in order to help them to build relationships network [10]. For example, University of Science and Technology of China library has applied this kind of Web 2.0 technology.

Web 2.0 has brought much new enlightenment and the applied direction for library. For example, it expedites the communication channel between library and users, and expands library service range, but there is no a mode can replace any other kind of service mode. Therefore, libraries should pay attention to mutual confluence of blog and other interactive service modes. They should adopt various ways of network interaction.

References

[1] Guo Jing, "Current Status and Analysis On Online Reference Services of Domestic Universities Libraries," Library Journal. issue 9, pp.55-58, 2002.

[2] Gu Jin and Zou Haibo, "A Tentative Discussion on Information Interactive Services for the Academic Libraries under Net Environment," Journal of Chang Chun Teachers College (Natural Science), vol.24, issue 5, pp.155-158, 2005.

[3] Si Jiaojiao, "Study on Network Interaction Mechanism Between Library and the User," Master Dissertation, Fujian Normal University, 2009.

[4] Wu Chao, "On the Application of Web 2.0 in Information Services," Modern Information, issue 8, pp.63-64,68, 2006.

[5] Shi Fenglan, "Discussion on Interactive Communication Pattern in Information Service of Digital Libraries," XINKECHENG, issue 8, pp.94-95, 2010.

[6] Ma Aifang and Zhao Zhiying, "Application Status of Classified Web 2.0 Technologies in University Libraries in China", Information Studies: Theory & Application, vol 33, issue 8, pp.91-95, 2010.

[7] Hu Shuli, "Blog: A New Way of Network Interactive Services for University Library," Modern Information, issue 10, pp.141-143, 2007.

[8] Lou Xiuming and Ding Pengyu, "Analysis on Application Status of RSS Technologies in Shanghai Libraries", Library Journal, vol 28, issue 2, pp.53-57, 2009.

[9] Qin Tiehui & Sang Xiaoqi, "A Study on Characteristics of Social Network Services and its Impact of Competitive Intelligence", Document, Information & Knowledge, issue 6, pp. 11-15, 2007.

[10] Library 2.0 Studio, "Library 2.0: Upgrade Your Services", Beijing: Beijing Library Press, pp.194, 2008.

978-1-935068-76-1 © 2011 SciRes.

Data Mining Workbench and Its Application in Design Archives Management and Position System

Tao CHEN, Zhanli FENG, Dunru REN, Liqiu CHANG, Qi YUAN
Design and Research Institute of Anshan Iron and Steel Group Corporation
Anshan City, Liaoning Province, China
Email: qq5547661@yahoo.cn

Abstract: Data mining is a whole process, including data acquisition and processing, using algorithms and application of the results for decision-making. In these sections, data acquisition and processing is the core of data mining. In the data acquisition and processing section, It is an improvement from manual process to application of software tool. But application of data mining workbench is another new leap. This paper discusses the architecture and the technology characteristic of data mining workbench and the application example of design archives management and position system for data mining workbench

Keywords: workbench; software development; design Archives; barcode

1 Introduction

"Data Mining is a valuable process that obtain valid, novel, potentially useful and ultimately understandable patterns from a lot of data in a database, a data warehouse or other repositories." With the further development of data mining, many data mining software tools continue to emerge. There are tools of Data mining for specific problem areas or common problem areas. For example, IBM's AdvancedScout system; SKICAT systems of California Technology Institute; QUEST system of IBM Almaden Research Center; MineSet systems of SGI and so on.

We believe that data mining is a whole process, including data acquisition and processing, using algorithms and application of the results for decision-making. In these sections, data acquisition and processing is the core of data mining. "Handy not as good as tool", in the data acquisition and processing section, It is an improvement from manual process to application of software tool. But application of data mining workbench is another new leap. This paper discusses the structure and technical characteristics of data mining workbench that is used in data collection and processing section for the direct application. Combining with design Archives management and location system (i.e. lending DB and archives DB is combined. The groups is found that often use the design Archives. Service strategy and storage solutions are produced), we describe a specific application of data mining workbench.

2 The Characteristics of Workbench

Data mining workbench has the following technical characteristics:

- The "manual labor" of data acquisition and processing in Data Mining approach into a "mechanical" or "automatic" mode.
- Common software tools and fully functional components is integrated in data mining workbench. These tools and components is compatible with data in the database, and have a good interface with database.
- It is visualization operation, direct application- oriented issues, transparency, intuitive and easy to use.
- Improving efficiency of data collection and processing, reducing errors in data entry and processing.
- Processing data by batchwise, orderly and Normatively.
- The components and a variety of tools are integrated in a platform. It is easy to use for non-professional.

3 The Structure of Workbench

The structure of data mining workbench is shown in Fig 1. The workbench consists of six parts. That is the working database, data import parts, configuration management, data processing part , the splitting theme part and the algorithm platform.

3.1 Working Database

Working database is the basis of the workbench and all work is carried out in the working database. Working Database include the buffer database and the results database. There are various types of databases in Practical application of information management systems, such as SQL server, Sybase, Informix, VF, MDB, Oracle and so on. The MDB database has good compatibility and flexibility. In order to unify data format, the Working database select MDB database. Various operations is made in the buffer database. The final results of the data are stored in the results database.

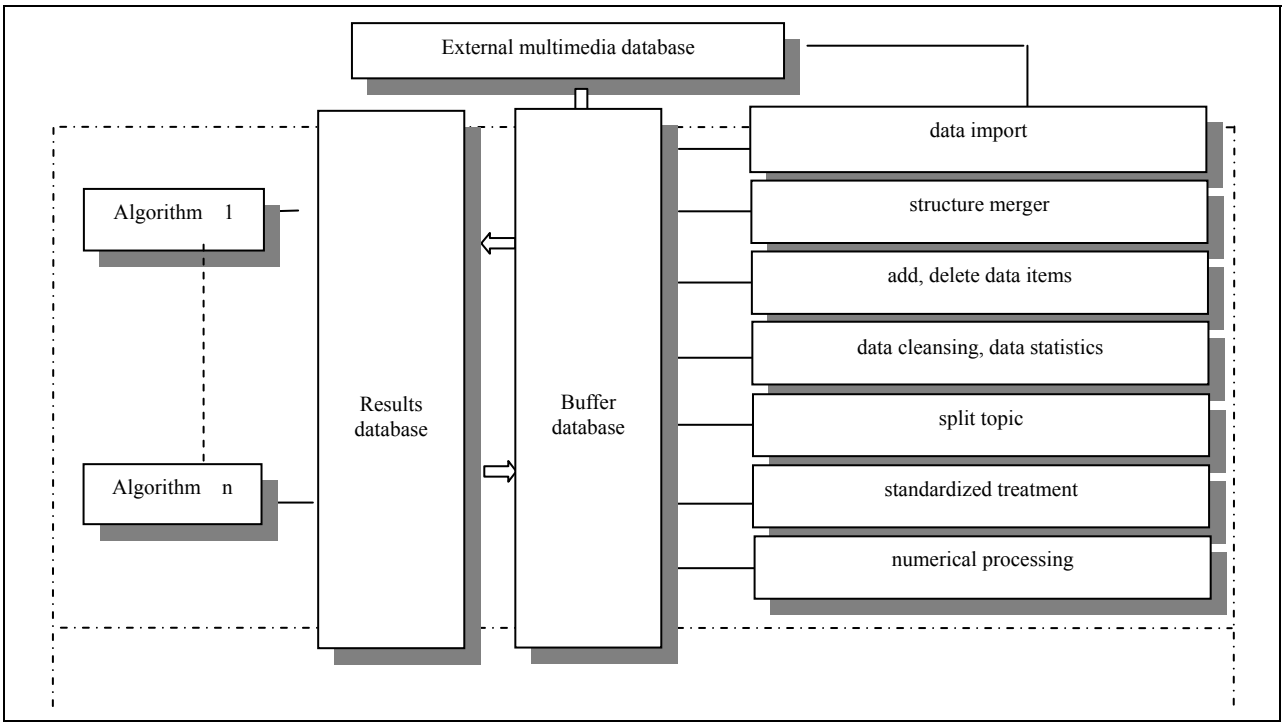


Figure 1. The structure of data mining workbench

3.2 Data Import Parts

Data import parts complete data import that the data is converted from kinds of different structures database into MDB database. The data is stored in the buffer database to prepare for its operation and the necessary treatment

3.3 Structure Management Parts

Structure management complete the processing of data fields in the buffer database, including structural consolidation, add data items, delete data items and other operations data items.

3.4 Data Processing

In the data processing part, We make the implementation of data cleansing, statistics, standardized treatment, numerical processing.

3.5 Splitting Topic Part

Splitting topic part complete the extraction of some key words in the text, and split into different effective fields.

3.6 Algorithm Component Platform

There are kinds of common algorithm in the algorithm Component Platform. such as clustering, association rules, Classification, characteristics pick-up, text abstract, clustering and Classification of text.

4 An Application Example of Design Archives Management and Positioning System

With the rapid development of computer, network and database technology, information systems has penetrated into all spheres of society. In these systems, a huge amount of data is stored. It is a very important issue How to make use of these data to provide decision support for enterprise decision-makers. In addition to using the existing query system that get some visual information, The people have to mined substantial, undiscovered and subsistent data relationships. This section discusses an example of design archives management and positioning systems of Design and Research Institute of Anshan Iron and Steel Group Corporation for the applications of data mining workbench.

4.1 Background Information

Archives department of Design and Research Institute of Anshan Iron and Steel Group Corporation is a operating center in enterprise engineering Design., Sets of files are saved in the archives department. Including research reports, budget estimate, project plan, preliminary design, calculations, contract documents, construction drawings and so on. Based on the characteristics of engineering design, Old files of 40% to 80% are used in new design project. The reuse rate is extremely high. With increase

978-1-935068-76-1 © 2011 SciRes.

in the number of the current file, there is a problem of shortage of warehouse capacity. It is an important research topic how to properly manage and use the design files, both to meet the designers to quickly find and to make good use of storage space.

4.2 Implementation Process by Workbench

The main menu of Data mining workbench is shown in Figure 2.

Step 1. In the figure 2, select term 1. Then selecting database name, the all fields are exported from the archive lending database and multimedia database. The DB1.mdb and DB2.mdb database is formed in the buffer database.

Step 2. In the figure 2, select term 2. The redundant fields are removed from DB1.mdb and DB2.mdb database. Combining DB1.mdb with DB2.mdb,the HDB.mdb database is formed in the buffer database.

Step 3. In the figure 2, select term 3. The irrelevant or weakly related fields are deleted in the HDB.mdb database. (borrowing card number, file number, etc.). Some useful fields are added (loan number, loan time, keywords, etc.)

Step 4. In the figure 2, select term 5 and term 3. According to project characteristics, split name fields, extract keywords and add new fields.

Step 5. In the figure 2, select term 6 and term 7. Call the appropriate modules for data cleaning. When the volume does not borrow, the record is deleted on the database. Non-standard word is uniformed. such as the "three burning project", "the third sintering machine project", "three burning renovation project" that is exchanged into the "three burning projects". Finally, the numerical treatment of the field is made.

Step 6. Call cluster analysis and association rules for formation of the results.

4.3 Interpretation and Use of Results

Result 1. Through association rules analysis, we found an association of "computer and civil engineering". This is because the design of computer network system requires the plans of civil engineering as reference diagram.

Result 2. Through cluster analysis, we found design files to commonly used. The design files That are not often used is stored in the other storage densely and compressively to year as the threshold value. In order to file management and position, We configure location identification and the barcode in each file of multimedia database. Shown in Fig 3.

Result 3. Cluster analysis found the staffs that often use archives are concentrated in metallurgy equipment department, electrical department and civil engineering department. It is necessary that to study special service.

Figure 2. Main menu of Data mining workbench

Figure 3. The barcode of management and position

5 Conclusion

The real example of design archives management and location system in Design and Research Institute of Anshan Iron and Steel Group Corporation shows that using the workbench can make the data mining process of

mechanization, automation, intelligent and can greatly improve efficiency.

Acknowledgments

We acknowledge support from science and technology department of design and research institute. Special thanks Ms. Hui feng-ju and Ms. Yie ying-chun for their help and advice.

References

[1] Mao Guojun, Duan Lijuan, “data mining principles and algorithms,” Tsinghua University Press, 2007.
[2] Xu Lanfang, “database design and implementation” Shanghai Jiaotong University Press, 2006.
[3] Liu Tianhui, “Visual Basic Programming Guide”, Tsinghua University Press, 2006.

On Construction of Regional Industry Competitive Intelligence System Based on Triple Helix Innovation Theory

Min SHI[1], Jian LUO[2], Xuekui XIAO[3]

[1]*Hunan Science and Technology Information Institute, Changsha, China, 410001*

[2]*College of Business, Hunan Agricultural University, Changsha, China, 410128*

[3]*Hunan Science and Technology Information Institute, Changsha, China, 410001*

Email: 631725@qq.com, luojianwork@gmail.com, sm8@21cn.com

Abstract: With competition increasing, a growing number of local governments are interested in establishing regional industry competitive intelligence system to promote regional innovation. At first, the Triple Helix Innovation theory and its adaptability to regional industry competitive intelligence system are analyzed. Then, regional industry competitive intelligence system is defined. Next, functions and elements of the system are analyzed. And at last, the model of the system is constructed to provide a theoretical model for construction of regional industry competitive intelligence system in China.

Keywords: Regional innovation; Triple Helix Innovation theory; Industry competitive intelligence; Industry competitive intelligence system

1 Introduction

Our government and the academic circles have paid much attention to promotion function on the development of economy and industry of regional competitive intelligence system in developed countries. At present, some regions in our country have begun to explore ways to construct regional construction industry competitive intelligence system, but rare theoretical research on this area has been conducted. Regional economic development is mainly derived from the regional innovation and new research shows that innovation system in China has applied the Triple Helix model. Therefore, this article is intended to research into construction model of regional industry competitive intelligence system based on Triple Helix Innovation Theory.

2 Analysis on the Triple Helix Innovation Theory and Its Adaptability to Regional Industry Competitive Intelligence System

2.1 The Triple Helix Innovation Theory

In 1995, the Triple Helix Innovation theory was proposed by Henry Etzkowitz & Loet Leydesdorff. They applied three-helix philosophy to explain the complex interactions taking place among different levels and innovation subjects. They thought that universities, industries and governments can organize, host and participate in innovation. They co-edited "University and the Global Knowledge Economy: the Triple Helix of governments, university and industries".

According to the Triple Helix theory proposed by Henry Etzkowitz & Loet Leydesdorff , there are three relation models between governments, industries and universities, generally referred to as "national socialist model", "laissez-faire model" and "Triple Helix model" [1]. These three models can be distinguished because of the huge differences between each body's relationship and status. Three of them merge together and in the mixed overlap system. They can play each other's role. Therefore, hybrid and mixed organizations are founded during the interaction of these three subjects. Incompleteness of the system is maintained for flexible transformation. The Triple Helix model is shown in Figure 1.

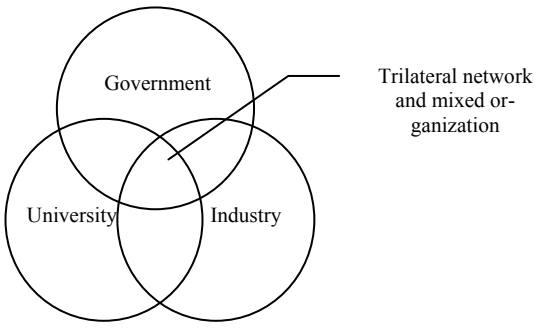


Figure 1. The Triple Helix Innovation Model

2.2 Adaptability of the Triple Helix Innovation Theory to Regional Industry Competitive Intelligence System

Firstly, one of the most important functions of regional

National Natural Science Foundation Project. (*70972119*); Hunan Natural Science Foundation Project, "Morden Competitive Intelligence Theory and Methodology" (*2007JJ5113*); Hunan Science and Technology Infrastructure Platform Project, "Competitive Intelligence Platform in Hunan Province" (*2007TP2002*)

industry competitive intelligence system is to promote regional innovation. For example, competent intelligence in Japan covers government, industry and academia, which could hold technology and management trends in other countries timely and develop information sharing of government, industry and academia by their own country to promote innovation ability.

And then regional industry competitive intelligence system could develop the interactive of industry and research institution. The competitive intelligence service supports decision-making for regional innovation. The government could grasp industry development and make policy effectively by this system. Companies could gain regional industry policy, tendency information and foster research ability efficiently. Research institution could obtain industry development and companies' innovation demands. Therefore, regional industry competitive intelligence system could advance the triple helix innovation developed to the high-level.

Endingly, regional industry competitive intelligence system fits the characteristics of trilaterally mixed organization. The junction of the three main subjects form the Trilaterally mixed organization, which is called interface organization by Yingdong Pan and Da Ying professor. Trilaterally mixed organization can be launched spontaneously and cooperatively. Service targets include three main subjects or even the whole society. The theory promotes the widespread cooperation and innovation in the Triple Helix, facilitates knowledge sharing and knowledge applications in a great extent, and further promotes industry development. The performance organizations include consulting institutions, associated agency and intermediary agent. So the regional industry competitive intelligence system should be taken as trilaterally mixed organization to be researched in.

In conclusion, the regional industry competitive intelligence system promotes the technology innovation of regional industry. The regional industry competitive intelligence should be studied as the regional innovation system according to its characteristics. Research into the regional industry competitive intelligence based on the Triple Helix theory is not only consistent with our current model, but also maximizes efficiency when integrating it into the regional innovation system at the beginning of the design.

3 The Position and Function of Regional Industry Competitive Intelligence System

3.1 The Position of Regional Industry Competitive Intelligence System

The regional industry competitive intelligence system exists in the form of trilaterally mixed organization in the triple helix innovation system. According to Pan Donghua and other scholars' analysis on the role played by the mixed organization in the triple helix innovation system, the triple helix of industry, government and university possesses different social functions and attributes, it has different organizational goals and operative methods, which determines triple helix's different roles and functions in the innovation process as well as the heterogeneity of its knowledge entropy [2]. The demand for innovation & cooperation and the drive by market interests highlight the heterogeneity of knowledge entropy, and this highlighted heterogeneity further promotes the flow of knowledge in the triple helix.

Mixed organization is an important carrier of knowledge flows and entropy increase. Meanwhile, in order to achieve the overall negative entropy value, we must continue knowledge exchange and integration of learning both inside and outside the triple helix, and the process is realized by the mixed organization, an important carrier. The trilaterally mixed organization strengthened connectivity of innovation network and promoted more loop formation. The production and operation of this organization reduced the transaction costs in the industry innovation process, promoted the emergence of balance between competition and cooperation, and boosted the choice and decision making of technical solutions to this social capital. As a carrier and center of knowledge and resources, mixed organization, with the process of organization and coordination, resolves the conflicts in innovation and completed the selection and decision-making of the optimal technical solution to social capital.

As a result, the regional industry competitive intelligence system, as a typical trilaterally mixed organization, should be positioned according to the role of mixed organization. Therefore, regional industry competitive intelligence system aims to center on regional innovation, promote transfer of heterogeneous knowledge among innovation subjects, and form a stable knowledge flow model[3].

Of course, if the regional industry competitive intelligence system wants to give play to its information function and turns exogenous drive into endogenous one[4], it must break the limitations of the region and conduct full information exchange with external environment. Therefore, the regional industry competitive intelligence system not only is a mixed organization of the triple helix regional innovation model but also introduces exogenous knowledge of regional innovation. The role of the regional industry competitive intelligence system is shown in Figure 2.

3.2 The Function of Regional Industry Competitive Intelligence System

According to the position research in triple helix model, regional industry competitive intelligence system in China, as a trilaterally mixed organization to promote regional innovation, is promote information exchange between government, industry and university and develop the evolution of triple helix innovation model. At the same time, as a part of regional innovation system, com-

978-1-935068-76-1 © 2011 SciRes.

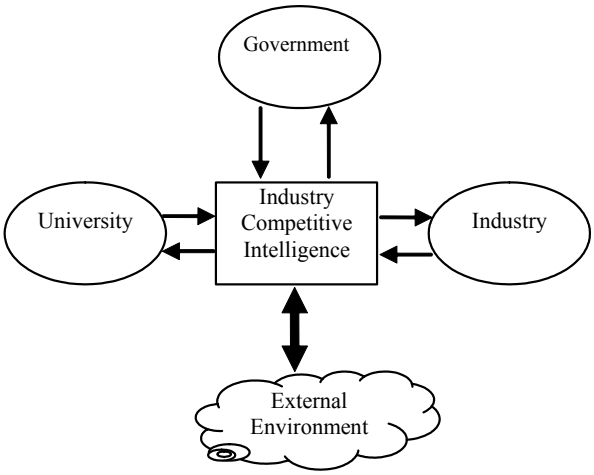


Figure 2. The role of industry competitive intelligence in the regional innovation model

petitive intelligence system helps to complete information exchange between regional innovation system and external environment.

1) The information function of suppoting government' industry decision-making

The government's decision-making in industry is the cornerstone of the regional industry development and a catalyst of regional industry innovation. Especially in China, government decision-making can directly promote the industry. Therefore, it is essential to the development of regional industry that whether the government's decision-making on industry is correct or not. Although current government's decision-making capability in China has been improved to some extent after many years of development, but at this stage, it is also difficult to make decisions influenced by an increasing number of factors. Government needs an industry intelligence support agency to provide regional industry decision support timely and accurately. Therefore, to provide information support government's decision-making on industry is an important guarantee to promote regional innovation.

2) The function of Improving enterprise competitive intelligence system

Currently, although a few enterprises have established their own competitive intelligence system, little support of industry competitive intelligence when the strategic decision-making. Therefore, the establishment of regional industry competitive intelligence system can carry out industry research and competitive intelligence services to meet the common needs of the industries within the region, while also achieves the purpose of improving the existing competitive intelligence systems. For those small and medium enterprises which are unable to construct their own competitive intelligence systems, they can directly take use of regional industry competitive intelligence systems to collect and analyze information and quickly lift their capabilities in intelligence applications.

3) The guiding function on research direction in university

The research capacity of university promotes regional innovation. These problems, like how to understand the industrial development purpose of government and the technological innovation demand of enterprises and how to industrialize the scientific research achievements， are important to university's integration into the regional innovation system . At this point, as one kind of the triple helix innovation model, the regional industrial competitive intelligence system should play the bridge function. Therefore, the regional industrial competitive intelligence system must have the guiding functions on research direction in university in order to accelerate the scientific research achievements transformation and advance the regional innovation capacity.

4) Continuous monitoring function of the regional industry

Besides its endogenous service and information exchange function, regional industry competitive intelligence system also has the function of collecting and analyzing information from external environment. Regional industry competitive intelligence, as a triple helix model, both spread generalized industry knowledge within the region and serves as a starting point to obtain generalized industry knowledge from the external environment. Long-term and continuous monitoring of the related information inside and outside China on key industries within the region is the basic requirements to ensure results of the regional industry competitive intelligence system.

4 Analysis on Elements Composing the Regional Industry Competitive Intelligence System

Functions of the system need to be supported by corresponding elements. Regional industry competitive intelligence system has similar information gathering function with enterprise competitive intelligence system. But the former, as the third-party intelligence service agency, has essential distinction compared with the later. Therefore, when designing service elements, it is necessary to consider not only transmission tunnels of different targets, but also service forms and contents of different service subjects. Functional elements should be segmented according to functional requirements, as shown in Table 1.

Therefore, elements of the regional industry competitive intelligence system include information gathering element, information analysis element, intelligence service element, and construction platform element.

4.1 Information Gathering Element is the Basic one of Industry Competitive Intelligence System

There is no difference between industry competitive intelligence and enterprise competitive intelligence in the way of gathering information but in contents and focuses

Table 1. Segmented functional elements of regional industry competitive intelligence

	Information Collection Elements	*Information Analysis Elements*	*Information Service Elements*	*Construction Platform Elements*
Government Industry Decision-making nfor-mation Support Function	•	•	•	•
Enterprise Competition Improvement Informa-tion System Function	•	•	•	•
University Scientific Research Direction Guidance Function	•	•	•	•
Regional Key Industries Continuous Monitoring Function	•	•		

of collecting information. Contents of industry competitive intelligence collection are more macro than that of enterprise competitive intelligence. It focuses on government decision-making requirements, supplements inadequacies of enterprise competitive intelligence system, as well as directs scientific research in colleges and universities.

4.2 Information Analysis Element

Information analysis is conducted based on the needs of clients. Generally, competitive intelligence collects fragmented and relatively large number of information that can not be directly served as contents of services for clients. According to the demand and monitoring program of competitive intelligence, in order to ensure the quality of service, it is a key to process the collected information.

4.3 Intelligence Service Element

Regional industry competitive intelligence system, as a mixed organization, provides services for government, industry and university as the third party, and the service function can promote the service efficiency. Regional industry competitive intelligence system is characterized by universal competitive intelligence function and targeted decision support function, so services should be divided into different forms of services like regular services, advisory services etc. Intelligence service element is the ultimate expression of each function, only can good intelligence service be done; the effectiveness of regional industry competitive intelligence system can be demonstrated.

4.4 Construction Platform Element

Although competitive intelligence system is a trilaterally mixed organization, according to overseas experience and our present situation, the regional industry competitive intelligence system had better take science & technology sector of the government as the leading sector and region's nonprofit organizations as principal construction agencies. In this way, the system can serve the government and enterprise as well as direct scientific research of universities.

5 Model Construction of Regional Industry Competitive Intelligence System

Even regional industry competitive intelligence system is independent existence, the information collection and analysis of it are similar with company intelligence system, which are absolutely necessary to the establishment of industry competitive intelligence system. But the function implement and service content are different from company intelligence system. Firstly, in service content, the system should guide the technology exploration. Secondly, the service objects include government, universities and company (but not company employees). Thirdly, the service forms include supporting information monitoring and decision making. So the function would not be realized just by one intelligence service system. According to the analysis of elements of regional industry competitive intelligence system, the most important work is the appropriate installation of the platform subjects. Now the technology & science government system, which is the unique appropriate organization to integrate industry and technology resources, could establish the regional industry competitive intelligence system to promote industry policy making, guide the research institution technology exploration. Therefore, the system model which is in the charge of technology & science government system could be consist of information collection system, analysis system, service system, achievement conversion system and information technology supporting system. (see Figure 3)

Regional industry competitive intelligence system which is in the charge of technology & science government system, should be burdened with the promotion of information and knowledge flow. The external driving force could transform to internal driving via information sharing to advance the regional innovation system and economy development. During the system establishment, the most problem is constructing information source, which could be organized and integrated by technology & science information service institution. The external information gaining methods and theory could be referenced in “Study on Industrial Tracking Roadmap as Used in Competitive Intelligence” [4].

978-1-935068-76-1 © 2011 SciRes.

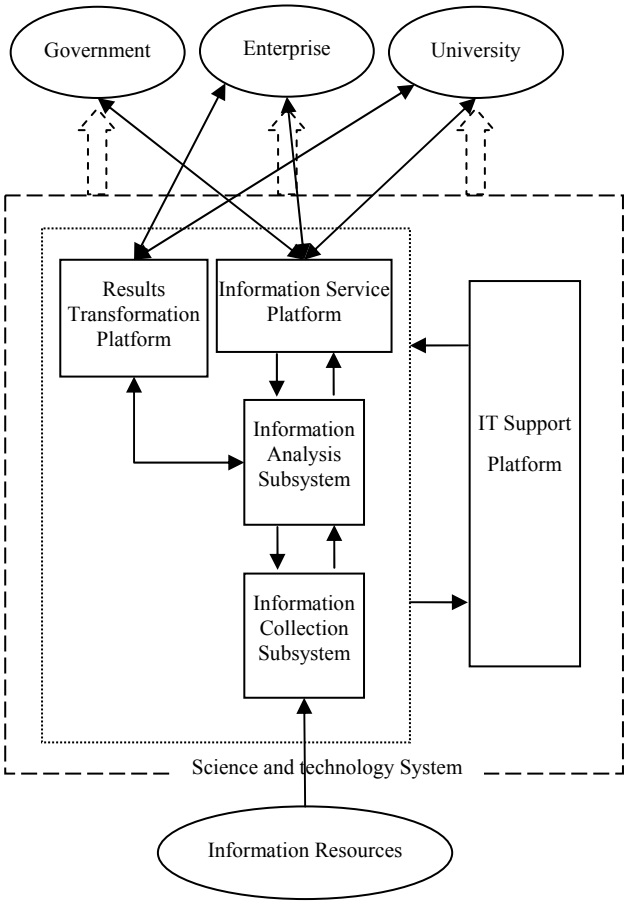


Figure 3. Regional Industry Competitive Intelligence System Model

6 Conclusion

This article first positions the regional industry competitive intelligence system based on the Triple Helix innovation theory and identifies it as the mixed organization of triple helix innovation system, and it can provide external information for the system to promote knowledge integration among the three innovation subjects. Then, its proper features and elements are analyzed according to the position. Finally, the regional industry competitive intelligence system model is constructed to provide theoretical basis for establishment in China of regional industry competitive intelligence system.

References

[1] Etzkowitz H, "The triple helix: University-industry-government innovation in action," Routledge, 2008.
[2] Pan Dong-Hua, Yin Da-Wei, "Triple helix interface organization and knowledge transfer in the innovation," Studies In Science of Science, vol. 005, pp. 1073-1079, May 2008.
[3] Sun Liping, "Transition and Fracture： Social Structural Changes in China since the Reform,": Tsinghua University Press, 2004.
[4] Shi Min, Xiao Xuekui, Study on Industrial Tracking Roadmap as Used in Competitive Intelligence, Journal of the China society for science and technical information, 2010(1): 184-192.

Scientific Research

The Service Platform for Domain-specific Resources Based on Vertical Search Technology

Cuixia MU[1,2], Hongwei HAO[1], Xucheng YIN[1], Dianyin WANG[1]

[1]*Department of Computer Science, School of Computer and Communication Engineering, University of Science and, Technology Beijing, Beijing, China, 100083*

[2]*Department of Computer Science, China Women's University, Beijing, China, 100101*

Email: nwuc_mucuixia@163.com; hhw@ies.ustb.edu.cn; xuchengyin@ustb.edu.cn; wdyl8686@126.com

Abstract: According to the distribution of domain-specific resources on the Internet, the system framework of the service platform for domain-specific resources based on vertical search technology is designed, which consists of resources domain knowledge base, domain-specific resources collection, resources processing, resources structured storage and web user interface. The system towards exam resources is implemented. Practical application shows that the system works well. It can search precisely exam resources from the Internet, and effectively process and store them. In addition, web user interface provides classification retrieval and customization information service.

Keywords: vertical search; Web crawlers; information extraction; information denoising

1 Introduction

The explosive growth of information on the Internet resulting in information overload has caused people submerged in the sea of information. But the universal search engines[1,2,3] can not fully satisfy the specific users' needs for domain-specific resources, because they have the drawbacks of complexity, inefficiency, slow update, no subject restriction, low accuracy of query, etc. As artificial intelligence technology further mature and information services diversification, the vertical search engines have become the research focus. Vertical search engines [4-7] are the domain-specific professional search engines, which reduce the search scope to increase query accuracy and improve the capacity for tracking web resources. Therefore, Vertical search engines can provide precise, efficient and all-around information service for specific-domain users.

According to "Statistical Report on Internet Development in China" published by CNNIC (China Internet Network Information Center) , up to Dec, 2010, the number of net citizens in China has reached 457 million, and the popularity rate of Internet has climbed up to 34.3%, and the number of search engines users has reached 375 million. The time and cost for users to access to effective information greatly increase with the information overload, so the accuracy and intelligence of search engines should be improved greatly.

Currently, the Web sites related with education resources are the most important platforms, where users can find what they need and communicate with each other. But it is difficult for users to easily and quickly search the complete related resources from the Internet because of the "mixed large and small settlements" distribution of exam resources. Therefore, we need an information service platform to provide complete exam resources quickly, precisely and efficiently.

In this paper, we try to construct the service platform for domain-specific resources based on vertical search technology. And the service platform towards exam resources is implemented and runs well.

2 System Framework Analysis

In order to implement the service platform for domain-specific resources based on vertical search technology, we first analyze the processing of domain-specific resources information and then describe the system framework.

2.1 Domain-specific Information Processing

Domain-specific resources information will undergo five kinds of forms, from Web pages on the Internet to that on the service platform, as in Fig. 1.

1) Web pages on the Internet

Most of domain-specific resources on the Internet exist mainly in either of two forms, and one refers to Web pages' content seen directly, and another refers to of hyperlinks' URLs through which the resources files such as text files, image files and other files on Web servers, can be downloaded.

2) Web pages or URLs locally stored

The collection subsystem of the domain-specific resources service platform gathers the Web pages or URLs related with domain-specific resources on the Internet, and then arranges and stores them locally.

3) Body text files of pages, URLs and their titles locally stored

Body text of pages can be extracted and then locally saved as text files. The Web pages' titles, URLs and their titles can also be extracted and stored in a certain file.

978-1-935068-76-1

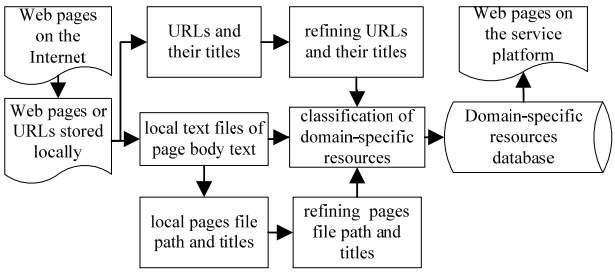


Figure 1. Domain-specific resources processing

4) Structured data in domain-specific information resources database

In order to provide efficient information service, the pages body text, URLs, titles and other identification information are stored in the domain-specific information database.

5) Web pages on the service platform

The users can search and make full use of the domain-specific resources information and the Web pages of search results can be shown on the Web user interface.

2.2 System Framework

The system framework [8,9] of the service platform for domain-specific resources contains five function modules, as shown in Fig. 2.

1) Resources domain knowledge base and Chinese word segmentation

There are a large number of domain-specific keywords, already arranged and classified, which are used to guide domain-specific resources collection, processing, storage and retrieval. Chinese word segmentation is the important step of text analysis and an assurance for other four function modules working efficiently.

2) Domain-specific resources collection based the focused crawlers

Two focused crawlers are designed and run at the same time. PRSpider designed independently which contains crawling, URL extraction, URL filter, page body text extraction, topic relevance calculation, information denoising and so on, aims to gather the newest resources regularly from some fixed famous sites. Extended Heritrix is responsible for locating and downloading domain-specific resources on a massive scale from the whole Internet.

3) Domain-specific resources processing

The service platform must effectively process Web pages downloaded from the Internet and then provide resources information service. And this module includes pages body text extraction, topic relevance calculation of body text and information denoising.

4) Structured storage of domain-specific resources

The module contains tagging, classification and storage of the domain-specific resources information. Tagging the domain-specific resources information refers to extract and arrange the identification information related

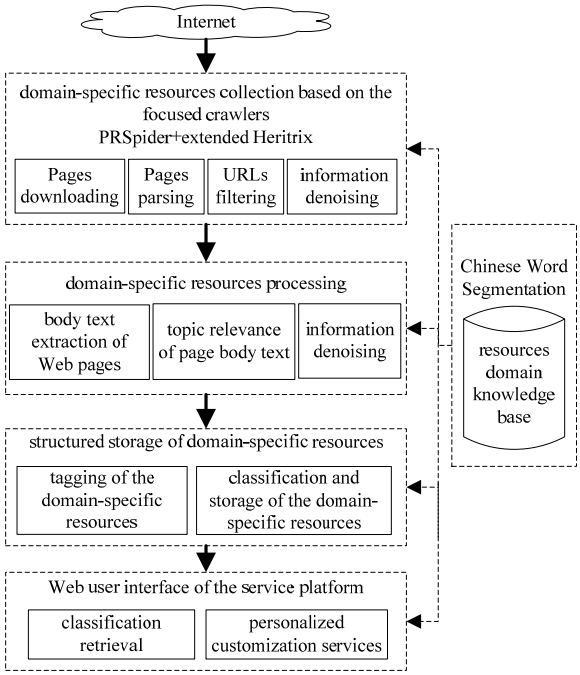


Figure 2. System framework

with domain-specific resources, such as titles, URLs and anchor text. Classification and storage of the domain-specific resources refer to store a mass of resources information after classification.

5) Web user interface of the service platform

This module includes classification retrieval and personalized customization services for domain-specific resources. Classification retrieval can reduce the retrieval scope to improve retrieval accuracy. Personalized customization services refer to provide the appropriate domain-specific resources information to particular users according their needs.

3 System Implementation

In this paper, taking example for exam resources, we design and implement the service platform for domain-specific resources based on vertical search technology on Windows XP using Java, MySQL5.1 and Tomcat5.5.

3.1 Resources Domain Knowledge Base and Chinese Word Segmentation

Chinese word segmentation is the important process of text analysis, and the necessary step to collect the domain-specific keywords. The system uses ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) developed by Institute of Computing Technology Chinese Academy of Sciences. We add these keywords in resources domain knowledge base to the user dictionary of ICTCLAS. So we can implement Chinese word segmentation which strongly supports domain-specific resources collection, process, storage and retrieval.

The domain-specific resources knowledge base can be used with keywords as the basic unit to guide the domain-specific resources collection, process, storage and retrieval, and so we emphasize on collecting and classifying domain-specific keywords to construct it. Domain-specific keywords are usually classified into four types which are specialized, internal, pages navigation and banned keywords. The following takes example for exam resources keywords.

(1) Specialized keywords

For example, related with exam resources, specialized keywords include exam resources types, exam subjects and so on, and they can effectively distinguish the domain-specific resources. In addition, there exist synonyms, acronyms, abbreviations, slang terms and different languages for some specialized keywords. So we need to extend specialized keywords, classify them and establish the corresponding relationship between them.

(2) Internal keywords

Internal keywords refer to the terms which appear frequently in documents related with domain-specific resources and rarely in others. Take example for exam resources, "multiple-choice questions" and "questions answer" belong to internal keywords.

(3) Pages navigation keywords

Pages navigation keywords usually appear in the particular locations of Web pages, as the quick visual indicators. At the same time, they can effectively help the crawlers to crawl and information denoising because of their strong distinguishing capacity.

(4) Banned keywords

Banned keywords include politically sensitive keywords, pornographic violence keywords and so on. In this paper, we adopt the generally accepted banned keywords in search engine fields.

3.2 Domain-specific Resources Collection Based on the Focused Crawlers

Take into account the depth, comprehensiveness, timeliness and accuracy of resources information, two focused crawlers, PRSpider and extended Heritrix are designed. The key tasks of crawlers include pages downloading, pages parsing, URLs filtering and information denoising [10,11]. Additionally, the implementation of information denoising and topic relevance of page body text will be described in section 3.3.

(1) PRSpider+Extended Heritrix

In PRSpider, as shown in Fig. 3, beginning from some given URL seeds, pages downloader downloads continuously Web pages based on the waiting URL queue, and then stores them locally. Next, the new URLs and pages body text are obtained through URL parser and body text parser, and then URLs of Web pages relevant to the topic are added to the waiting URL queue, but URLs, as the download links of documents, are tagged and stored. At last, pages body text denoising, classification, tagging and storage guided by domain-specific resources knowledge base, will be performed. As shown in Fig. 4, Extended Heritrix uses relatively loose rules of URLs extraction and filtration to preliminarily estimate the topic relevance of new URLs. However, page body text extraction and its topic relevance will be performed in domain-specific resources processing.

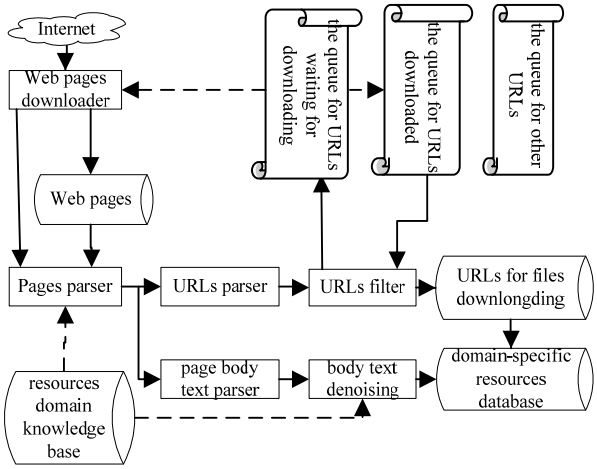


Figure 3. PRSpider general structure

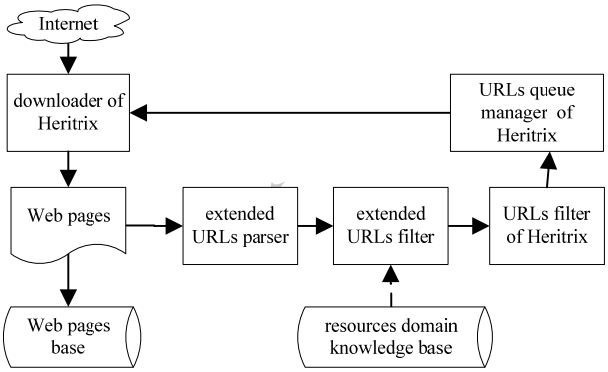


Figure 4. Extended Heritrix general structure

(2) Pages downloading

First we need to choose some URL seeds for the crawler before pages downloading. Exam resources, for instance, we can choose home pages or exam channels of relevant Web sites. Here, as many seeds as possible are collected by manual or automatic operation.

Web pages downloading is the process that the crawler locates the Web pages according to the URLs, sends the access request to the corresponding server, receives files from the server, and then locally stores them in certain formats. We mainly collect files which have the extensions such as .html, .htm or .shtml, rename and then store them locally based on their URLs.

(3) Pages parsing

Pages parsing is the process that the crawler extracts particular information from the source files of Web pages.

978-1-935068-76-1 © 2011 SciRes.

Here, the new URLs, page body text and other relevant information are extracted. First determine the page type (theme page or navigation page) according to the URL, and then respectively extract URLs or page body text. A URL and its anchor text can be extracted according the particular format of URL in the source file. Page body text extraction will be detailed in section 3.3.

(4) URLs filtering

First, preliminarily estimate the topic relevance of new URLs depending on the corresponding anchor text, page title and page structure. And then classify URLs into three types according to their transfer protocols, file paths, file extensions and anchor text: URLs pointing to Web theme pages will be added into the waiting queue, URLs pointing to files downloading such as WORD files and text files, will be stored, and irrelevant or invalid URLs will be abandoned.

3.3 Domain-specific Resources Processing

The service platform must effectively process Web pages downloaded from the Internet and then provides efficient resources information service. And this module includes page body text extraction, topic relevance calculation of body text and information denoising [12,13].

(1) Body text extraction of Web pages

Shown in Fig. 5, first divide a Web page into three regions: file head block, text title block and body text block, and then respectively extract information relevant to body text, such as the page title, page abstract, text title, creation time, body text and so on. Next, restore the URL of the Web page depending on its file path stored locally, generate the local path of each file corresponding to the page body text, and then integrate them into the tag file which is stored locally and includes all information relevant to page body text. Here, data integration can ensure in priority that body text and text title are not empty by replacing or duplicating between the page title, page abstract, page body text and text title.

(2) Topic relevance of page body text

Topic relevance of body text depends on URL, text title, page body text, page title and so on, as shown in Fig. 6.

(3) Information denoising

Information denoising goes through the procedure of domain-specific resources collection and processing, including pages parsing, pages type determining, topic relevance and so on, as shown in Fig. 7.

When determining pages type, information denoisng is to identify and then abandon noise pages such as error pages, navigation pages, irrelevant pages and so on. And noise pages in specific domain can be effectively identified based on URL, page title, page content and the size of page file.

When parsing pages, information denoising includes body text denoising and characters cleaning. Body text denoising has been included in the extraction process of page body text, and body text noise includes HTML tags

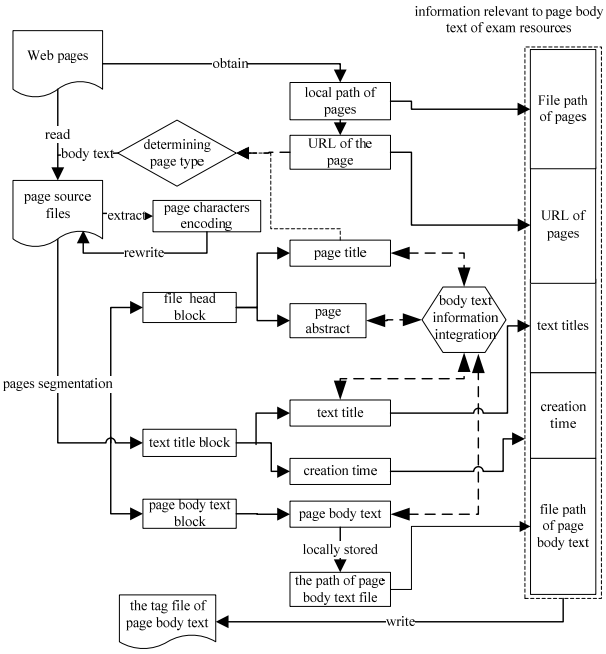


Figure 5. Web text extracting

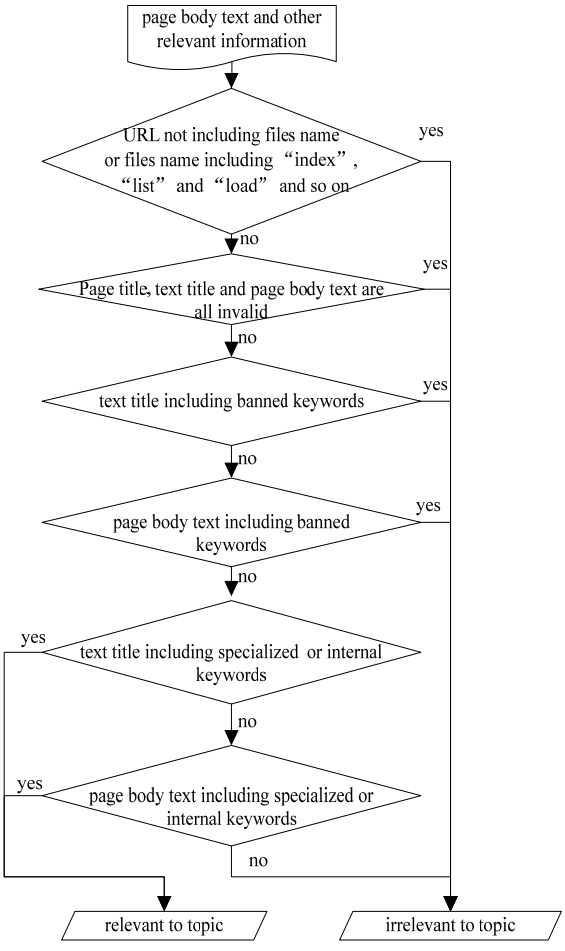


Figure 6. Estimation of opic relevance of body text

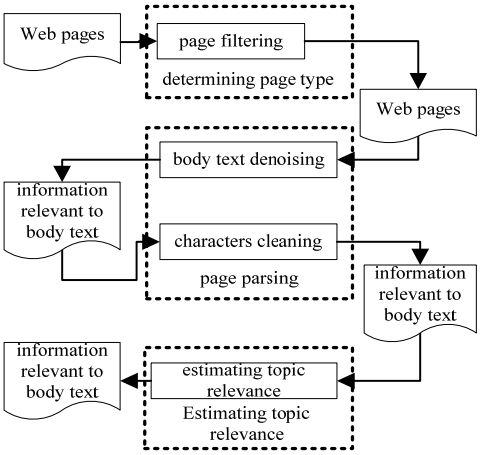


Figure 7. Information denoising

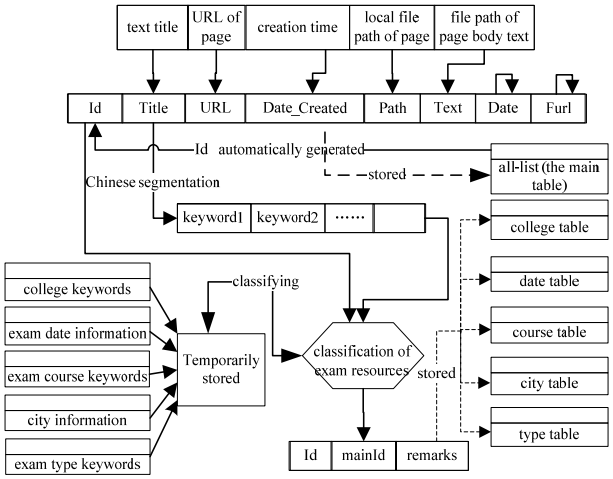


Figure 8. Information classification and storage

disturbing page segmentation, other text irrelevant to body text such as "copyright" and other HTML tags. Characters cleaning is to delete redundant characters, replace illegal characters, and correct wrong characters.

When calculating topic relevance of page body text, information denoising is to identify and abandon irrelevant page body text, as shown in Fig. 6.

3.4 Structured Storage of Domain-specific Resources Information

(1) Tagging of the domain-specific resources

The module contains tagging, classification and storage of the domain-specific resources information. After extraction, page body text is locally saved as a text file and relevant information such as text title, creation time, URL, page file path and the file path of page body text, is all written to the local tag file of page body text. Additionally, the information relevant to URLs pointing to files downloading is also stored locally in the other tag file.

(2) Classification and storage of the domain-specific resources

Classification and storage of the domain-specific resources information are explained, taking example for exam resources, shown in Fig. 8.

We design an exam resources database by combining the main table with the secondary tables. The main table is to store exam resources information including the text title, page URL, father page URL, page downloaded time, page creation time, page file path, body text, record number and so on. And the secondary tables such as a college table, a date table, a course table, a city table and a type table related with exams, are to store classification information of each record of the exam resources information. The secondary tables consist of three attributes: Id, mainId and remarks, and mainId columns are to store the identification numbers of the exam resources records in the main table, and remarks columns are to store the keywords which have been used to classify the exam resources information in the main table.

Depending on the keywords classified and tagged in the domain-specific knowledge base and the text titles of the exam resources records in the main table, the corresponding records are classified and stored to the appropriate secondary tables. URLs pointing to exam resources downloading are handled as the same.

3.5 Web User Interface of the Service Platform

Web user interface of the service platform includes classification retrieval and personalized customization services for domain-specific resources. Classification retrieval can reduce the retrieved scope to improve retrieval accuracy and efficiency, based on the classification information in the secondary table. Customization services refer to provide the appropriate domain-specific resources information to particular users according their personalized needs.

4 Conclusion

In this paper, in order to meet precise, deep and timely retrieval, we study the domain-specific resources service platform based on vertical search technology. In a week, the system implemented towards exam resources runs and downloads data about 20GB which contains 600000 Web pages and URLs for downloading, and about 450000 exam resources records are parsed and stored in the exam resources database. Practical application shows that the system works well. It can search precisely exam resources on the Internet, and effectively process and store them. Web user interface provides classification retrieval and personalized information services. Additionally, two issues need further study. The first is the system framework with scalability and good portability which is easily applied to different domains. And the second is to mine the latent knowledge with the sharp increasing of data in the database.

References

978-1-935068-76-1 © 2011 SciRes.

[1] Sergey Brin and Lawrence Page, "The anatomy of a large-scale hypertextual Web search engine," Computer Networks and ISDN Systems, Vo.l 30, pp. 107-117, 1998.
[2] Jon M. Kleinberg, "Authoritative sources in a hyperlinked environment," Journal of the ACM (JACM), Vol. 46(5), pp. 604-632, 1999.
[3] P.Srinivasan, F.Menczer, G.Pant, "A general evaluation framework for crawlers," Information Retrieval, Vol. 8(3), pp. 417-447, 2005.
[4] M.Chau and H.Chen, "Comparison of three vertical search spiders," IEEE Computer, Vol. 36(5), pp. 56-62, 2003.
[5] B.Novak, "A survey of focused web crawling algorithms," Proceedings of SIKDD, pp. 55-58, 2004.
[6] Michael Chau, Hsinchun Chen, Jialun Qin, Yilu Zhou, Yi Qin, Wai-Ki Sung, Daniel McDonald, "Comparison of two approaches to building a vertical search tool: A Case Study in the Nanotechnology Domain," JCDL'02, pp.135-144, 2002.
[7] Michael Chau, Jialun Qin, Yilu Zhou, Chunju Tseng, Hsinchun Chen, "SpidersRUs: automated development of vertical search engines in different domains and languages," JCDL'05, pp. 110-111, 2005.
[8] Michael Chau and Hsinchun Chen, "Using content-based and link-based snalysis in building vertical dearch rngines," ICADL, pp. 515-518, 2004.
[9] Shicong Feng, Li Zhang, Yuhong Xiong, Conglei Yao, "Focused crawling using navigational rank," CIKM'10 Proceedings of the 19th ACM international conference on Information and knowledge management, pp. 1513-1516, 2010.
[10] Shuyi Zheng, Pavel Dmitriev, C.Lee Giles, "Graph-based seed selection for web-scale crawlers," CIKM'09 Proceeding of the 18th ACM conference on Information and knowledge management, pp. 1967-1970, 2009.
[11] Dirk Ahlers, Susanne Boll, "Adaptive geospatially focused crawling," CIKM'09 Proceeding of the 18th ACM conference on Information and knowledge management, pp. 445-454, 2009.
[12] Takuya Maekawa, Yutaka Yanagisawa, Yasushi Sakurai, Yasue Kishino, Koji Kamei, Takeshi Okadome, "Web searching for daily living," SIGIR'09 Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, pp. 27-34, 2009.
[13] Li GuangLi, Liu JueFu, "Research and implement of vertical search engine," JOURNAL OF INTELLIGENCE, vol. 28(10), pp. 144-147, 2009.

Scientific Research

Competitive Intelligence Service and Interactive Mode Based on the Technology Innovation Demands of SME

Weisi LI[1], Min SHI[2], Xuekui XIAO[3]

Competitive Intelligence Center, Hunan Scientific and Technical Information Institute, Changsha, China

Email: Liweisi_710@sohu.com, 631725@qq.com, sm8@21cn.com

Abstract: With the global economy integration and market competition, Small and Medium Enterprises which are lack of money and technology are meeting the serious challenges. In this paper, based on the theory of features of the technology innovation in SME, 129 high-tech enterprises in Hunan are surveyed. And then the competitive intelligence service and interactive mode are proposed to establish an interactive platform, which includes cooperation of industry, universities, research institutes, technological financing, talent introduction, achievements commercialization, technology introduction and new-product development six information interactive models for these enterprises, governments, investors and experts.

Keywords: high-tech SME; technology innovation; competitive intelligence; information interactive

1 Introduction

With the global economy integration and market competition, high-tech Small and Medium Enterprises (SME) which are lack of money and technology are meeting the serious challenges. Information service cannot be supported to high-tech SME effectively and efficiently, so the SME is difficult to fit the rapidly changing environment. The construction of competitive intelligence service system for high-tech SME can help them foster information ability and accommodate market circumstance. In recent years, there are many scholars have brought up the idiographic models and experiences for SEM. Z.A. Banaszak[1] introduced constraint programming (CP) as the technical framework to offer support means for decision- making of SEM. M. Levy and P. Powell[2] pointed out that information system of SEM must take the organizational reform into account. Johansson[3] researched processes and factors of decision-making on ASP Technology for SEM. Zhang Yucai and Song Xinping [5]designed the information service system including enterprise environment, information management within companies, consultative, evaluation, financing, training and intellectual property electronic system for SME. Wang Jin [6] introduced the work practice of scientific & technical information service in Sichuan province.

Based on the theory of technology innovation in SME, the experience and practice of Hunan Competitive intelligence Center is described in this context. We put forward a framework of competitive intelligence service and interactive mode to support the high-tech SME.

2 Technology Innovation Information Demands of High-tech SME

2.1 Technology Innovation Theory of High-tech SME

Different theoretical viewpoints of technology innovation ability were pointed out. D. L. Barton[4] considered the most important factors of technology innovation ability are senior technical personnel, technology system, management system, attitudes and values of companies. Fu Jiaji[5] thought the technology innovation ability is consist of investment, management, technological selection, research & development, manufacture, marketing and innovation output ability.

Taking the features of technology innovation in SME into consideration, Hunan Competitive intelligence Center surveyed 129 high-tech companies and analyzed their innovation restricting factors and development demands.

2.2 Technology Innovation Restrict Factors of High-tech SME

High-tech SME which are lack of money and technology always develop slowly and difficultly. According to the survey of the 129 High-tech companies, 109 interviewees showed that the lack of funds and resources hindered the development mostly. 108 interviewees said they are short of the support by governments, and then the talent gap, information asymmetry and technical limitation are also the restricting factors. (See Figure.1)

1) Financing channels. The financing environments and channels are not open completely to high-tech SMEs. It's always considered that the SME credit guarantee and investment returns are the most problems by banks, funds, venture capitals and other financing institutions, therefore, they are unwilling to loan or invest in SMEs.

2) Support of policy. The financial support of governments is also quite limited in our country. The infor-

National Natural Science Foundation Project. (*70972119*); Hunan Natural Science Foundation Project, "Morden Competitive Intelligence Theory and Methodology" (*2007JJ5113*); Hunan Science and Technology Infrastructure Platform Project, "Competitive Intelligence Platform in Hunan Province" (*2007TP2002*)

978-1-935068-76-1

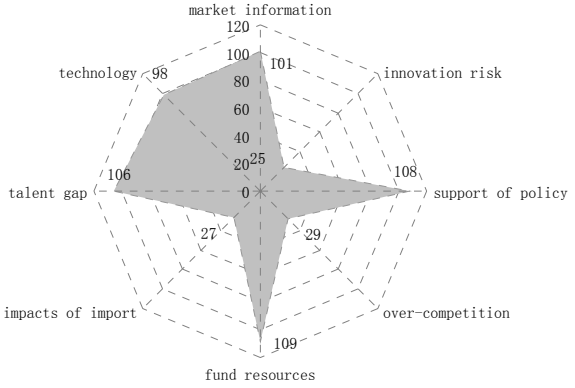


Figure 1. The restricting factors of high-tech SMEs' development in Hunan province

mation, which includes government project examination and approval, tax reduction and exemption, is so unavailable and inefficient that the SME often lose the opportunity to get assistance. If we can offer more information channels and consultations, they could enjoy more development opportunities.

3) Talent gap. It's one of the most serious problems of SME. Considering the career plan and development, senior technical and managerial personnel prefers engaged in large-scale enterprise instead of SME.

4) Market information. Interviewees showed that they cannot get the market information timely, especially like the tendency of industry chain development, the price and yield of production, activities of competitors. As a result, when making decision for business strategy and market exploitation, they are short of information support.

5) Technology research and development. Many SMEs have reached to a plateau for years, without production process flow improvement, feeling hard in technology import, imitation, adsorption and re-innovation.

2.3 Technology Innovation Information Service Demands of High-tech SME

According to the survey of 129 high-tech companies, we find that SMEs are in urgent need of competitive intelligence, mostly in fund, talent, market and technology.

1) Financing information demands. Financing from internal reserves and from external reserves are two basic ways for an enterprises to accumulate funds. In SME, however, information asymmetry hinders the companies financing mostly. In the survey, 97.67% interviewees hope to gain information, and 89.92% hope to be invested by venture capital investment. Most administrators considered the information about getting bank loan, venture capital investment, governmental project is most needed.

2) Talents information demands. SMEs are limited by organization structure and property scale. 97.67% interviewees hope to be supported by recruitment services. It is said that the cooperation of industry, academe and research institutes is the best way to contribute to develop innovation, but the companies don't know how to look for a corporate partner to develop technology and bring it to the market.

3) Market information demands. The information demands are concentrated on products price, yield, activities and strategic management of competitors. 98.44% companies expect to gain market information.

4) Technology information demands. In this survey, 79.97% companies expect to gain patent, literature of science and technology. They require monitoring technology tendency, patent application at home and abroad, constructing patent early warning system to promote technology import, adsorption and re-innovation.

3 Competitive Intelligence Service and Interactive Mode for High-tech SME

Traditional channels in SMEs to gain experts, technology, market information are from social network, which is in low-level efficiency and hinder the integration of resources. SME have no idea to attract venture investment, technology experts, so the funds, human, technology resources cannot be integrated effectively. Hunan Competitive intelligence Center has designed a system framework to integrate resources into the platform and establish communication in high-tech SME, investors, experts, governments, universities and research institutes. (see Figure.2)

3.1 Competitive Intelligence Service and Interactive Mode for High-tech SME

Communication and interaction among high-tech SME, investors, experts, governments, universities and research institutes could be established by the competitive intelligence platform. SMEs submit their innovation demands

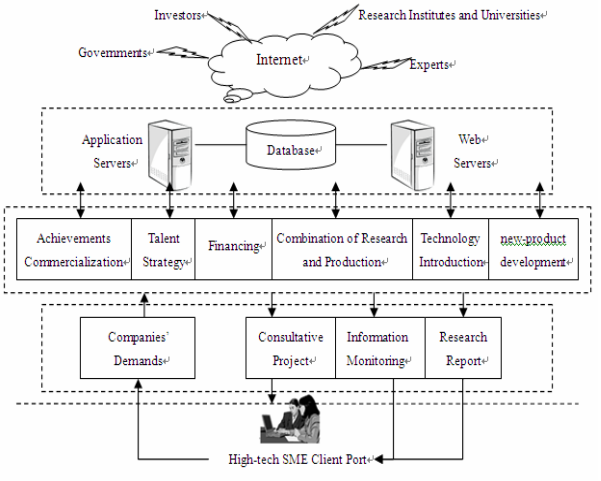


Figure 2. Competitive intelligence service and interactive mode based on the technology innovation demands of SME

throughout the client port, which includes commercialization of scientific and technological achievements, technical R&D cooperation, science and technical financing, technology and talent introduction. Then the information of investment institutions, technical institutes, experts, patents, literatures are collected, indexed and analyzed. Endingly the competitive intelligence feedback, which is in the form of project consultation, monitoring survey, and research report, is pushed to SME based on the demands. The interactive modes of this platform consist of cooperation of industry, universities, research institutes, technical financing, talent introduction, achievements commercialization, technology introduction, and new-product development.

1) Cooperation of industry, universities, research institutes. In this mode, the information of universities, institutes, and key state laboratories is collected and analyzed based on neural network model to find out the correlation between companies and research organization. Then the alternative cooperation partners could be submitted to SME by the client port.

2) Technological financing. On this platform, the company's credit and innovation ability evaluation model is established, which can be reference basis for making decision to invest or not by banks and venture investments.

3) Talent introduction. In this mode, experts' information could be integrated in this system, including the experts' research field, paper, patent, achievements. We look for and monitor the qualified scientists and technicians to support talent strategy to SME.

4) Achievements commercialization. The new technology and research project information would be tracked and pushed to SME, to help them to commercialize the new technological achievements.

5) Technology introduction. In this mode, project, patent and literature information at home and abroad is supported to SME to promote technology digestion, absorption and re-innovation.

6) New-product development. The capacity of the market, development of industry chain, product exploitation of competitors could be tracked, analyzed and forecasted to offer marketing strategy support.

3.2 Competitive Intelligence Service Suggestion for High-tech SME

In order to promote SME's information gaining ability, reduce their cost of information collection and disorder of manufacture and management, science & technical information system should be established by information service institute.

1) Strengthen the sense of information value of high-tech SME, guide them to make technological demands. Many SMEs have reached a plateau for years. The managers without innovative and information consciousness considered the development and promotion of SME is unnecessary. Science and technical information service institute should broaden the ways to exchange information with SME, strengthen their innovative consciousness, guide them to make technological demands. It contributes to integrating resources and promoting competitive advantages of SME.

2) Promote the technology and mechanism innovation. In information age, high-tech SME should utilize Internet and product resources to promote the technology and products quality. Besides it, SME also could apply automatic and flexible manufacture system to improve process flows and administrative system, establish competitive intelligence system management.

3) Strengthen the information infrastructure development. The governments should concentrate funds, create conditions for SME. Competitive intelligence servicing and sharing platform and client pots must to be planned as soon as possible. The SME should be encouraged to realize informationization, establish network with research institutes, investors, and other innovation objects.

4) Establish external information service. The competitive intelligence service institute and consultative agency should be encouraged established to promote external information service. Limited by capital scale, SMEs are more suitable to gain information from external service, which is supported by governmental policies and measures. So the establishment of policy measures to promote external information service is an effective way to improve the research and interactive ability in SME.

4 Conclusion

Technology innovation theory of high-tech SME is proposed firstly to support the high-tech company survey, which showed that funds, talent, market and policy information are the most restricting factors and innovation demands. According to the demands, interactive platform for these enterprises, governments, investors is suggested to be established, including cooperation of industry, universities, research institutes, technological financing, talent introduction, achievements commercialization, technology introduction and new-product development six information interactive models. The competitive intelligence service and interactive mode are theoretical and practice basic support for the other information service work to SMEs.

Acknowledgment

Weisi LI, Min SHI and Xuekui XIAO thank high-tech companies of Hunan Province to complete the technology innovation survey, thank ISTIC to offer this opportunity to make international academic exchanges.

References

[1] Z.A. Banaszak, M.B. Zaremba, W. Muszyński. CP-based

decision making for SME [A]. Prague: Preprints of the 16th IFAC world congress [A], 2005.
[2] M. Levy, P. Powell. Information systems strategy for small and medium sized enterprises: an organisational perspective [J]. Journal of Strategic Information Systems,2000,(9):63-75.
[3] Johansson B, Carlssons A. Evaluation of an application sew ice provider: an SME view[C]. Madrid Spain: The Proceedings Of The 10th European Conference On Information Technology Evaluation[A]. 2003.
[4] Barton D L. Core capability and core rigidities: A paradox in managing new product development [J]. Strategic Management Journal, 1992, 13: 111-125.
[5] Fu Jiaji. Technology Innovation[M]. Beijing: Tsinghua University press, 1998:321.
[6] Zhang Yucai, Song Xinping. A study on high-tech SMEs-oriented information service system [J]. Science and Technology Management Research, 2008(11):241-244.
[7] Wang Jin. Analysis of the Science & Technology Information Service in the Small & Medium-sized Companies in Sichuan[J]. Soft Science,2008,108,(12):91-92.
[8] Hu Ying ,Zhong Weizhou. Relational Finance Research based on Value Analysis [J]. Circulation of Economy in China, 2009(11):71-74.

Study and Application on the Law of Scattering Distribution of Agricultural Scientific Documentation

——Demonstration Analysis from CAB Abstracts

Yun YAN, Xiaolu SU, Yuhong XU

Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing, China

Email: yan@mail.caas.net.cn

Abstracts: Selects the international well-known secondary document database-CAB Abstracts as research tool, and Bradford's Law Scattering as reference frame, this paper studies the scattering distribution patterns of periodical documents in different disciplines of Agriculture, by taking periodical documents as measure unit and large sample bibliometrical statistical analysis. The results show that there are some differences in scattering distribution patterns of scientific periodical paper in different disciplines of Agriculture, the differences could be measured by scattering degree.

Keywords: documents Distribution; Scattering degree; periodical; Collection Construction; Agriculture

农业学科科技文献离散分布规律研究与应用

——CAB Abstracts 实证分析

颜蕴，苏晓路，续玉红

中国农业科学院农业信息研究所，北京，中国，100081

Email: yan@mail.caas.net.cn

摘　要：以国际著名的农业权威二次文献数据库 CAB Abstracts 为研究工具，以布拉德福文献分散定律为参照系，选择期刊论文为计量单元，通过大样本量的文献计量统计分析，研究农业不同专业领域期刊文献离散分布规律。结果表明：农业学科不同专业领域科技期刊文献离散分布规律因专业领域的不同而有所差别，这种差别可用文献离散度指标来衡量。

关键词：文献分布；离散度；期刊；资源建设；农业

1 引言

布拉德福定律（Bradford's Law Scattering）的基本内容是“如果将科学期刊按其登载某个学科的论文数量的大小，以渐减顺序排列，那么可以把期刊分为专门面向这个学科的核心区和包含着与核心区同等数量论文的几个区”。如果用图像来描述，则取等级排列的期刊数量的对数 logN 为横坐标，以相应的论文累积数 R（n）为纵坐标进行图像描述，绘制出布拉德福分散曲线，分散曲线为是由对应核心区的上升的曲线和对应相继各区的平滑直线组成。布氏定律作为图书情报工作科学管理的基础理论，特别是文献资源建设工作中，应用其理论和研究方法确定核心期刊、制定文献采购策略和藏书政策、优化馆藏、指导期刊订购工作等均有一定的指导作用。

本文以世界权威农业文摘数据库 CAB Abstracts 的数据为分析对象，应用布拉德福文献离散分布定律，采用等级排列技术和文献计量学分析方法，研究农业学科 9 个专业领域科技期刊文献离散分布规律，以确定农业学科不同专业领域核心区域、次核心区域期刊数量及其占本领域期刊的比例，为不同专业领域期刊资源建设的深度和广度提供科学研究依据。

2 研究方法

选择国际著名农业文摘数据库 CAB Abstracts（OVID Technologies 网络平台）为计量分析对象，检索下载农业学科不同专业领域的期刊文摘记录，数据处理采用 SQL Server 建库工具和 Excel 电子表格工具进行，对抽取的数据按同种载体（期刊）文献进行归并，以各专业来源期刊载文量的大小，渐减顺序排列，进行以下统计和分析：

（1）对 9 个农业专业的下载数据按载文量多少进

978-1-935068-76-1

行渐减排序，绘制布拉德福文献离散分布点阵图型，观察不同农业专业期刊文献离散分布曲线差异；

（2)计算 9 个农业专业下载数据的单刊载文量占总载文量比例和期刊数量占该专业来源期刊数量的比例，进行渐减进行排序，分别计算 80%载文量、80%—90%载文量、90%—100%载文量三个区域期刊所占比例，并用百分比堆积柱形图来描述；

（3）以各专业占 80%载文量期刊数量与该专业总来源期刊数量的比例与帕累特“20/80”法则 20%相比较，分析对比农业学科不同专业领域的期刊论文的离散分布规律，研究建立农业学科各专业文献离散度参数。

3 数据统计

本项研究的数据来源为世界农业大型文摘数据库 CAB Abstracts，该数据库于 1973 年建立，是世界权威农业文摘数据库，收录了世界范围内 140 多个国家 1 万余种文献的信息，主要的文献类型有期刊、会议录、文集、年报、专利、专著等，该数据库制作精良，具备完备的主题词表和分类代码体系，目前数据量达 600 余万篇，其中期刊论文占总文献量的 80%以上，基本涵盖了世界范围内主要农业科技期刊。

为了保证数据样本的完整性，利用 CAB Abstracts 数据库学科分类标识系统 CABI Classification System（简称 CABCODE)分专业抽取 2004—2006 年 3 年出版的期刊论文数据，进行数据处理和统计分析。选择的专业分农业经济（CABCODE 为 CC)、作物科学（FF)、植物保护（HH)、土壤科学（JJ)、林业科学（KK)、动物科学（LL)、农业自然资源（PP)、农业工程（NN)、食品科学（QQ）等 9 个，将抽取的数据按同种载体（期刊）文献进行归并、聚类，总体情况如表 1 所示：

本项研究分析的数据记录总量为 714，357 条，数据规范、样本量大，利用其进行分析能达到较高的可信度。

4 结果分析

4.1 农业学科不同专业领域期刊论文的布氏分布曲线描述

将各专业按期刊按其登载某个专业的论文量的大小，以渐减顺序排列，得到“农业经济”、“作物科学”、“植物保护”、“林业科学”、“动物科学”、“土壤学”等 9 个专业领域来源期刊文献累计载文量的排序表，

Table 1. Number of agricultural documents and periodicals of 9 fields from CAB Abstracts
表 1. CAB Abstracts 9 个专业农业文献及期刊数量一览表

学科	文献量（篇）	期刊数量（种）
农业经济	36,971	3,065
作物科学	192,067	3,956
植物保护	74,617	4,078
土壤学	63,106	2,754
林业科学	54,240	2,920
动物科学	108,784	3,325
农业工程	17,845	1,762
农业自然资源	99,277	3,822
食品科学	67,450	3,487

如表 2 所示。(考虑到表格的篇幅和美观因素，表 2 仅对载文量前 13 位的期刊及其论文数量进行罗列，中间省略部分用 “……” 表示，并将不同专业论文来源期刊总量和载文总量统一排列在末行中，以方便观测。)

以期刊累积数量 n 的对数 logN 为横坐标，以累积载文量 R（n）为纵坐标，描绘农业不同专业领域期刊累积载文量的布拉德福文献分布曲线如图 1：

从图 1 可以看出，不同专业领域期刊载文量的布氏分布曲线形状较为相似，但其曲线的斜度，拐点的位置有所不同，说明不同专业领域文献离散规律有一定的差异。

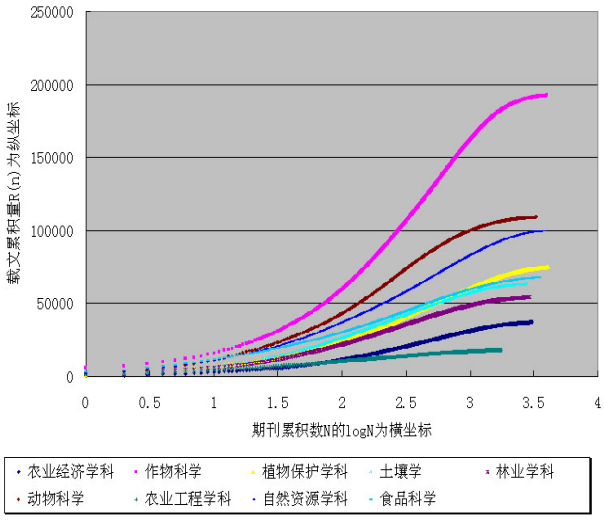


Figure 1. Distribution curve of periodical articles 9 fields of agricultural documents
图 1. 农业学科 9 个专业领域期刊文献布氏分布曲线

Table 2. Sort table of amount papers of 9 fields from CAB Abstracts
表 2. CAB Abstracts 9 个专业领域期刊文献载文量排序表

农业经济		作物科学		植物保护		林业科学		土壤学		动物科学		自然资源		食品科学		农业工程	
期刊数	累积载文量	期刊数	累积载文量	期刊数	累积载文量	期刊数	累积载文量	期刊数	累积载文量	期刊数	累积载文量	期刊数	累积载文量	期刊数	累积载文量	期刊数	累积载文量
1	418	1	5585	1	1218	1	1583	1	844	1	1505	1	1556	1	2677	1	964
2	824	2	7010	2	2077	2	2403	2	1632	2	2852	2	3012	2	4268	2	1539
3	1151	3	8422	3	2921	3	3071	3	2408	3	3974	3	4309	3	5600	3	2080
4	1448	4	9635	4	3487	4	3558	4	3073	4	5058	4	5423	4	6693	4	2463
5	1715	5	10726	5	4039	5	4005	5	3669	5	6110	5	6502	5	7694	5	2828
6	1949	6	11793	6	4567	6	4410	6	4238	6	7142	6	7431	6	8588	6	3188
7	2180	7	12803	7	5079	7	4778	7	4798	7	8077	7	8349	7	9340	7	3512
8	2410	8	13780	8	5589	8	5141	8	5347	8	8984	8	9151	8	10092	8	3775
9	2620	9	14750	9	6083	9	5503	9	5869	9	9866	9	9904	9	10839	9	4033
10	2811	10	15667	10	6553	10	5849	10	6371	10	10654	10	10625	10	11404	10	4289
11	2998	11	16582	11	6990	11	6179	11	6826	11	11425	11	11284	11	11886	11	4520
12	3181	12	17440	12	7423	12	6501	12	7275	12	12176	12	11921	12	12360	12	4700
13	3358	13	18282	13	7782	13	6816	13	7705	13	12927	13	12484	13	12783	13	4865
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
3065	36971	3956	192067	4078	74617	2920	54240	2754	63106	3325	108784	3822	99277	3487	67450	1761	17844

Table 3. Sort table of proportion amount papers for source journals 9 fields from CAB Abstracts
表 3. CAB Abstracts 9 个专业领域文献量占来源期刊比例排序表

序号	农业经济		作物科学		植物保护		林业科学		土壤学		动物科学		自然资源		食品科学		农业工程	
	占 3065 种来源期刊比例	载文所占比例	占 3956 种来源期刊比例	载文所占比例	占 4078 种来源期刊比例	载文所占比例	占 2920 种来源期刊比例	载文所占比例	占 2754 种来源期刊比例	载文所占比例	占 3325 种来源期刊比例	载文所占比例	占 3822 种来源期刊比例	载文所占比例	占 3487 种来源期刊比例	载文所占比例	占 1762 种来源期刊比例	载文所占比例
1	0.00033	0.01131	0.00025	0.02908	0.00025	0.01632	0.00034	0.02919	0.00036	0.01337	0.0003	0.01383	0.00026	0.01567	0.00029	0.03969	0.00057	0.05402
2	0.00065	0.01098	0.00051	0.00742	0.00049	0.01151	0.00068	0.01512	0.00073	0.01249	0.0006	0.01238	0.00052	0.01467	0.00057	0.02359	0.00114	0.03222
3	0.00098	0.00884	0.00076	0.00735	0.00074	0.01131	0.00103	0.01232	0.00109	0.0123	0.0009	0.01031	0.00078	0.01306	0.00086	0.01975	0.0017	0.03032
4	0.00131	0.00803	0.00101	0.00632	0.00098	0.00759	0.00137	0.00898	0.00145	0.01054	0.0012	0.00996	0.00105	0.01122	0.00115	0.0162	0.00227	0.02146
5	0.00163	0.00722	0.00126	0.00568	0.00123	0.0074	0.00171	0.00824	0.00182	0.00944	0.0015	0.00967	0.00131	0.01087	0.00143	0.01484	0.00284	0.02045
6	0.00196	0.00633	0.00152	0.00556	0.00147	0.00708	0.00205	0.00747	0.00218	0.00902	0.0018	0.00949	0.00157	0.00936	0.00172	0.01325	0.00341	0.02017
7	0.00228	0.00625	0.00177	0.00526	0.00172	0.00686	0.0024	0.00678	0.00254	0.00887	0.00211	0.0086	0.00183	0.00925	0.00201	0.01115	0.00397	0.01816
8	0.00261	0.00622	0.00202	0.00509	0.00196	0.00683	0.00274	0.00669	0.0029	0.0087	0.00241	0.00834	0.00209	0.00808	0.00229	0.01115	0.00454	0.01474
9	0.00294	0.00568	0.00228	0.00505	0.00221	0.00662	0.00308	0.00667	0.00327	0.00827	0.00271	0.00811	0.00235	0.00758	0.00258	0.01107	0.00511	0.01446
10	0.00326	0.00517	0.00253	0.00477	0.00245	0.0063	0.00342	0.00638	0.00363	0.00795	0.00301	0.00724	0.00262	0.00726	0.00287	0.00838	0.00568	0.01435
11	0.00359	0.00506	0.00278	0.00476	0.0027	0.00586	0.00377	0.00608	0.00399	0.00721	0.00331	0.00709	0.00288	0.00664	0.00315	0.00715	0.00624	0.01294
12	0.00392	0.00495	0.00303	0.00447	0.00294	0.0058	0.00411	0.00594	0.00436	0.00712	0.00361	0.0069	0.00314	0.00642	0.00344	0.00703	0.00681	0.01009
13	0.00424	0.00479	0.00329	0.00438	0.00319	0.00481	0.00445	0.00581	0.00472	0.00681	0.00391	0.0069	0.0034	0.00567	0.00373	0.00627	0.00738	0.00925
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
	1	2.7E-05	1	5.2E-06	1	1.3E-05	1	1.8E-05	1	1.6E-05	1	9.2E-06	1	1E-05	1	1.5E-05	1	5.6E-05

978-1-935068-76-1 © 2011 SciRes.

4.2 农业学科不同专业领域文献离散分布规律分析

长期以来，图书情报学家对布拉德福离散分布规律进行了大量探索，形成了多个经验公式和数学模型，但至今尚无统计的定论。考虑科技论文的分布受多种因素的影响，带有主观性和模糊性，本文不打算对农业专业领域文献的离散规律进行数学表达，而是想寻求一种简单的方法对不同专业领域论文离散分布情况进行比较。

采用不同载文量百分比来进行专业间文献离散情况的比较。首先通过对每一专业、每一种期刊载文量进行统计，同时计算期刊载文量占该专业载文量的比例，并分专业按期刊载文量渐减次序排列，得到 9 个专业中期刊所占比例与载文比例的数据表，如表 3。

根据帕累特“80/20 法则”（即 20%的文献涵盖了 80%的主要信息内容），利用表 3 中载文所占比例项的数据，将各专业期刊论文按累积载文量分别达 80%、80%—90%、90%—100%划分期刊载文核心区、载文次核心区、载文外围区，计算每个区域期刊数量占该专业论文来源期刊的总品种数的比例，对农业不同专业领域文献离散分布规律进行比较，如表 4 所示。

Table 4. The Scattering Distribution of accumulative number of papers for agricultural subjects
表 4. 农业学科各专业累积载文量离散分布

学科	核心区期刊比例（累积载文量 80%）	次核心区期刊比例（累积载文量 80%-90%之间）	外围区期刊比例（载文量 90%-100%）
农业经济和政策	27.50%	16.25%	56.24%
作物科学	21.21%	11.86%	66.93%
植物保护	23.79%	15.82%	60.40%
土壤科学	24.07%	11.76%	64.16%
林业科学	20.75%	14.75%	64.52%
动物科学	16.15%	11.19%	72.66%
自然资源	22.24%	14.81%	62.95%
食品科学	18.38%	14.94%	66.68%
农业工程	20.09%	18.45%	61.46%

为使分析结果更加直观，利用表 4 中 9 个农业专业文献三个区域期刊的相对比例数值，生成百分比堆积柱形图，见图 2 所示。

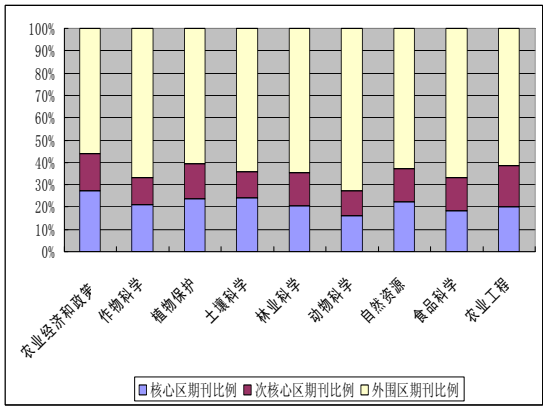


Figure 2. The Graphic description of Scattering Distribution for agricultural subject documents
图 2. 农业学科各专业科技文献载体分布规律图形描述

从表 4 和图 2 看出，农业学科各专业领域覆盖 80%论文（即载文核心区）期刊所占该专业来源总期刊比例差异较大，从高到低依次分别为农业经济 27.5%、土壤科学 24.07%、植物保护 23.79%、农业自然资源 22.24%、作物科学 21.21%、农业工程 20.09%、林业科学 20.75%、食品科学 18.38%、动物科学 16.15%。

4.3 农业学科不同专业离散度指标的建立

为了定量的比较不同专业间文献的离散规律，本项研究引入了“离散度”概念，即按覆盖 80%文献的期刊数量占该专业领域期刊数量的比例，与 20%这个经验定律相除，得到专业领域文献“离散度”指标，如表 5 所示：

Table 5. Scattering degree of 9 fields of agricultural documents
表 5. 农业学科 9 个专业领域学科文献离散度

专业	核心区期刊比例（累积载文量达 80%）	文献离散度
农业经济	0.275	1.375
作物科学	0.2121	1.061
植物保护	0.2379	1.119
土壤科学	0.2407	1.204
林业科学	0.2075	1.038
动物科学	0.1615	0.808
自然资源	0.2224	1.112
食品科学	0.1838	0.919
农业工程	0.2009	1.005

从表 5 中看出，农业学科不同专业期刊文献离散度差异较大，离散度越高表明该专业文献分布越分散，核心文献越分散，反之，离散度越低，该专业文献离散分布越小，核心文献越集中。

建议在期刊采集工作中，可根据不同学科或专业建立二级文献离散度指标，将其作为调节因子，在新增期刊遴选时与期刊收藏量、出版量、需求强度等几项指标综合考虑，确定不同专业文献的遴选策略，保证各专业间文献的收藏和保障水平达到平衡。

5 结论和应用

农业学科不同专业领域科技期刊文献离散分布规律因专业领域的不同而有所差别。农业经济、农业自然资源、土壤科学等专业论文量达 80%时所含期刊品种的比例较其他专业领域明显较大，这一现象与本单位在外文期刊遴选工作实践中的感性认识相吻合。

建立学科或专业离散度指标，并根据该指标来指导期刊资源建设，如圈定各学科资源建设的薄弱领域，确定不同专业科技期刊的入选比例，均衡学科和专业间的保障水平。

由于不同学科和专业科技文献离散分布规律复杂，尚难采用精确的数学方法进行判定，因此应用“文献离散度”来指导国外期刊遴选时并不能严格根据数值进行定量决策，应采用定量-定性生活相结合的方法。

本研究仅是对农业 9 个专业领域期刊文献进行离散分布的初步研究，研究的结论可能受 CAB Abstracts 文献收录的学科范围、来源期刊的数量影响。

致　谢

中国农业科学院农业信息研究所原在读研究生张禹同学参与部分数据下载和统计工作，在此致谢。

References（参考文献）

[1] Feicheng Ma, Rui Chen, Hong Yuan, Study on the Law of Scattering Distribution of Scientific Information——Demonstration Analysis from Document Level to Content Level (Ⅶ): Comparison and Conclusion [J]. Journal of the China Society for Scientific and Technical Information, 1999.18(1): 79-84.
马费成，陈锐，袁红.科学信息离散分布规律的研究　从文献单元到内容单元的实证分析（(Ⅶ）：总体研究框架[J]. 情报学报，1999.18（1）：79-84.

[2] JunPing Qiu. Bibliometrics [M]. Beijing: Science and Technology Literature Press, 1988.
邱均平. 文献计量学[M].北京：科学技术文献出版社，1988.

[3] Jianyong Zhang, Xiaomin Liu. Development Strategy of Western Periodicals Collection Based on Citation Statistic Analysis [J]. Library and Information Service, 2007, 51(5): 106-109.
张建勇，刘筱敏. 基于引文统计分析的西文期刊馆藏发展策略[J]. 图书情报工作，2007.51（5）：106-109.

978-1-935068-76-1 © 2011 SciRes.

Semantic Sensor Information Processing Using Cloud Computing

Dehua MA, Wantong LI, Chenggui LI
School of Communication engineering, Jilin University, Changchun, China
Email: wenxi.yongyuan@163.com

Abstract: Semantic sensor is now a hot spot of network science and technology. The semantic markers of sensor data, making machines automatically identify sensor information, is the path to automatic processing information and realizing Internet of Things. But, the semantic marks generate a great amount of sensor data. How to fast handle such a large amount of data without data losing became a key of semantic sensor network. According to this problem, we propose a Cloud Computing Architecture for Semantic Sensor Networks. This information oriented architecture is based on Hadoop platform. The architecture using distributed file system Hbase and parallel processing mode MapReduce of Hadoop solve the problem of center node of semantic sensor network in data storage and processing. The architecture consists of two parts: 1) data storage based on Hbase, 2) data processor using MapReduce.

Keywords: sensor network; semantic technology; Hadoop; Hbase; MapReduce

语义传感器网络数据的云计算处理模型

马德华 [1]，李宛桐 [2]，李程贵 [3]
[1,2,3] 吉林大学通信工程学院，吉林长春，中国，130012
Email: wenxi.yongyuan@163.com

摘　要：语义传感器网络是当今网络科技的一个热点，对传感器数据进行语义标记，令机器自动识别传感器信息，是自动化处理信息和实现物联网的必经之路。但是，语义标记后的传感器数据量巨大，如何快速少丢失的处理如此大量的数据也成为了语义传感器网络的一个关键。针对这个问题，本文提出语义传感器网络数据的云计算处理模型，这个面向信息的模型基于开源的云计算平台 Hadoop。模型利用 Hadoop 的基于分布式文件系统的 Hbase 和并行处理模式 MapReduce 解决语义传感器网络中心节点存储和处理问题。模型由两部分组成：1. 基于 Hbase 的数据存储，2. 基于 MapReduce 的数据处理。

关键词：传感器网络；语义技术；Hadoop；Hbase；MapReduce

1 引言

在如今数字化信息化的时代，信息的采集和处理变得越来越重要。无线传感网络的产生满足了人们对数字空间的需要，但由于传感器网络及其数据的异质性，只能对传感器数据进行纵向处理，应用程序无法对传感器数据共享和重用。Lionel[1]提出语义传感器网络，利用语义网中语义标注技术和本体推理技术，通过构建传感器领域本体，对传感器采来的数据进行语义描述和标注，进而利用 jena 和 SPARQL 实现无线传感器数据的操作和推理等智能处理。大型语义传感器网络的应用过程中产生了大量带有语义标记的数据，而且语义传感器数据智能化处理过程也需要处理海量数据。因此语义传感器网络的中心节点存储处理能力将会是限制其发展的瓶颈。

云计算平台拥有分布式文件存储系统、优越的并行处理能力及非结构化数据存储结构。目前，云计算的海量存储能力和并行处理能力已在网页检索、病毒查杀等领域广泛应用。基于云计算结构的一些特点和优势，本文提出了一种语义传感器网络数据的云计算处理模型，利用现有的理论与技术支撑实现语义传感器网络数据的云计算平台存储和处理。

2 背景

2.1 语义传感器网络

语义传感器网络是在传感器网络中加入语义技术，通过构建领域本体，对传感器采来的数据进行语义描述，进而利用 jena 和 SPARQL 推理机实现无线传感器数据的操作和推理。Vincent [2]提出四层的语义传

感器架构，并且通过建筑物火灾案例，证明语义技术对于传感器数据高效的提取和推理。Vincent 提出四层语义传感器架构：1. Sensor Network Data Sources，2. Ontology Layer，3. Semantic Web Processing Layer，4. Application Layer。

2.2 云计算平台

云计算是一种新型的计算模型，它将计算任务分布在大量计算机构成的资源池上，适合对大量数据进行存储计算。其中 Google 作为云计算最大的使用者，提出了自己的云计算架构，包括 GFS、MapReduce、Bigtable、Chubby[4]。其架构如图 1 所示。

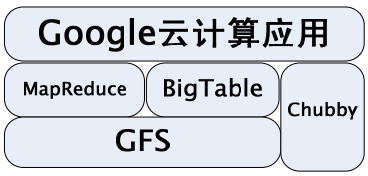


Figure 1. The architecture of clouding computing
图 1. 云计算架构

Google 的核心技术并未公开，本文利用开源云计算模拟框架 Hadoop。Apache Hadoop 作为一个开源项目，模拟了 Google 运行系统的主要框架，包括文件系统 HDFS、计算框架 MapReduce 及对于结构化处理的 Hbase 等[5]。

Google 文件系统（GFS）是一个大型的分布式文件系统。它为 Google 云计算提供了海量存储，并与 MapReduce 及 Bigtable 等技术结合十分紧密，为其它技术提供底层支持。HDFS 作为 Hadoop 的分布式文件系统，相应的实现了 GFS 的大部分功能。

MapReduce 是 Google 提出的一种并行的编程模式，用于大规模数据集的并行计算，特别适合于非结构化和结构化的海量数据的搜索、挖掘、分析与机器智能学习等。MapReduce 由 Map 和 Reduce 两部分组成。用户程序中的 MapReduce 函数库首先把输入文件分块，接着在集群机器上处理。在分派的执行程序中，有一个作为主控程序 JobTracker，其余的作为 JobTracker 分派工作的 TaskTracker。JobTracker 选择空闲的 TaskTracker 来分配 Map 和 Reduce 任务。Maper 函数的输入部分是一对<key,value>，是由源数据进行分片分配所得，输出须是另一对不同性质的<key,value>。在 Mapper 函数一般执行输入格式的映射，实现对数据的分析、选择和过滤。与之不同的，Reducer 的输出需要是一对与输入相同性质的<key,value>，它的输入是 Mapper 函数的输出。Reducer 函数对 Mapper 得到的数据进行 value 值组合或其他处理，使数据进行化简。MapReduce 的最大特点是把 TaskTracker 的任务分给不同的计算机同时进行处理，这改变了传统单核的处理机制，达到了资源利用和吞吐量的提高。MapReduce 的工作机制如图 2 所示。

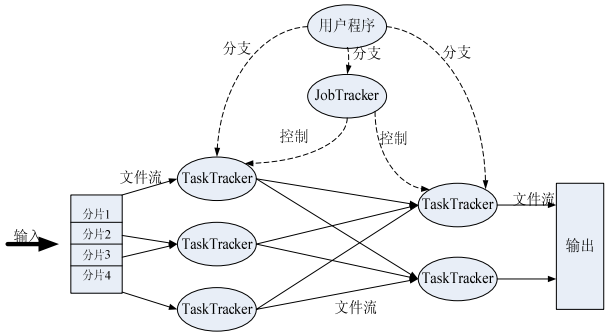


Figure 2. MapReduce Working Mechanism
图 2. MapReduce 工作机制

Bigtable 是一个分布式多维映射表，表中的数据可通过一个行关键字（Row Key）、一个列族（Column Family）以及一个时间戳（Time -Stamp）进行索引的。Bigtable 由行和列组成，表的单元格是行和列的交集，并且在单元格插入时，会自动分配时间戳[3]。 Hbase 是 Bigtable 的开源版本。是建立在 HDFS 之上，提供高可靠的高性能实时读写的数据库系统。主要用来存储非结构化和半结构化的松散数据。Hbase 以表的形式储存数据，表由 row 和 column 组成，表单元由{row key, column(=<family> + <label>), version}唯一确定。Row key 作为行键，也是 table 的主键，存储文件首先会按照主键进行排序。column 划分为若干列族，Hbase 表中的每个列，都归属与某个列族。必须在使用表之前定义。列名都以列族作为前缀。version 的类型是 64 位整型，version 可以由 Hbase(在数据写入时自动)赋值，标记数据写入时间。

3 相关工作

在模型的构思与搭建过程中，我们主要考虑两方面问题，一方面是语义传感器网络的构建，另一方面是云计算平台对语义网数据的处理。在语义传感器网络构建方面，Lionel[1]等人提出语义传感器网络（SSN），并在理论上对 SSN 进行研究，他们的研究表明 SSN 是一种异质传感器网络，SSN 可以动态的把语义信息标记到传感器数据上，因此不同应用程序就可以对传感器数据进行处理。在语义传感器本体构建

978-1-935068-76-1 © 2011 SciRes.

方面，Daniel[6]等人利用现有的本体创建工具，构建了通用的语义传感器本体,作为实现大型的语义传感器网络基础设施。他们构建的本体具有灵活拓展的优点。

同时越来越多的人关注云计算平台与语义技术的融合，他们的目光主要集中在语义网中 RDF 和 OWL 数据的存储和和推理这两方面内容。Mohammad [7]提出存储和检索大量 RDF 三元组的框架，利用 Hadoop 分布式文件系统存储 RDF，并提出一种把 RDF 查询语言 SPARQL 应用于 MapReduce 的算法。浙江大学孙剑林（音译）[7]等人在 Hbase 中用六张表(S_PO, P_SO, O_SP,PS_O, SO_P 和 PO_S)来存储 RDF 三元组。这样可以囊括所有 RDF 模式，而且他利用 MapReduce 实现 SPARQL 的基本 RDF 图处理。在 Hadoop 处理 OWL 应用中。于伟（音译）[8]等提出云计算平台管理本体的框架，框架用 MapReduce 对本体进行管理。

4 语义传感器网络数据的云处理模型

基于云计算的高并行，吞吐量大等特点，我们考虑采用云计算这种的网络结构对语义传感器网络产生的大数据量信息集进行处理。本文提出了一个语义传感器数据的云处理模型。语义传感器网络可以由各种不同功能的传感器节点、网关组成，通过不同的协议进行通信，创建不同领域本体，所以语义传感器网络是一个复杂的异质网络。但是这些不是我们要重点讨论的问题，因为经过语义标记的传感器数据能够自我描述并且都具有相似的结构，可参考 Vincent [2]等人提出语义传感器网络底层架构。所以我们不对云计算处理框架以外的部分做深入讨论。

4.1 基于 Hbase 的数据存储

Hbase 建立在 HDFS 之上，提供高可靠的高性能实时读写的数据库系统。主要用来存储非结构化和半结构化的松散数据。Hbase 以表的形式储存数据，表由 row 和 column 组成，表单元由{row key, column(=<family> + <label>), version}唯一确定。Row key 是 table 的主键，column 可划分为若干列族。Hbase 表中的每个列，都归属与某个列族。必须在使用表之前定义。列名都以列族作为前缀。例如 jilin:changchun，jilin:yanbian 都属于 jilin 这个列族。于是存储过程中我们应用列族和列完成对不同区域的标注。version 可以由 Hbase(在数据写入时自动)赋值，此时时间戳是精确到毫秒的当前系统时间，在这里 version 可作为对数据采集时间的标注。

为了进一步阐述我们提出的模型，以气象信息的语义传感器数据为例，说明云计算对语义传感器数据的处理模型。在传感器采集数据阶段，我们用语义信息标记传感器类型（温度、湿度等），传感器位置（城市），和数据采集时间等数据。把在传感器数据上标记的信息，如传感器类型，传感器位置，和数据采集时间分别写入到 row key，column 和 version。仅以以上三种类型列表如表 1。

Table 1. Storage of Semantic sensor network data in the hbase

表 1. 语义传感器网络数据在 hbase 中的存储

Row key	timestamp	Column family1		…
		jilin:changchun	jilin:yanbian	
temp	Jun.	value1	value3	
	Feb.	…	…	
	Mar.	…	…	
humidity	Jun.	value2	value4	
	Feb.	…	…	
⋮				

4.2 基于 MapReduce 的数据处理

MapReduce 是一种并行的编程模式，用于大规模数据集的并行计算，通过 Map（映射）Reduce（化简）对数据进行处理。在模型中的 MapReduce 处理数据过程，首先对 Hbase 中的传感器数据进行处理，Map 处理过程以 Hbase 中的数据作为输入数据，通过对输入数据列表中的每一个元素应用一个函数创建了一个新的输出数据列表。以求解 B 城市当年一月温度传感器数据最小值为例，对 MapReduce 的处理过程进行说明。

Algorithm 1. Map

Map (InputKey , InputValue)

MISSING <— 9999 ;

For $Mapper_i$ in AllMappers

Get TempSensori and $TempData_i$;

$InputKey_i$ <— $TempSensor_i$;

$InputValue_i$ <— TempDatai ;

If (SensorDatai ≠ MISSING)

Context.write($InputKey_i$, $InputValue_i$);

Algorithm 1 描述了利用 Map 函数，从 Hbase 中各种传感器类型的数据中过滤出温度类型的数据。我们将丢失的数据定义为 MISSING。在 Map 算法中，对每一个执行 Map 工作的 TastTracker，从 Hbase 中读入

几行的具有一种或几种 Row Key 的语义传感器数据，赋值给该 TastTracker (key,value)对,然后输出选出的温度数据（key，value）对。对 Row Key 进行筛选后，继续利用 Map 函数对 column 进行城市的筛选，算法与 Algorithm 1 类似，然后再对 version 进行时间上的筛选。

Reduce 把数据聚集在一起。Reduce 函数接收来自输入列表的迭代器，它会把这些数据聚合在一起，然后进行所需的计算，返回需要的输出值。以下是对温度传感器数据最小值的 Reduce 算法。Algorithm 2 描述了利用 Reduce 函数将所求的温度最小值解出来。

Algorithm 2. Reduce

Reduce (InputKey,InputValue, MinValue)
MinValue <— 100 ;
For $Reducer_i$ in AllReducers
  Get $MapOutputkey_i$ and $MapOutputvalue_i$;
$InputKey_i$ <— $MapOutputkey_i$;
$InputValue_i$ <— $MapOutputvalue_i$;
If ($InputValue_i \neq$ MISSING)
MinValue<—the smaller between MinValue
  and $InputValue_i$;
Context.write($InputKey_i$, MinValue) ;

在 Algorithm 2 中，首先将 MinValue 赋值为 100。对每一个执行 reduce 工作的 TastTracker，将 map 过程的输出（key,value）对作为输入，这里因为是针对 B 城市的温度进行某一时间段的比较，所以 key 选定为 B 城市的的温度。对于相同的 key，对 value 进行迭代，比较输入 value 与 MinValue 的大小。迭代完成，输出要求的最小温度（key,value）对。

完成 MapReduce 后，准备进行数据转换，将处理结果转换成 RDF 格式数据，将 MapReduce 处理结果转成 RDF 主要有以下两点考虑：1.RDF 是 W3C 推荐标准，因此互联网上任何有权限的用户可以通过能够解析 RDF 数据的应用程序都可以对它读取和处理。2. 考虑到本模型在 MapReduce 处理过程中实现了部分 SPARQL 逻辑推理功能，为了更进一步对数据进行智能化处理。我们把 MapReduce 处理的数据结果转换成 RDF 数据，应用程序可以调用 JenaAPI 对 RDF 进行逻辑推理。

5 结束语

本文提出语义传感器网络数据的云计算处理模型，这个面向信息的标记数据存储和处理模型基于开源的云计算平台 Hadoop。利用 Hbase 对标记数据进行高效存储，通过 MapReduce 并行处理海量数据，可以实现部分 SPARQL 推理功能。在模型利用 MapReduce 处理数据过程，对数据智能处理并不完善，下一步工作致力于对 MapReduce 处理 RDF 类型传感器数据的研究，进而实现更复杂的推理功能。

References（参考文献）

[1] Lionel, M. N., Zhu, Y. Semantic Sensor Information Description and Processing. “Semantic Sensor Net: An Extensible Framework”. 2005.1144-1153.

[2] Vincent Huang, Muhammad Kashif Javed. Semantic Sensor Information Description and Processing “The Second International Conference on Sensor Technologies and Applications” 2008. 456-461.

[3] Daniel O'Byrne, Rob Brennan. Implementing the Draft W3C Semantic Sensor Network Ontology” IEEE International Conference on Pervasive Computing and Communications Workshops” 2010.196-201.

[4] P. Liu. Cloud computing [M]. Beijing: Electronic Industry Press, 2010.3.
刘鹏.云计算[M].北京：电子工业出版社，2010.3.

[5] Tom White. Hadoop: The Definitive Guide [M]. USA: O’Reilly, June 2009.

[6] Mohammad Farhan Husain, Pankil Doshi.Storage and Retrieval of Large RDF Graph Using Hadoop and MapReduce” Cloud Computing - First International Conference” 2009.681-686.

[7] Jianling Sun，Qiang Jin. Scalable RDF Store Based on HBase and MapReduce2010 3rd International Conference on Advanced Computer Theory and Engineering (ICACTE)” 2010.680-685.

[8] Yu Wei, Chen Junpeng. Ontology Management in Cloud Computing” 201O Second International Conference on Computational Intelligence and Natural Computing (CINC)” 2010, 57-59.

978-1-935068-76-1 © 2011 SciRes.

Knowledge Service Outsourcing: Formation, Implementation and Risks

Chigang LOU[1], Peng CHENG[2]

[1]*Information Management Department, Central China Normal University, Wuhan, China*

[2]*Hubei Provincial Department of Science and Technology, China*

Abstract: Knowledge service outsourcing is the way to expand knowledge service providers' niche and avoid the intense competition. This article explores the formation, implementation and risks of knowledge service outsourcing. The outsourcing of knowledge service providers constantly changes their own specific information niche to gain a symbiotic win-win situation.

Keywords: knowledge service; outsourcing; implementation

1 Introduction

Knowledge service providers are expanding the information niche with the development of knowledge customers' needs. They create opportunities for acquiring more knowledge and continue to absorb information from the environment. However, while the type and quantity of knowledge resources utilized by the knowledge service providers are increasing, there is a certain stage of this expansion that will inevitably lead to have a niche overlap resulting in the knowledge resource competition. In order to obtain more customers, knowledge service providers have a policy to have some knowledge service outsourcing to reduce the loss during the business expansion. In the information environment, the outsourcing of knowledge service providers constantly changes their own specific information niche to gain a symbiotic win-win situation.

Considering the cost savings and efficiency factors, knowledge service providers intend to contract with external specialized agencies. Outsourcing provides the opportunity to re-adjust knowledge service providers' niche, re-configure their resources to build their own comparative advantage and focus on developing the core business. While their non-core businesses transferred out, knowledge service providers could compensate and improve their weakness, and enhance their knowledge service for the rapid response capacity of the environment.

2 Knowledge Service Outsourcing Formation

In order to meet the needs of knowledge users, knowledge service providers always want to possess more knowledge resources. But the scarcity of knowledge causes the knowledge niche overlap with others. With the development of businesses, knowledge service providers try to narrow the niche overlap and input more resources to extend special knowledge niche.

There is three periods for the outsourcing formation of knowledge service providers[1] as can be seen in Figure 1: a) the knowledge expansion period of niche breadth; b) unsuitable period for knowledge niche; c) specific knowledge niche development period.

2.1 Knowledge Expansion Period of Niche Breadth

In this period, knowledge service providers tend to maximize the width of their niche to possess more knowledge resources. They invest human resource, financial resource and material resource. Figure 2 shows that the knowledge types and quantity of knowledge service provider A are greatly increased in this period.

The knowledge services rely on using a variety of information technology, mining and integration of knowledge resources to achieve an effective value of knowledge as much as possible. The knowledge niche space occupied by them is abundant at this time.

2.2 Unsuitable Period of Knowledge Niche

With the increasing scale of knowledge type and quantity is becoming more obvious, knowledge services usually extend under the circumstances. A certain gap between the needs of knowledge users and the knowledge niche shortage of knowledge service providers will eventually emerge. In order to meet the demand and diminish

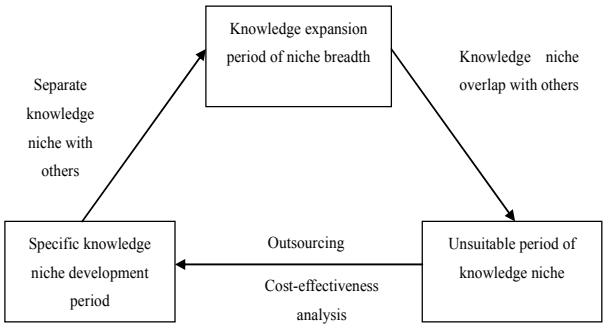


Figure 1. Knowledge service outsourcing formation periods

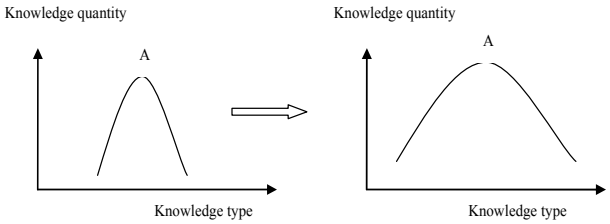


Figure 2. Knowledge expansion period of niche breadth

the imbalance, knowledge service provider often need to invest more money during the development. The cost will increase rapidly, and the effectiveness of knowledge services will descend. The knowledge niche becomes unsuitable. Figure 3 shows that knowledge service provider A has more knowledge niche overlaps with other providers B and C due to the scarcity of knowledge resources in this period. Knowledge service providers find it is difficult to independently provide all the knowledge resources for users, and can only play parts of role in knowledge service chain.

2.3 Specific Knowledge Niche Development Period

Under the pressure to reduce the cost and improve the efficiency, knowledge service providers realize that one way to gain more beneficial result is to narrow the knowledge niche overlap and to improve usage and efficiency of a certain type of knowledge resources. Figure 4 shows that the specific knowledge niche development transfers knowledge niche space to avoid the fierce competition in this period. Separation from others knowledge niche can reduce the overlap and develop special knowledge resources that strengthen knowledge service providers' core competitiveness.

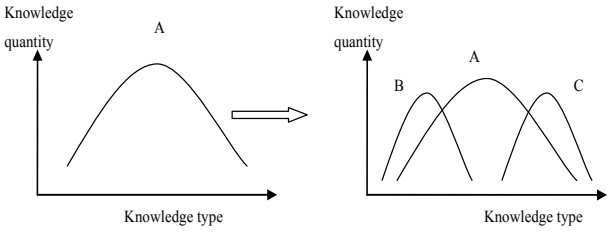


Figure 3. Unsuitable period of knowledge niche

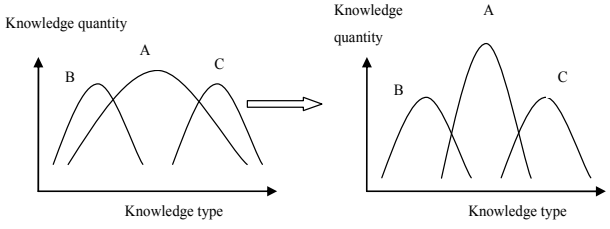


Figure 4. Specific knowledge niche development period

Outsourcing is the result of specific knowledge niche expansion of knowledge service providers, and it is a choice after they balance the cost and efficiency of knowledge development. Different knowledge service providers rely on their unique knowledge resources and occupy difference knowledge niche of businesses, resulting in the balance of knowledge service ecology. This relatively stable state help to reduce the waste of knowledge resources temporarily and improve the level of knowledge service providers' operation and benefits.

The three periods are not separate while knowledge service providers pursue outsourcing, but interwoven and interdependent. Knowledge expansion of niche breadth is throughout all periods, but the expansion region of niche breadth is different. When knowledge type and quantity of knowledge service providers are not meet the requirements of the knowledge customers, they expand related knowledge of development strategies, as well as all kind of knowledge. When knowledge service providers develop specific knowledge niche, the non-overlapping regions of knowledge niche width are expanded. A certain amount of knowledge service outsourcing helps differentiate development and encourages organization to focus on core business.

3 Implementation of Knowledge Service Outsourcing

When knowledge service providers' niche is unsuitable for the environment, they can choose service outsourcing. In general, the primary factor of the assessment in outsourcing is cost accounting. Only if the costs for providing the service are greater than outsourcing costs, the outsourcing decision can be made to expand their knowledge niche[2]. As shown in Figure 5, there are three stages for implementation of outsourcing: outsourcing preparing stage, outsourcing contract signing and implementation stage and outsourcing last stage.

Knowledge service providers consider a series of stages while intend to have parts of services outsourcing. There are also outsourcing process during each stage, specifically they are:

a) Determining outsourcing demands

Before outsourcing, evaluation of knowledge services should be planned to determine which service will be outsourced. The contents of evaluation include information technology, financial status, knowledge needs and risks.

b) Making plans and strategies for outsourcing

The decision of outsourcing should be ensure that the strategic plan and objective are identical. The strategies of outsourcing include goals to outsourcing, outsourcing objects, the scope of outsourcing, outsourcing timetable, costs, model and process. Successful plan and strategy choice help knowledge service providers to avoid negative effects of outsourcing and promote knowledge strategy development implementation.

978-1-935068-76-1 © 2011 SciRes.

c) Outsourcing service provider selection

Outsourcing service provider should provide sufficient information to ensure that the knowledge needs and expectations can be met in the outsourcing process. After acquiring information, knowledge service providers select one of potential outsourcing service providers according to strategic plan and knowledge needs.

d) Signing a contract of outsourcing

After contract negotiation finishes, an outsourcing contract will be signed. The contract is critical for the success for outsourcing because it covers the various detail aspects of outsourcing including the basic content, management, technical requirements, price, financial and legal norms of outsourcing. In order to avoid irregularities in the contract, the contract should be through both sides legal advice audit. Moreover, the service level agreement must be clear to ensure the quality of outsourcing.

e) Outsourcing transition process

Outsourcing transition process is the key process between the establishment of outsourcing relationship and implement of the outsourcing[3]. This process is a complex phase which involves outsourcing related knowledge transformation and knowledge service providers human resource relocating. Typically the scope of outsourcing transition includes: personnel management, exchange of knowledge, knowledge management and quality control. Effective outsourcing transition process needs knowledge service provider have human resource knowledge, communication skills, knowledge transfer and quality control capabilities.

f) Outsourcing project implementation

After the outsourcing transition process, the knowledge service outsourcing starts to implement. This process have four different parts: outsourcing project management, outsourcing relationship management, outsourcing change management and risk management[4]. The relationship management and contract change management should be paid more attention in order to meet the needs of service legal agreement.

g) Outsourcing project evaluation

When the outsourcing project is finished, knowledge service providers ask outsourcing provider to submit the final knowledge products or services. Knowledge service provider should evaluate the outsourcing products or services based on their own assessment of knowledge needs and the outsourcing contract. They compare the quality of outsourcing products or services and costs paid. Stable outsourcing relationship depends on the satisfaction of project appraisal and the future cooperation will be expected.

In the process of knowledge service outsourcing, knowledge needs of knowledge service providers may not be static. In the stage of outsourcing contract signing and implementation and outsourcing last stage, the change of knowledge environment or knowledge customers' needs may lead to update the contract of knowledge outsourcing[5]. Therefore, some adjustment of contract will adapt the outcome of outsourcing. This change makes knowledge service providers' niche more suitable to the knowledge environment.

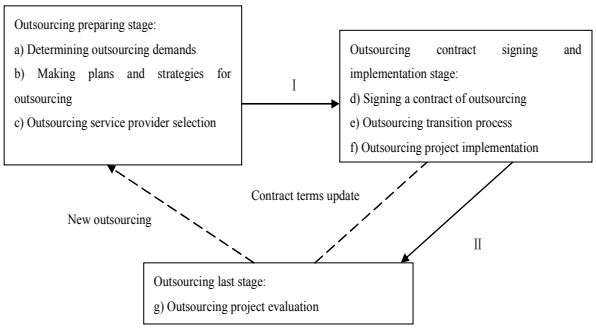


Figure 5. Implementation stages of knowledge service outsourcing

4 Risks in Knowledge Service Outsourcing

Risks in knowledge service outsourcing can be seen as information failure and are not expected by knowledge service providers, but still will occur on a certain probability. Risks can be endogenous or exogenous. Exogenous risk is the risk that cannot be controlled by knowledge service providers, such as natural disasters, the level of technology development. Endogenous risk is the risk that results from knowledge service providers' strategy, decisions or behaviors. Knowledge service providers should control the endogenous risks in outsourcing.

The risks in outsourcing last stage lie in the choice of outsourcing assessment tools. Knowledge outsourcing risks are unexpected loss. The size of loss depends on probability and size of information failure. Each risk exists by a certain probability. Total risk of knowledge service outsourcing can be calculated as following:

$RO = \sum_i P(RP_i) \times D(RP_i)$, $P(RP_i)$ is the probability of the risk i, $D(RP_i)$ is the loss when the risk i happens.

Each stage of outsourcing may have a series of risks that brought by the judgment and selection. Risks in outsourcing preparing stage focus on collecting information related to outsourcing. Risks in outsourcing contract signing and implementation stage focus on lacking of outsourcing management knowledge and tools, as well as quality control process. Risks in outsourcing last stage focus on the choice of outsourcing quality assessment. Risks in outsourcing are the expected loss which depends on probability of failure and its scale. Specifically outsourcing risks of each stage are:

a) Risks in outsourcing preparing stage: lacking of market information; unfamiliar with outsourcing plans and strategies; lacking of outsourcing provider information. Due to asymmetric information, knowledge service

providers cannot attain all information of outsourcing providers. They can only make decisions based on the information collected from outside environment.

b) Risks in outsourcing contract signing and implementation stage: lacking of outsourcing tools; quality control process failure; costs related risks: unexpected transition and management expenses, knowledge switching costs, contract amendments expenses, disputes and legal costs; outsourcing service providers related risks: breach of contract costs, hidden costs such as lack of communication; security control risks: key knowledge resource controlling failure, knowledge autonomy risks, etc.

c) Risks in outsourcing last stage: lacking of assessment indicators and tools; lacking of assessment quality certification model. If knowledge service providers cannot get enough accurate assessment about service outsourcing, they may find it difficult to develop next outsourcing planning and strategies.

5 Conclusion

The outsourcing of knowledge service providers is the result of adaption to knowledge environment. Knowledge service providers reduce the extra cost to expand the overlap knowledge niche with others, and focus on the development of non-overlapping areas. Outsourcing can alleviate the level of competition and make the knowledge ecological environment more harmonious. However, the risks of knowledge service outsourcing cannot be ignored. The losses can be minimized by outsourcing risk control system.

This article provides theoretical analysis of outsourcing in knowledge service providers including formation, implementation and risks. Knowledge service providers may use the ideas to establish their core competencies and develop their special niche. The future research work could extend grounds of knowledge service outsourcing and the risk management mechanism to effectively prevent knowledge service business loss. How to identify the probability and loss to calculate outsourcing risks is the possible issue that can be solved in the future research.

References

[1] F.S. Santana, M.F. de Siqueira, A.M. Saraiva, and P.L.P. Correa, "A Reference Business Process For Ecological Niche Modeling," Ecological Informatics, Vol. 3, Issue 1, pp. 75-86, January 2008.

[2] Mark Beasley, Marianne Bradford, and Bruce Dehning, "The Value Impact Of Strategic Intent On Firms Engaged In Information Systems Outsourcing," International Journal of Accounting Information Systems, Vol. 10, Issue 2, pp. 79-96, June 2009.

[3] M.Robinson, R.Kalaota, and S.Sharma, Global Out-sourcing. Alpharetta, GA: Mivar Press, Inc., 2005.

[4] B.C. Adeleye, F. Annansingh, and M.B. Nunes, "Risk Management Practices in IS Outsourcing: an Investigation into Commercial Banks in Nigeria," International Journal of Information Management, Vol. 24, Issue 2, pp. 167-180, April 2004.

[5] S. Domberger, P. Fernandez, and D.G. Fiebig, "Modelling the Price, Performance and Contract Characteristics of IT Outsourcing," Journal of Information Technology, Vol. 15, Issue 2, pp. 107–118, February 2000.

Research on Reliability of Knowledge-updating in Network Knowledge Resource Management System

Zhe YIN[1,2], Yunfei GUO[1], Maosheng LAI[2]
[1]*Mathematics Department, Yanbian University, Yanji, China*
[2]*Department of Information Management, Peking University, Beijing, China*
Email: yinzhe@ybu.edu.cn

Abstract: The definition of reliability of knowledge-updating in network knowledge resource management system (NKRM system for brief) has been given in this paper, which has important reference value to the evaluation of NKRM system. Besides, two kinds of models in NKRM system have been introduced. And one of the most common methods to improve reliability is the redundant technique. However, reliability and efficiency are inconsistent each other. By treating each knowledge as one part of the system, this paper proposes method of joint reserve of knowledge that may resolve this problem.

Keywords: reliability; redundant; knowledge resource management; joint reserve

1 Introduction

Research on reliability has become more and more important in resent years. Cohen[1] divided needs into urgent needs and ordinary ones. Moore[2] did the classification according the functions of spare parts. Because of the influence of spare parts to manufacture and economy, many scholars have studied the amounts of the spare parts needed. P.Flint[3] provided the advice that we should develop the fellowship and the resource sharing to reduce the cycle time. Besides, Foote[4] studied stocks prediction, and Luxhoj and Rizzo[5] obtained the method of amounts of spare parts needed of the same set based on the set model. Kamath[6] used the Bayesian method to predict the amounts of spare parts needed.NKRM system has been introduced in this paper, which is shown as follows:

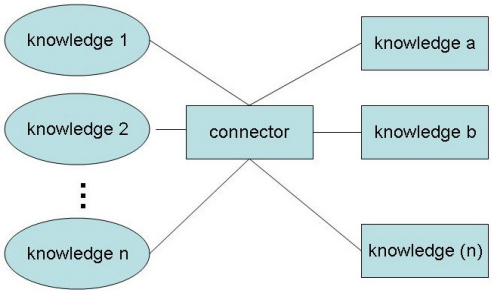


Figure 1. Network knowledge resource management system

Network knowledge resource management (NKRM) system is a system that can collect, process and organize a knowledge of all the information systems, which usually has a computer system supporting.

Method of joint reserve of knowledge in NKRM system has been proposed to improve the reliability and efficiency.

2 Self-updated Model and Calculation of Its Demand of Knowledge

2.1 Reliability of Knowledge-updating

Probability that the percentage of amounts of knowledge-updating are more than a certain value, that is: $A(x)=P(X>x)$, where X denotes the percentage of amounts of knowledge-updating.

Assume that reserve of knowledge of a certain new NKRM system starts from the zero moment, assessment will be done every $a(a>0)$. If the management system is normal(that is , all the knowledge are the newest), we will continue use these knowledge; otherwise, we will do self-updating to the knowledge, that is, after the temporary self-updating which cost time $b(0<b<a)$, management system recovers "normal" and proceeds reserve of knowledge, and this kind of model is called self-updated model, which is as follows:

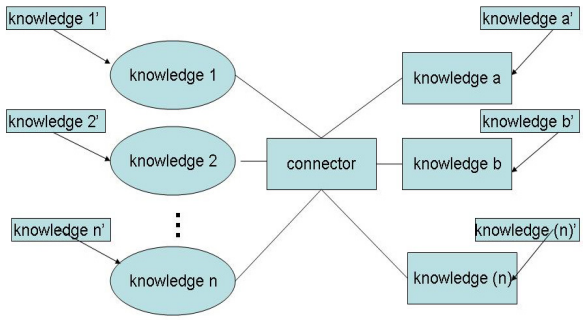


Figure 2. Self-updated model

For briefly expressing, state random variable X_t is introduced:

$$X t = \begin{cases} 1, & \textit{system is normal at time } t \\ 0, & \textit{system fails at time } t \end{cases}$$

Let $F(t)$ be the distribution function of random variable Z which is the first failure time of NKRM system. From the assumption of the model we can know, when $b = 0$, availability of system at time t (probability of system is in the normal state at time t) $A(t) = P(X_t = 1) = 1 - F(t)$; when $b > 0$, note $a_k = k \cdot a (k = 0,1,.....)$, $b_k = k \cdot a + b (k = 1,2,.....)$, $A(t)$ satisfies these formulas according to the model assumption:

$$\textit{when } a_0 < t \le a_1, A(t) = 1 - F(t) \tag{1}$$

$$\textit{when } a_k < t \le b_k (k > 1)$$

$$A(t) = P(X_t = 1 \mid X_{ak} = 1) P(X_{ak} = 1) = \frac{1 - F(t)}{1 - F(a_k)} A(a_k) \tag{2}$$

$$\textit{when } b_k < t \le a_{k+1} (k \ge 1)$$

$$\begin{aligned} A(t) &= P(X_t = 1, X_{ak} = 0) + P(X_t = 1, X_{ak} = 1) \\ &= P(X_t = 1 \mid X_{ak} = 0) P(X_{ak} = 0) + P(X_t = 1 \mid X_{ak} = 1) P(X_{ak} = 1) \\ &= P(X_t = 1 \mid X_{bk} = 1) P(X_{ak} = 0) + P(X_t = 1 \mid X_{ak} = 1) P(X_{ak} = 1) \\ &= \frac{1 - F(t)}{1 - F(b_k)} (1 - A(a_k)) + \frac{1 - F(t)}{1 - F(a_k)} A(a_k) \end{aligned} \tag{3}$$

From (2) and (3) we can see that $A(t)$ can be obtained once we calculate all the $A(a_k), k = 1,2,...,[t/a]^*$. Let $t = a_{k+1}$ in (3), when $k \ge 1$, we have:

$$A(a_{k+1}) = \frac{1 - F(a_{k+1})}{1 - F(b_k)} + \left[\frac{1 - F(a_{k+1})}{1 - F(a_k)} - \frac{1 - F(a_{k+1})}{1 - F(b_k)} \right] A(a_k)$$

According to the formula above and $A(a_1) = 1 - F(a_1)$, we can calculate $A(a_k), k = 1,2,...,[t/a]^*$, and then $A(t)$ can be obtained.

2.2 Calculation of Demand of Knowledge

Suppose there are N systems starting the reserve of knowledge from time $t = 0$ at the same time on the same condition, besides, the state of these N systems is independent each other. Then minimum of probability that there are at least N normal regional sub-systems at the given moment during the reserve period is not less than P_0 is:

$$M = \min \left\{ m; \sum_{j=N}^{N+m} \binom{N+m}{j} A(t)^j (1 - A(t))^{N+m-j} \ge P0, 0 \le t \le T0 \right\} \tag{4}$$

M is just the amount of the reserve of knowledge we needed.

3 Knowledge-introduced Model

The basic assumption is similar with the self-updated model, the difference is that the updating process will be finished by the introducing other knowledge from else-where when the knowledge should be updated, which is shown as follows:

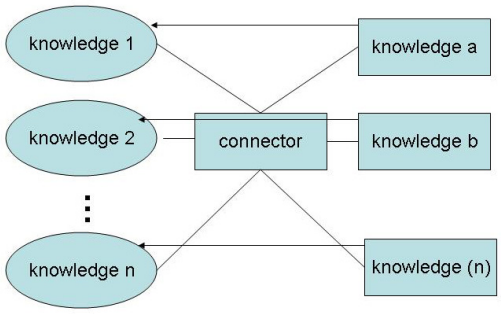


Figure 3. Knowledge-introduced model

4 Calculation of Amounts of Joint Reserve of Knowledge Needed

Generally, a knowledge updating of NKRM system in different regions is carried out independently. However, considering the problem of reducing the costs as possible as we can on the condition that the joint reserve of knowledge model is satisfied, we propose a new reserve way – joint reserve of knowledge, that is the method that

978-1-935068-76-1 © 2011 SciRes.

amounts of joint reserve of knowledge needed are calculated uniformly after determining sum of NKRM system in each region. Assume that there are n systems which belongs to similar areas, and once there is a new knowledge such as knowledge 1', knowledge 2', …, knowledge n', these similar systems can update the new knowledge at the same time, that is, the new knowledge can be shared in these similar systems, which is as show as figure 4:

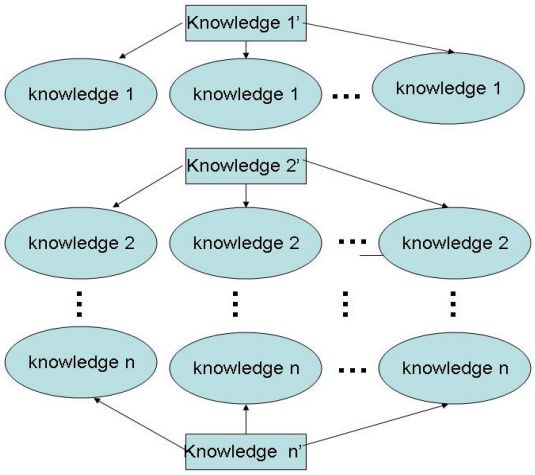


Figure 4. Joint reserve of self-updated model

In the "self-updated" model, suppose there are N_1 systems in region A, starting the reserve of self-updating from time $t=0$ at the same time on the same condition, besides, the state of these N_1 systems is independent each other. Then minimum of probability that there are at least N_1 normal systems during the reserve period isn't less than P_0 is:

$$M1=\min\left\{m1;\sum_{k=N1}^{N1+m1}\binom{N1+m1}{k}A(t)^k(1-A(t))^{N_1+m_1-k}\geq P0,0\leq t\leq T0\right\}$$

At the same time, there are N_2 systems meeting "self-updated" model, the conditions are the same as region A, then amounts of reserve of knowledge needed-- M_2 is:

$$M2=\min\left\{m2;\sum_{k=N2}^{N2+m2}\binom{N2+m2}{k}A(t)^k(1-A(t))^{N_2+m_2-k}\geq P0,0\leq t\leq T0\right\}$$

According the idea of joint reserve of knowledge, sum of these N_1+N_2 systems can be treated as N joint reserve of NKRM systems, then amounts of knowledge needed-- M is:

$$M=\min\left\{m;\sum_{k=N}^{N+m}\binom{N+m}{k}A(t)^k(1-A(t))^{N+m-k}\geq P0,0\leq t\leq T0\right\}$$

5 Calculation Instances

5.1 Case 1

Suppose there are 10 ($N=10$) same NKRM systems, which obey Weibull distribution-- $w(m,\eta)$, where examination time interval $a=1.0$, training time $b=0.2$, repair rate $P_0=0.9$ and predetermined period of reserve of knowledge $T=20.0$. In the "self-updated" model, taking some groups of parameter to simulate, results are shown as Table 1:

Table 1. Simulation result of "self-updated" model

m	η	T	N_1	N_2	N	M_1	M_2	M	Confidence level $1-\alpha$
1.5	30.	20.	4	6	10	6	8	12	0.9
1.5	30.0	10.0	5	5	10	3	3	4	0.9
1.5	30.0	10.0	4	6	10	2	3	4	0.9
1.0	25.0	20.0	5	5	10	11	11	19	0.9
1.0	25.0	20.0	4	6	10	9	13	19	0.9

From Table 1 above, we can see $M<M_1+M_2$.

5.2 Case 2

Next, simulation of "knowledge-introduced" model is also carried out on the same condition, whose results are show as Table 2 below:

Table 2. Simulation result of "knowledge-introduced" model

m	η	T	N_1	N_2	N	M_1	M_2	M	Confidence level $1-\alpha$
2.0	30.0	20.0	4	6	10	4	6	9	0.9
2.0	25.0	20.0	5	5	10	8	8	14	0.9
2.0	25.0	20.0	4	6	10	7	10	14	0.9
1.0	20.0	10.0	5	5	10	6	6	11	0.9
1.0	20.0	10.0	4	6	10	5	7	11	0.9

From Table 2 we can also obtain that $M<M_1+M_2$.

6 Conclusions

Through treating each knowledge as one part of the system, this paper proposes method of joint reserve of knowledge that can improve reliability and efficiency. Besides, we can get the conclusion that amounts of knowledge needed of joint model are less than those of model in which knowledge are updated or introduced separately, which shows the importance of the knowledge-sharing.

References

[1] Cohen M.A., P.R.Kleindorfer, H.Lee, etal, "Multi-item service constrained(s, S) policy for spare parts logistics system," Naval Research Logistics. Vol. 39, pp. 561-577, 1992.

[2] Moore R, "Establishing an inventory management program," Plant Engineering. Vol. 50, pp. 113-116, 1996.

[3] P.Flint, 'Too much of a good thing: Better inventory management could save the industry millions while improving reliability," Air Transport World. Vol. 32, pp. 103-106, 1995.

[4] B.Foote, "On the implementation of a control-based forecasting system for air-craft spare parts procurement," IIE Transactions. Vol. 27, pp. 210-216, 1995.

[5] J.T.Luxhoj, T.P.Rizzo, "Probabilistic spaces provisioning for repairable population models ," Journal of Business Logistics. Vol. 9, pp. 95-117,1988.

[6] Kamath K. Rajashree, T.P.M. Pakkala, "A Bayesian approach to a dynamic inventory model under an unknown demand distribution," Computers & Operations Research. Vol. 29, pp. 403-422, 2002.

[7] Pradhan D K, "Roll—forward check pointing scheme：a Novel Fault tolerant Architecture," IEEE Trans. Vol. 43, pp. 1163-1174, 1994.

978-1-935068-76-1 © 2011 SciRes.

Information Resources Development towards Knowledge Service

Developing and Using the Patent Information of Enterprises under the Knowledge Economy

Jie YUN

Institute of Science and Technology Information of Jiangsu Province, Nanjing, Jiangsu, China

Email: yunj@sti.js.cn

Abstract: This paper introduces the basic concepts of patent information, discusses the relationship between knowledge-based economy and the development of patent information. And it also expatiates the important effect development and utilization of the patent information have exerted in the technology innovation and patent strategy. What's more, it analyzes its technological, economic and legal value, and then investigates typical cases on how enterprises towards knowledge service develop and utilize patent information.

Keywords: patent information; knowledge-based economy; development; technology innovation; knowledge service; patent strategy

1 Introduction

Knowledge economy are featured as knowledge, while it is still a new economic model which builds on knowledge and information in the form of production, distribution and application. Its essence is the intellectual property rights, which are centered on patents and plays an important role in the knowledge economy. In the 21st century, mankind is moving to a new information era so that knowledge economy is bound to dominate the international economy. In this situation, accelerating the development and utilization of patent information, promoting the integration of knowledge and economy, researching on the importance of development and utilization of patent information for an enterprise under the knowledge economy, implementing the enterprise patent strategy and how Chinese enterprises enter into the international market and maintain their core- competitiveness have become China's important topic.

2 The Basic Concepts of Patent Information

Every technological achievement mankind made and has become a wisdom treasure for inventions. On the basis of patent documentation or patent literature, it can form into varied patent-related information through the means of decomposition, processing, indexing, statistics, analysis, integration and transformation. Patent information is characterized as technical, legal and economic. More than 90% of the world's latest inventions, quickly or initially spread out by patent documents [1]. According to World Intellectual Property Organization (WIPO) statistics, on average, companies can shorten 60% R & D time and saving 40% of R & D costs if effectively using patent information; among the average R & D outputs in the world, the economic value of patents was more than 90% compared with other activities[2]. Therefore, without development and utilization of patent information, the effective functioning cannot be realized in the knowledge economy.

3 The Relationship between Development and Utilization of Patent and the Knowledge Economy

Development and utilization of patent information is the important material foundation for the information construction. First, the development of knowledge-based economy depends on the level of information, but also on the breadth and depth of development and utilization of information resources [3]. Development and utilization of patent information, is to conclude the technical information, legal information and economic information of patent information, by collection, collation, analysis and extraction of patent information, at the same time transfer the individual, irregular, fragmented patent information into the systematic, complete, valuable patent competitive intelligence. Then it provides decision basis for enterprise managers, with this can make clear technological innovation for enterprises, direct product development and patent strategy and competitive strategy for enterprises. Especially in today's knowledge economy, R&D of new technologies and new products are always closely related to patent applications, thus patent information has become an important information source for social and technological innovation and new product development. Full development and utilization of patent information, as well as the potential value hidden from patent information, have ensured competitive advantage of today's enterprises.

Second, information construction and knowledge economy are interdependent and complementary. Information is an important part of the knowledge economy, but also an important means of implementing knowledge-based economy. The core issue of information con-

struction is fully developed and utilized of the information resources, instead of simply listing messages. Under the background of knowledge economy, development and utilization of patent information is essential for the scientific and technological progress. For one thing, it promotes all-around development of human beings and social progress, fully develops and utilizes patent information resources, which will help people understands and masters more updated information, broadens their horizons and improves the technical innovation so as to promotes scientific and technological progress, and accelerates economic development and social civilization. For another other, it is helpful to enhance the core competitiveness of enterprises. Under the economic globalization, there will be increasingly fierce competition between countries, or between businesses in the fields of economy, science and technology; however, competition cannot do without the patent creation and usage capability. Particularly those major core patents with global and strategic features are what people fight for. So who dominates in the development and utilization of patent information resources, who will excel at independent innovation, and take the initiatives in the market competition.

4 Major Role in Development and Utilization of Patent Information for Enterprises

In recent years, as the government and enterprises take more attention on intellectual property rights, our country's patent applications have significantly increase , indicating independent innovation capability of enterprises obviously promote also, but we should also clearly see that R & D expenses consumed by more than 20,000 large and medium enterprises in China covers only 0.81% of sales accounts, only equal to one-tenth of the developed countries. China has 9.28 million registered enterprises, but the enterprises which owns real core technologies, or independent intellectual property rights take up around three ten-thousandths, 98.6 percent of the enterprises have never applied for a patent. As can be seen from the above data, China's enterprises need strengthen technological innovation capability as a whole. Therefore, full development and utilization of patent information resources, in-depth mining potential value hidden in the patent information, will greatly improve the ability of technological innovation, and promote the implementation of enterprise patent strategy, and further the strategic objectives of national intellectual property rights. At present, the major roles of development and utilization of patent information for China's enterprises are mainly reflected as the following:

4.1 Acknowledging the Dynamic Competition in the Industry by Full Use of Patent Information

By means of collecting all the patent information in a technical field, major competitors in this business can be determined after grouping and sorting the patentees or inventors; patent distribution maps can be drawn for categorizing patent information from all the major competitors in the industry, which can be assessed for its technical strength and R&D capabilities, therefore, the competitor's technology strategy and development direction can be accurately determined. By studying the implementation and the legal status of the rivals, and observing whether patents inter-related with utility model and design patents appear or not, areas of patent applications can be judged thereof, so that potential markets of these patents will be identified also. In terms of researching into patent citations of rivals, technology development strategy of competitors are readily evaluated, while activities and changes in technology innovation are easily understood.

4.2 Analysis and Tracking the Relevant Technical Field by Patent Information

Through the collection and analysis of patent information, the development process of a technology can be understood with the help of patent analysis tools, the new trends of this technology can be determined, future development path can be predicted, technology innovative ideas of personnel can be enlightened. Technical difficult issues in existing technologies can also be exploited for the second time, and new technology solutions can be further developed and become an exclusive intellectual property; through patent information analysis, life of a technology cycle can be accurately judged, clearly showing the changing process of emerging, growth, maturity, declining of a technology. Study on life cycle of a technology can help enterprises understand the trends in technology development, the new direction in technology development, and the possibility of whether this technology can be used in industrial production or it can form emerging markets or not, while it can forecast the next development path of the technology and provide decision making evidence for business managers. Through the analysis of a Patent-Classification-Matrix in a certain technical field, the R&D focus, difficulty and technical gaps can be determined, when combined with the actual situation of the enterprises targeted to determine the direction of business R&D investment to ensure that R&D has the right direction to achieve the ultimate goal of independent innovation.

4.3 Mapping out Patent Strategy by Full Use of Patent Information

Through the collection and analysis of patent information reflected in the actual situation, companies can formulate patent development strategy and competitive strategy for their own at an early date, and determine the direction of market development. Through the analysis and comparison of the technical information reflected in patent information, the latest development and changes in the technology at home and abroad can be acquired, and

978-1-935068-76-1 © 2011 SciRes.

innovative products leading domestic counterparts can be worked out by putting investment in R&D when compared to various programs of technological innovations and predicting the market outlook; by the analysis of the law information as reflected in patent information, enterprises are able to timely understand the legal status of domestic and foreign patents during the import and export trade of technology or products, avoiding patent barriers and blocking that foreign companies set up, and avoiding them to get into trouble of infringement litigation and huge compensation. Through the analysis of economic information reflected in patent information, the technical R&D focus of major competitors, the field of product distribution and market share intentions are easily acknowledged, so that it can offer enterprises with decision making for developing their own patent strategy and competitive strategy, just as the saying goes "Know ourselves, know yourself."

5 Useful Attempt in Development and Utilization of Patent Information

Development and utilization of patent information can supply enterprises with a wide range of support for technological innovation and economic activities, while understanding information retrieval and analysis skills under knowledge-economy situation has been the key to R&D staff and managers of enterprises. For instance, I have once received a large lawsuit case (worth more than 10 million RMB) commissioned by a company, which is a patent infringement case that the foreign patentees suited Chinese enterprises for its patent infringement. The case relates to a kind of production technology on corrugated pipe production line, in which the defendant is a professional manufacturer of such equipment, whose scale of investment alone reached 10 million RMB. Once the infringement was sentenced, the Chinese enterprise will be forced to stop production, facing bankruptcy, not only thousands RMB investment down the drain, but also to compensate all loss of foreign companies. If the defendant would reach the turning point in the proceedings, the only resolution was to find the existing technology the same with foreign companies, negating the novelty of its patent claims to make patent prosecution invalid, completely remove the crisis.

During the process of investigating this project, I listened carefully to technical presentations for the defense. After carefully analyze the patent documents, this case is known to be involved in the large industrial equipment and technology, which covers broadly, thus it is difficult to precisely define the technical theme. Even if using conventional means is still hard to achieve the purpose, information discovery methods and ideas must be applied, constantly adjusting and searching a breakthrough point for information retrieval and retroactive expansion of the new information element, in order to obtain breakthrough results. When relying on information resource platform domestically and internationally in Engineering Documentation Center in Jiangsu Province, I repeated many attempts to retrieve the theme by constantly adjustment and multi-faceted adoption, and retrieved with the logical group by multiple searches and entrances, through a number of orthogonal tests among international patent classification and multiple technology key words, in-depth and attentively comparing and analyzing various search results.

Finally in the vast ocean of information at home and abroad, I accurately detected a French company who once applied the earlier patent documents in the United States. Through careful reading that patent claims and comparing the patent claims of this case, I found that although the description of the two patents are in different ways but the essence of each technical characteristics shares similarity with each other. Thus it can be regarded as negating its novelty as X-Files, according to this we can support sufficient evidence to the SIPO Patent Reexamination Board charging that the patent the plaintiff made was invalid. Because I highlighted search results and directed at weak points of the vitals, the plaintiff party quickly withdrew its the prosecution of patent infringement, the commissioning party (Chinese enterprises) changed its status of a defendant to a plaintiff, and had access to a bright future for litigation. In this complicated case, the information discovery methods had been combined with patent information retrieval methods, efficiently developing and using patent information resources, avoiding enterprises to get into trouble in patent infringement litigation and a huge loss for compensation, safeguarding technological innovation of enterprises and economic activities, namely, winning abundant economic benefits.

6 Conclusion

As the most important scientific and technological resources under the current knowledge economy, the effective use of patent information has been positive factors to push economic and social development. Meet the challenges resulted from the knowledge economy, fully developing and utilizing patent information for serving the enterprises in technological innovation, as well as enhancing the ability of technological innovation and core competitiveness of enterprises, are practical problems needed to be resolved and commonly concerned by our government and enterprises.

References

[1] Jianrong Li. ZHUANLI XINXI YU LIYONG [M]. Intellectual Property Press, 2006.
[2] Jiangling Zhu, Jianzhong Zhou. Research on the role of patent information in the knowledge economy development. Library & Information Science Tribune, 2000(1):45-46.
[3] Yunlong Wang, Xueyun Zhao. Development of the library information network under knowledge economy era[J]. Modern Information, 2002(6): 137-138.

Development Research on Sci-Tech Information Sharing Service Based on Knowledge Discovery

——A Case Analysis of HBSTL

Zengzeng TANG[1], Nianjun FU[2], Fei XUE[2]

[1]*Huazhong Agricultural University Library*

[2]*Information Management Department of CCNU, Wuhan, China*

Email: tzz1182@mail.hzau.edu.cn

Abstract: Based on the typical case of HBSTL, the article analyzes the present status of sci-tech information sharing service and points out the necessity of applying knowledge discovery. And finally, according to the supply and demand theory in economics, from the perspectives of "supply" of sci-tech information resources, users' "demand" and the supply-demand balance, we explore the path of applying knowledge discovery for sci-tech information sharing services.

Keywords: knowledge discovery; sci-tech information; sharing service

The sci-tech information sharing service is endued with new significance in the age of knowledge economy, thus it is necessary and feasible to introduce the idea and methodology of knowledge discovery for the purposes: firstly, reflecting the values of the two concepts, i.e. knowledge discovery and sci-tech information sharing service; secondly, profoundly revealing the present status of sci-tech information sharing service based on knowledge discovery; thirdly, promoting the development of sci-tech information sharing service industry by knowledge discovery.

1 Present Status of Sci-Tech Information Sharing Service in Hubei Province

1.1 Overview

Table 1. Existing Sci-Tech Information Sharing Service Platforms in Hubei

Service Platforms & Providers	Target Users
CALIS Wuhan University Library	All universities in the region
CASHL Wuhan University Library	All universities in China
Hubei Standard Information Service Platform; Hubei Institute of Standardization	Social organization, enterprise and individuals
Patent Database Service Platform; Hubei Intellectual Property Office	Demander for patents; Establishing cities' service platform
CSTNet Wuhan Subcenter; Wuhan branch of the National Science Library, CAS	Base in central and southern, open to all Demanders in China
HBSTL; Information Management Department of CCNU	Base in central and southern , open to all Demanders in China

Among existing sci-tech information sharing service platforms in Hubei Province, HBSTL is organized by Hubei Science and Technology Information Research Institute[1] and jointly established by CALIS Central China Center (Wuhan University Library), Wuhan Library of National Science Library of Chinese Academy of Sciences, and CERNET Central China Center (Huazhong University of Science and Technology Library) etc. This article probes into the case of Hubei Science and Technology Information Sharing Service Platform (hereinafter referred to as "HBSTL"), the key science and technology infrastructure platform of Hubei Province.

1.2 Present Status of Core Business

The sci-tech information sharing service of HBSTL is a paid service open to any organizations and individuals with access to internet (Figure 1).

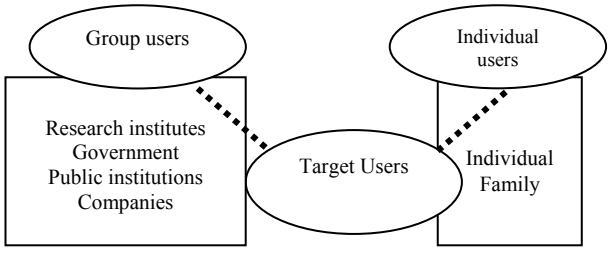


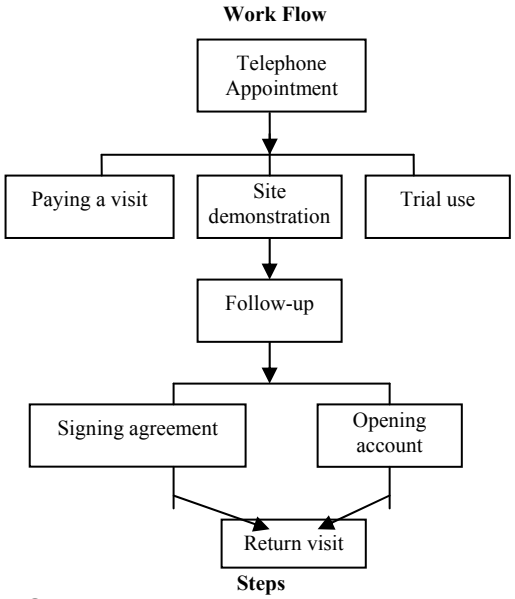
Figure 1. Target Users

Recently the Ministry of Culture and the Ministry of Finance jointly release the Opinions on Free Admission to Art Galleries, Public Libraries, Cultural Centers in China, requiring that before the end of 2011, all public libraries in China should admit free, all public facilities and places open free and all basic services are free to all citizens[2]. Free admission to public libraries is only a beginning which marks the coming of an age of free information

978-1-935068-76-1

service for common good. Then, serious consideration must be given to the prospect of core business of sci-tech information sharing services. Whether the existing service patterns can adapt to this new age?

Meanwhile, the science and technology information sharing service is at the start end of the user information chain without an organic relation with users, so that it has limited bearing on users. For example, the routine work of HBSTL includes managing the secondary service station sites, directly developing users and uniformly managing them, as presented in Figure 2.



Steps

①According to existing organization structure, select potential users and make telephone appointment

②Pay a visit, demonstrate the platform's function at site and clarify how to use it. Open a trial account for potential user to get familiar with the platform

③Further understand user's demand for specific database and answer the questions

④ Sign the agreement, make clear the rights and information services and open an account

⑤Follow up user's use of the platform and know the trend of demand for database promptly

Figure 2. Work Flow of Developing Users

In fact, the work flow of developing users and the services rendered by the platform indicate that it is a type of paid self-service. That is to say, resources are provided to users with the support of the platform. In this process, users are guided to get familiar with the functions of retrieval and reading on the platform. After the potential users become the current users, the focus is only on keeping up-to-date to demand for database, instead of involving in the critical process how users make use of the information resources provided on the platform.

2 Necessity of Knowledge Discovery

Development sees two possibilities. One is the adaptive development in the sense of biology, i.e. to change its own functions and service pattern with the change of environment, characterized by "change to follow changes", and it is a passive adaptation. Another is the development in the sense of sociology, not only "change to follow changes", passively following the change of environment, but also "change to promote changes", actively changing the environment. For either possibility or the combination of the two possibilities, "knowledge" always lies in the core of development.

2.1 Current Trend

The mist in the development of the science and technology information sharing service industry rests in a macro-environment. Hence, only with a new insight and investigation of the industry under the background of the new age, the development can be nudged in the right direction on the whole and continues to develop in the new age and new environment.

At present, Chinese economy is transitioning from agricultural economy into industrial economy and from industrial economy to knowledge economy. The tide of the knowledge economy brings about changes to the whole economic structure. Traditional factors of production become less important, but knowledge turns into the most creative and valuable core resources with increasingly larger contribution to organizations, thus being the source of the core competitive advantages of any organization. As pointed out by American scholar Peter F. Drucker, knowledge capital is the "third resource" in addition to financial capital and labor to facilitate the continuous development of enterprises. The tide of knowledge economy also radically changes the society in that people's pursuit, creation, possession and use of knowledge are totally unprecedented. 21st century is a knowledge-based century when the popularity and possession of knowledge reflect the trend of socialization. Thus, transition of sci-tech information sharing to "knowledge"-cored service is an inevitable trend.

Currently, information sharing service of all provinces in China bases itself on the database service industry and extends towards subject consultation and subject service etc. No doubt all these are pointed to knowledge, which is the impetus for sustainable development of sci-tech information sharing service. This is like a tree, of which products and services are branches and fruits, various factors are leaves, and the root is the provider all nutrients it needs.

It should be noted that, the shift of the focal point of the sci-tech information sharing service from information to knowledge doesn't necessarily mean the traditional information service is not as important as the database service industry, and yet it doesn't indicate that the information service is obsolete or out-dated, but that the information service needs to develop toward knowledge service and responds to new problems in the new age.

2.2 Market Force

The present development of sci-tech information sharing service industry is mainly driven by two powers: power of country, i.e. policies of information service; market power, including users and competitors. In case the policies and system are adjusted in stages and the welfare information service starts, the traditional sci-tech information service organizations will be divided, some take the opportunity to invigorate their original core in the new age and others decline due to difficulty in adaptation to changes.

Today's sci-tech information services are not confined to traditional sci-tech information organizations or universities, but distributed in industries and sectors, made up of sci-tech information agencies of governments and private and foreign information organizations. Each of these powers has its strong points in the market competition. As the market segments, the traditional sci-tech information providers gradually lose their battlefield. Currently the sci-tech information sharing service market hasn't undergone the all-around competition yet. For instance, the operation of HBSTL to a large extent benefits from existing policies and the advantageous position of its organizer Hubei Science and Technology Information Research Institute, which is a key information service provider constructed and supported during the seventh and eighth "Five-Year Plan" of China and the largest comprehensive sci-tech information organization in Hubei Province, or even in the whole central and southern China.

Relied upon policy support, naturally we can survive, but with an eye to users and competitors, we can achieve sustainable development.

From the viewpoint of the author, under the background of knowledge and network, sci-tech information sharing service witnesses four trends in user demand. Firstly, specialization, which means the information provided shall be typical of representative figures in the specialty realm and be at the forefront of this industry; secondly, broadness, which indicates the information shall not be limited to specific industries but highly comprehensive; thirdly, all-embracing, which refers to the all-around information solution obtained through analysis, screening, condensation, processing and integration; fourthly, who. The information requested by user and its source shall be checked, true and correct. The four trends cover the description of content of information demanded by users and users' requirements for form of information. At least these four points are satisfied, then the sci-tech information sharing service can effectively solve problems of users. Therefore, it is a problem of concern for sci-tech information sharing service providers to mine the mass data and values hidden behind information. This is the very problem knowledge discovery will solve.

3 Path of Knowledge Discovery Application

Sci-tech information sharing service is lack of fixed pattern and process, aiming at the individualized and high-end needs of users, integrating information resources, technical equipment and service capacity of sci-tech information providers, and improving the level of knowledge service based on the traditional information services, such as reference service, selective dissemination of information, and document delivery etc. As the basis and precondition of knowledge organization and knowledge service, knowledge discovery is a process of extracting implicit, useful, undiscovered and potentially valuable rules, information or knowledge [3]. In practice, this process needs to utilize the mining technology of information organization and analysis, through interaction with user, repeatedly explore the database or knowledge base to discover novel and useful intelligence or regularity, and after interpretation by humans, turn the discovered regularity information to be useful information or knowledge.

We explain the role of knowledge discovery in the sci-tech information sharing service industry by using the supply and demand theory in economics, as presented in Figure 3. Future sci-tech information sharing service market is highly competitive, and for winning the battle, the core competitive ability must be established from two aspects: externally, mining user's demand; internally, mining sci-tech literature information resources, thus ensuring the provision of high-quality information supply and securing both demand and supply. "Balanced supply and demand" is the state that economic theory attempts to achieve. For demand of specific users, the sci-tech information sharing service is more suitable for "matching supply with demand".

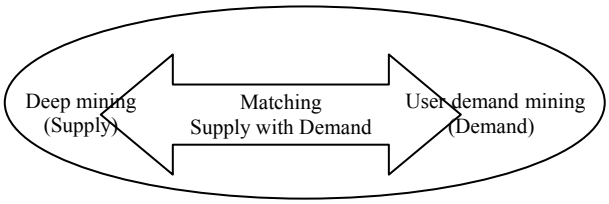


Figure 3. Role of Knowledge Discovery in Sci-Tech Information Sharing Service

3.1 Supply—Deep Mining of Sci-Tech Literature Information Resource

A Chinese saying goes like this: "if you know the enemy and know yourself, you need not fear the result of a hundred battles". For sci-tech information organizations, the deep mining of science and technology information resource makes up a part of the organizational knowledge construction, so the knowledge mining is not limited to "supply", but looks forward to prospect, with the knowledge construction of the whole organization in mind.

Deep mining of science and technology information resource should make clear statement of knowledge status

978-1-935068-76-1 © 2011 SciRes.

of the organization. Knowledge map is used to depict the knowledge and its distribution in the organization and constructs multiple forms of knowledge base on the basis of self-knowledge. In the case of HBSTL, the organizer of the platform, Hubei Science and Technology Information Research Institute has a multiple business chains in parallel and clarifies the business chains according to logical relation, inheritance and closeness between businesses. The information product center is responsible for the operation of the platform, but technologically relied on network and resource departments and intellectually dependent on the strategic planning department for concrete intelligence work. Around the sci-tech information sharing service business chain, the information of related departments are organized and incorporated, their knowledge systems are accumulated, and further the basic knowledge base is constructed with the ability to support deep mining of sci-tech literature information resources.

It is our principle to critically absorb and multi-directionally investigate various competing theories or several methods in the field of knowledge discovery. For solving specific problems, we are not confined to fixed method but lay stress on the effectiveness of the solution

3.2 Demand—Users and Demand Mining

In the knowledge-based context, the division of users and demand in a sense has the merits and significance of refinement and may highlight the direction of knowledge discovery. However, how to divide them? In respect of content, such as sci-tech demand and humanities and social science demand, this is a division of specialty; by time, it is divided into normal demand and temporary demand; by function, general demand and strategic demand.

The mining of user and demand utilizes the functions of classification and clustering of knowledge discovery to figure out the individual differences of users and reflect the similarities of users respectively.

Presently the individualized information of users of sci-tech information sharing service is acquired from two sources: personal information of users registered and submitted by users, explicit and easily available; behavior record information of user visiting the platform throughout the use of the platform, implicit. During information retrieval, according to keywords input by user, through machine learning, identify and predict user's search habits or preferences; with the data of user's visits and downloads, locate the interested specialty area of user by the cluster analysis of content; in the light of visit time, estimate the time distribution of user demand through the time series analysis; understand the function of user demand from the features of the data class of the association analysis; base on series data of user retrieval and use information to construct the vector model of user's individualized interest by the fuzzy set theory[4]. Follow up in a long term for perfection and adjustment. The clustering application mostly seeks out some common features of existing users from individualized data, e.g. common focal points of users at different stages, hereby estimates and predicts target user groups or analyzes and attracts users purchasing the service of other competitors, find out the problems of sci-tech information sharing service and put forward corresponding improvement measures.

The example of HBSTL vertically compares regional users, integrates and mines the economic and social development and volume and content of download by users in the past in secondary service station regions, clarifies whether the user demand in this region has the possibility of potential development, and determine the focus of future demand. It may also horizontally compare regional users and demand, have knowledge of time period and major fields of regional users and configure the best pattern for regional user development and management.

3.3 Matching Supply with Demand

The core of matching supply with demand is the heuristic and collaborative mechanism, not absolutely demand-oriented and not simply supply-first. If there must be an emphasis, it can be divided into directional and non-directional mining.

The directional mining is goal-driven. In the early stage of deep mining of sci-tech literature information resource, aiming at the explicit demand and interests of user, direct users to establish the vector model and individualized subject tree of information class. Users may at any time pay attention to their own individualized customization. This "directional mining" employs the user participation and apriori knowledge heuristic, effectively avoiding blindness of knowledge discovery.

The deep mining of sci-tech literature information resource is for the purpose of meeting users' need, but it doesn't mean being limited to demand only. "Non-directional mining" integrates the content mining and use mining of the sci-tech information resources and preliminarily construct the classification of user visit patterns and the prediction of users' interests. As for HBSTL, according to different demands, build a spatial vector model of users' interests and push the information and knowledge possibly to be mined by users to them. In virtue of users' use, the spatial vector model and the knowledge class vector model in the system will realize the self-organization evolution, automatically applying new vector models to periodically filter knowledge so as to gradually approach to users' demand.

Matching supply with demand guided by data mining is a kind of active push service and also reflects the trend of information service — individualized.

3.4 Employee Management

Technology and tool are certainly essential, but the core of the data mining is human. All supply/demand links of sci-tech information sharing service are closely

related to employees, as presented in Figure4. Employees are both the manager of user's explicit knowledge, the insider of implicit knowledge of user demand, and the mediator of matching supply with demand. The achievement of sci-tech information sharing service until now profits from employees' personal quality to a considerable degree, especially the morality. Since the present business pattern is simple and the service flow is mature, the resignation of employees has little impact on the business and consequently the employee management should also be simple. However, in the long run, simple management pattern will surely hold back the improvement of sci-tech information sharing service.

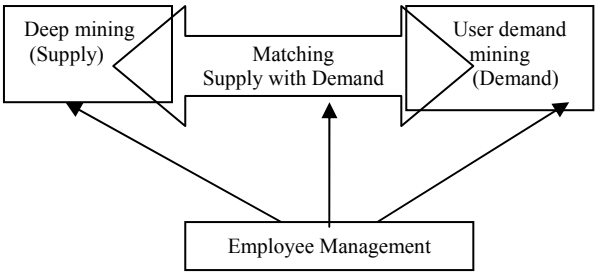


Figure 4. Necessity of Employee Management

Data mining can optimize the employee management. Firstly, clarify the requirements for employees' quality. Extract the moral qualities and regard them as professional ethics; the professional qualities are more important, including specialty knowledge and skills necessary for sci-tech information sharing work. The clear quality requirements not only can point out the clear direction for employees' efforts, but also are the intellectual guarantee for smooth data mining. Secondly, objectively evaluate employees' performance. Conduct knowledge management of employees, record real-time details of users [5], inclusive of explicit information, such as industry of user and database user selects, and the implicit knowledge, such as user's habit of retrieval and demand features. The related data can be used as source of data mining and the evaluation criteria of employees' job performance.

Traditional information and modern knowledge service are not mutually exclusive but embracing. By knowledge discovery, the challenge confronting with sci-tech information sharing service are identified and analyzed and various theoretical and practical problems are investigated and discussed in depth. Knowledge discovery itself is a developing process, so its application in sci-tech information sharing service will surely experience repeated exploration and practice.

References

[1] Information Management Department of CCNU. http://www.hbsti.ac.cn/index.jsp,2011-03-10

[2] Art Gallery、Public Libraries and Cultural Center in China will all Fully Cost-free and Opening.China daily. http://www.chinadaily.com.cn/hqgj/jryw/2011-02-10/content_1741201.html

[3] Li Yang, Jiang GuoRui. Two Important Technique of Knowledge Management:Data Miningand Knowledge Discovery. Journal of Information, 3rd ed ,2007,pp.53-55 (Ch).

[4] Wu, TK; Huang, SC; Meng, YR; Liang, WY; Lin, YC. Rough Sets as a Knowledge Discovery and Classification Tool for the Diagnosis of Students with Learning Disabilities. Internationali Journal of Computational Intelligence Systems,1st ed., vol. 4. 2011, pp.29-43.

[5] Nkambou, R; Fournier-Viger, P; Nguifo, EM. Learning task models in ill-defined domain using an hybrid knowledge discovery framework. [J]. Knowledge-Based Systems, 1st ed., vol. 24, 2011, pp.176-185.

978-1-935068-76-1

Evaluation of Policy Values Based on Document Context

A Comparative Analysis of MIIT and MOST Information Policies

Lei PEI[1], Jianjun SUN[2]

Information Management Department, Nanjing University, Nanjing, China

Email: plei@nju.edu.cn, sjj@nju.edu.cn

Abstract: Taking policy documents as materials, this paper advocated a quantitative analysis method by word frequency calculating. Creating a information policy value model TEOSC and its vocabulary, the paper took MIIT and MOST as examples, collected all information policies and divided the policy document into words and words, calculated the word frequency to mining potential rules. Then it mapped the words with value vocabulary, and examined the common values and differences. At last, it gave some advice to both information policy development and analysis methodology acceptance.

Keywords: policy values; document context; comparative analysis

1 Introduction

The general goal of policy research is to understand the effects, impacts, and outcomes of a policy within a greater political and social context (Jaeger, 2007). It can determine whether a policy and its components are effectively meeting programmatic, ethical, societal, economic, and intellectual goals. It also focused on improving the policy's performance and increasing its positive impact on society.

While, moved by technology innovation, the volume of information-related legislation and growing social awareness of information access (Trauth, 1986). As one of the important means to conduct and control information affairs, information policy has to meet strategy demands and social formulation. Social scholars, economists, lawyers, librarians, information experts and other disciplines' researchers have been spontaneously writing critical papers from different views, taking policy documents as important materials or tools. But they usually chose the policy as a whole, no matter whether the policy has revealed anything else. So we have to improve such analysis with the development of the following aspects.

First, the aim and methodology of policy research is more and more clear. In policy research, outcomes include providing information about performance, identifying and critiquing underlying goals and assumptions, and offering recommendations (Dunn, 1994).So researchers have to draw out the real meaning of the policy first of all, and introduce scientific methodology to assure the persuasion of their conclusions and suggestions. Gap analysis, comparative analysis, experts' comments or social value review, as main methodologies, were widely used in policy research, while quantitative and content-oriented methods rarely used. In 1979, Manifesto Research Group was formed in Berlin to analyze party manifestos within a common framework (Budge, Robertson, and Hearl 1987). Through several iterations of data coding efforts, the party manifesto data have been collected and expanded to most major political parties in 25 Western countries throughout the postwar period (Budge 1992; Budge et al 2001). From then on, many scholars begun to take advantage of these coding rules to analyses economic policies or cultural policies, while information policies were ignored. So it was necessary to take a comprehensive quantitative view of information policies.

Secondly, more and more information policies are being put out. Until now, hundreds of information policies have been put out and calculated to be databases or retrieval systems. In China, Legislative Affairs Office of State Council built a law full text database, which contains nearly 57393 documents over 20 years, besides of 281 information policy documents. And there are also many policy documents scattered in minister websites, according to the implementation of Government Information Disclosure Act. These documents can afford abundant materials for quantitative analysis.

For the third reason, in a broad view, with the arising of new technology environment and international attention of information polices, changing of policy contents, international comparison of information policy , information policy implementation and performance evaluation of resource allocation were hot topics in LIS research nowadays. Prior research highlights the fact that information policy values are very important to catch the macro demands of the era, but this work was ignored.

At last, LIS researchers have been widely using new methods to improve former theories recently, such as text mining, auto category, networks analysis, word frequency, word structure and words relationship between networks. These mature analysis methods and tools can greatly improve the using of quantitative analysis in policy documents research, while the application in information policy research is distinct limited.

In a word, a number of scholars discussed the information policy practices or theory scheme by or on policy documents, but only few researchers paid attention to information policy values. Content-oriented quantitative

analysis of information policy values are not only to introduce a new method, but also to pick up the focused values, to discover which values we have ignored or need to be strengthened..

2 Literature Review

2.1 Information Policy Values Category

Information policy exists since the early 16th century (Browne, 1997a). However, information technology using a modern approach was traced back to the 1960s and expanded in the early 1970s (Arnold, 2007). While governments first began to use the term 'information policy' around the late 1970s and the early 1980s. In 1976, a first formal statement of National Information Policy appeared in the USA. Therefore, this relatively new field faces a few problems such as no agreed definition (Browne, 1997a), various issues that form the information policy (Arnold, 2004), and rare attentions to information policy values.

Policy values, as Rein defined (Rein,1976), is comprise normative propositions that affirm what our public policy ought to be, and the normative and moral assumptions that underlie present practice. Guba suggested policy values included assumptions or basic beliefs underlying different paradigms of enquiry; theories or hypotheses describing or explaining phenomena; perspectives, such as the perspective of a particular discipline; social and cultural norms within a society or cultural group; personal or individual norms imposed by the individual on himself or herself which may or may not be the same as the dominant social and cultural norms(Guba,1984). Whatever kind of conceptions they defined, rare papers focused on information policy values. Browne,in his widely cited papers (Browne,1997a,1997b), giving an encyclopedic overview of the conception, the boundary and research methodology of information policy, gave quite positive comments on policy text reading and analysis, but he also sighed that:

In summary, there has been little analysis of the substance of the values underlying information policy, particularly to draw out the unique values, or combinations of values, that might characterize information policy.

But information policy dimensions and issues category was researched quite early. In 1974, Lamberton classified information policy issues into four groups: scientific and technical information, social science information, service information and needs towards information technology (Lamberton, 1974). In 1986, Trauth introduced an I-P-O model and built a information policy research framework, which contained the main policy values such as freedom of information, privacy, equality of access, software protection, computer crime, copyright and etc (Trauth, 1986). In 1990, Overman and Cahill provided a starting point for identifying values which remain constant regardless of social and political context (Overman and Cahill,1990). They checked 16 federal information policies and classified seven main values: (1) access and freedom; (2) privacy; (3) openness; (4) usefulness; (5) cost and benefit; (6) secrecy and security; (7) ownership.

In 1997, Browne introduced a value-critical method, accepting the idea of Overman and Cahill, as well as Webster who conceptualized Information Society from which underlying information policy values. They preferred information policy value compose of five types: (1) technological, which emphasizes the technical innovation and the convergence of telecommunications and computing; (2) economic, with its focus on the role of information in the broader economy; (3) occupational, which argues that contemporary occupations are predominantly focused on dealings with information; (4) spatial conceptions stress the linking of physical locations through the development of networks; (5) cultural, which emphasizes the extent of media and information pervasiveness in our daily lives (Browne, 1997a, 1997b).

In 2010,Yusof took a critical classification of information policy issues, lying out more than 40 items. He point out that information policy underlying values can be divided into 6 classes: (1) Technical and Scientific Information; (2) Library; (3) Information and communication technology (ICT); (4) Social issues; (5) Government information; (6) Economy(Yusof,2010). Yusof's approach to identifying issues, listing and classification is that used by Rowlands (1996) and Braman (2006). Clusters in information policy issues and their arrangements are based on frequency of use (Hernon and McClure, 1987) or content analysis..

In China, Jiang (1999), Ma (2003), Du (2005), Haiqun Ma (2007), Zhu (2002)and tens of information researchers built information frameworks, but most frameworks were organized by content or industry. Nobody focused on policy values category except Haiqun Ma (2007), but his opinion was limited to the dilemma of efficiency and fair.

2.2 Policy Category in Manifesto Research and Content Analysis

The Manifesto Research Group was formed in 1979 to analyze party manifestos in 19 Western countries comparatively and within a common framework (Budge, Robertson, and Hearl 1987). The project started at the beginning of the 1970s with the idea of comparing parties' programmatic strategies in Britain and the United States during the post-World War II period (Robertson 1976). Since then, the data collection has been continually updated for all new national elections and extended to 29 members of the Organization of Economic Cooperation and Development (OECD) and all member states of the European Union (EU). In addition, the project documented 23 Central and Eastern European (CEE) party systems. IThad used trained human coders to code 2,347 party manifestos

978-1-935068-76-1 © 2011 SciRes.

issued by 632 different parties in 52 countries over the postwar era (Volkens 2001).

Through several iterations of data coding efforts, the party manifesto data have been collected and expanded to most major political parties in 25 Western democracies throughout the postwar period (Budge 1992; Volkens 1995, and Budge et al 2001). Many scholars of political parties have begun to take advantage of these data. Of particular interest to some researchers are new measures of party ideology as summary measures of party preferences on a single ideological dimension (e.g., Laver and Budge 1993, Kim and Fording 1998, Gabel and Huber 2000, Laver and Garry 2000, and McDonald, Mendes, and Budge 2004). Recently Kim and Fording expanded party manifesto research by developing measures of voter and government ideology (Kim,1998, 2001b, 2002). The key method is "quasi-sentence" analysis and coding.

For better data and codes, the classification scheme of the Manifestos Project captures the whole content of manifestos issued for national elec-tions in a comprehensive, reliable, and efficient way (Budge/ Robertson/Hearl 1987; Budge et al. 2001; Klingemann et al. 2006; Volkens 2007). This classification covers 56 categories in seven policy domains. The categories specify general policy preferences towards specific issues that vary between parties, party systems, and over time. This widely used scheme divided policies into 7 categories: (1) external relationship;(2) freedom and democracy;(3) political system;(4) economy;(5) welfare and quality of life;(6) fabric of society;(7) social groups.

But this category was not exactly match the status of China's political atmosphere, so Chinese scholars build new categories and vocabulary for further study. Tu Duanwu, who study education policy in China, built a two-divided framework: material and visible values, mental and invisible values(Tu Duanwu,2010).Material values contains economy(经济) ,power and authority(权力),knowledge(知识),technology(技术)、 welfare(福利), while mental values contains terminology(专有称谓), horner(名誉)、 mental mechanism(意识形态)、 social goals (目标)。

Furthermore, Laver,Benoit and Lowe proposed a comprehensive method associated with computer science, linguist science and politics. Typical methods contain content analysis, event data extraction, wordscoring, experts investigation and etc.(Laver and Garry,2001). From a technical aspect, data or information acquisition was efficient, based on content analysis. Word frequency of raw texts were also used to index and classify, but the approach was not be proved efficient in policy documents analysis yet.

2.3 Document Subject Headings and GIR Catalog Systems

Document subject headings are used in e-government organization and digital document management. While Subject heading is a word or phrase from a controlled vocabulary which is used to describe the subject of a document or a class of documents. Subject headings may be represented in individual bibliographical records or they may just exists in classification systems or as separators in card catalogues(Wellisch, 1995).Wellisch writes that indexers before the advent of thesauri relied on lists of terms variously known as subject heading lists, keyword lists, term lists, and similarly labeled aids to consistency in indexing.

The latest Document subject headings of State Council was delivered in December 1997.The vocabulary consists of 15 classes, 1049 keywords, two main parts :sub-table and schedule .The main table has 13 categories of 751 keywords, sub-table has 2 categories of 298 keywords. Vocabulary is divided into three layers: domains, classifiers and generic words. In this vocabulary, information related subjects were not well defined, which scattered in several classifiers such as 02D telecommunication, 10C culture and etc..Culture Policy Library also delivered a vocabulary, and it contains a category named OC information industry, with 50 generic words. But the vocabulary can't cover all the fields of information policy.

Another achievement is the standardization of government information resources catalog system. In 2007, Chinese government published a series of standards to formulate the construction of e-government, which signed as SAC GB/T 21063:2007:Government Information Resource Catalog System. Especially SAC GB/T 21063-4:2007(Part 4: Government Information Resource Classification), has a very detailed classification system and coding rules, but it also can't be used to measure policy values.

3 Research Methods and Materials Prepare Your Paper before Styling

3.1 Definitions

Information policy. The governments' attitude, opinions or behaviors laid down by documents, such as *Laws and Regulations, Summaries of Laws and Regulations, Admin. Regulations, Legal Documents, Departments Rules and Judiciary Interpretations* and etc.

Information policy values. The common preferences and political beliefs of policy makers and other people involved in policy-making process, and the organizational objectives and policy orientation. According to the definition, it concludes assumptions or basic beliefs; theories or hypotheses describing or explaining phenomena; perspectives, such as the perspective of a particular discipline; social and cultural norms within a society or cultural group; personal or individual norms imposed by the individual on himself or herself which may or may not be the same as the dominant social and cultural norms.

Information policy values classification. A controlled vocabulary used to describe the subject of a policy paragraph, a policy document or a class of documents. It is that the vocabulary must be organized by levels.

3.2 Methodology

Qualitative design was chosen as being suitable to the needs of the present research, which is exploring issues and classes of information policy values which have not been fully developed (Creswell, 2009;Yusof,2010). The smart coding approach used here to identifying issues, listing and classification policy values according to relative academic backgrounds, used by Overman and Cahill, Tu Duanwu and MRG (Overman and Cahill, 1990;Tu Duanwu,2010) . Clusters in information policy issues and their arrangements are based on frequency of classifiers in document materials. From a technical aspect, data or information acquisition was based on content analysis or wordscoring as conducted by Laver, Benoit and Lowe (Laver and Garry, 2000; Laver, Benoit and Garry, 2002; Benoit and Laver, 2007).

The research covered four main steps:

Step1: collect all the information policy documents, and calculate the word frequency, classify and make a information policy values classification (IPVC);

Step2:Collect relative policy documents form the website of Ministry of Industry and Information Technology (MIIT)and the website of Ministry of Science and Technology(MOST), calculate the word frequency according to IPVC;

Step 3: Compare the two samples, find out the common values and their differences;

Step 4: According to IPVC, give some advice to information policy development.

3.3 Data Collection

Retrieved from Legislative Affairs Office of State Council full text database, we can get 281 documents (See Table 1), nearly 1,368,404words, using 4775 words or phrases. The published year covers between 1983 and 2011; most policy (75.4%) was delivered after 2000. The subjects were focused on information openness, information management, information resources development, information system security, telecommunication, e-government affairs, knowledge property and ownership, sharing, service and etc. This is the reference sample of this research.

Secondly, we login in the information portal of MIIT and MOST, and got 47 information policy documents in MIIT and 27 in MOST. Test set is described as Table 2.

4 Results and Content Analysis

4.1 Creation of Information Policy Values Classification

Table 1. Information documents sample sets

Documents 281	TotalBits 2,914,810	Total Words: 1,368,404	TotalParagraphs:16,893	Using Words: 4775
Time distribution	Before 1995	1996-2000	2001-2005	2006-2010
	17	52	105	107
Subject distribution（subjects crossing)	Openness	Management	Development	Security	Telecommunication
	47	14	33	37	52
	e-government	Informatization	Ownership & Sharing	Service	Others
	32	35	12	15	46

Table 2. In formation documents testing sets

	Documents	Total Bits	Total Words	Total Paragraphs	Using Words
MIIT	47	469,173	215,962	3,549	4,346
MOST	27	270,286	121,059	1,901	3,659

Remove the useless word in word frequency statistics table, such as “的”(36487 times), “第”(3633 times), “本”(2955times) and etc. Then we got 1963 words using more than 5 times. After coding these words by former classification (See Table 3, several words belong to none), we found it was not equivalent or balanced, which means these classifications are not exactly matching the practices of China nowadays. For an example, spatial information policy is seldom mentioned in Browne’s classifications. Therefore, we created a new model, choosing technological, economic, occupational, social, cultural and law affairs as domain words, which could be named TEOSC model.

Table 3. Words using according to several classifications

	Class	*Technological*	*Economic*	*Occupational*
Browne’s classification	Words	234	817	302
	Calculated	27,602	100,940	35,719
	Class	Spatial	Cultural	
	Words	61	152	
	Calculated	1,872	17,681	
Overman and Cahill’s classification	Class	Access and freedom	Privacy& ownership	Openness
	Words	105	24	61
	Calculated	4,957	1,892	13,417
	Class	Usefulness	Cost and benefit	Secrecy and security
	Words	180	123	54
	Calculated	6,644	2,429	2,797
Yusof’s classification	Class	Tech. & Sci.	Library	ICT
	Words	246	41	107
	Calculated	28,932	2,793	37,664
	Class	Social issues	Government information	Economy
	Words	152	121	817
	Calculated	17,681	27,392	100,940

978-1-935068-76-1 © 2011 SciRes.

Table 4. Chinese information policy values mapping to TEOSC

Technological	information security, information management, networks management, software, technology acceptance
Economic	information market management, telecommunication service, media industry, publishing, financial information, statistics
Occupational	information organization innovation, e-governme nt affairs, government information openness, information development, information resources management
Social	Informatization
Cultural&Law	knowledge property, cultural information sharing, library and archive management

4.2 Comparative Analysis

Based on TEOSC, we created a more detailed value vocabulary, which consists of 20 classifiers (See Table 4). Then we chose test sets, cut into word-by-word text, calculated the word frequency, we got the most frequent using words in two department (See Table 5). Then mapped all the words to TEOSC value vocabulary, coded and calculated, getting the picture of MIIT and MOST's policy value distribution (See Table 6).

Table 5. The most frequent using words/phrases in MIIT and MOST's information policies(TOP 20)

MIIT			MOST		
Words	Chinese	T	Words	Chinese	T
Telecommunication	电信	1577	Science&Technology	科学技术	680
Service	服务	1303	National	国家	568
Business	业务	1136	Ought to	应当	535
Ought to	应当	1064	Technology	技术	468
Management	管理	936	Science&Technology	科技	461
Regulation	规定	920	Regulation	规定	418
Communications	通信	879	Applying for	申请	414
Organization	机构	821	Patents	专利	391
Information Indusry	信息产业	698	Programs	项目	316
Department	部门	677	Management	管理	316
Electronic	电子	588	Knowledge	知识	287
Organization 2	单位	574	State Council	国务院	275
Internet	互联网	559	Property	产权	275
In charge of	主管	539	Department	部门	257
Operator	经营者	529	Organization	机构	249
Networks	网络	397	Research	研究	248
Domain name &URLs	域名	389	Organization	组织	241
Operation	经营	389	Organization	单位	236
Standard	标准	388	Trademarks	商标	231
National	国家	381	Plan	计划	229

Table 6. MIIT and MOST's value distribution according to TEOSC

	Technological	*Economic*	*Occupational*	*Social*	*Cultural&Law*
MIIT	*340*	*333*	*631*	*255*	*257*
	12,364	*12,585*	*18,314*	*6,096*	*9,867*
	20.88%	*21.25%*	*30.92%*	*10.29%*	*16.66%*
	Technological	*Economic*	*Occupational*	*Social*	*Cultural&Law*
MOST	243	*378*	*654*	*258*	*327*
	4913	*5483*	*9096*	*4571*	*6344*
	16.16%	18.03%	29.91%	15.03%	20.86%

In table 5, we found that some most frequent using words in two departments are in common, maybe these words representing the common value of information policy values. From table 6, we found that the number of coding words in two departments is nearly the same, which covers 41.8% and 50.8% of all the using words. In MIIT, 1816 words can be coded, and calculated to 59,226 times. In the five independent domains, occupational values took the biggest proportions, nearly 31%; social values least, only 10% and so on. In MOST, 1860 words can be coded, and calculated to 30,407 times. In the five independent domains, also occupational values covered the biggest proportions, nearly 30%; social values also least, only 15% and so on.

According to comparative analysis, we found out three common points .(1) The total words coding- able are nearly the same, about 1800 words, which means these 1800 words maybe covers the most values of information policies. Although they are not the same, these 2000 around words could be drawn out to comprise the generic words of TEOSC vocabulary.(2) The value structure of information policies is nearly the same, which is quite different with our pretention. Occupational values covered the most words, nearly 30%; social values least. It means that social values and public affairs should be strengthened in coming information policies.(3) In both test sets, we found many words such as "ought to", "organization", "regulation", "national " are in common, "information resources" and "informatization" are also high frequent using words, more than 100 times. These common oriented values may drive the two department collaborated to solve several political troubles.

At the same time, we found some details were different. For an example, their attitudes to social and cultural values were quite different. MIIT paid more attention to economic values and less to social and cultural values, which MOST more to cultural and less to economic and technical values relatively.

5 Conclusion and Discussion

Data may tell us a more real story about information policies, while researchers always run after the changing data. The job we did in this paper, tried to demonstrate two important opinions:

First, policy documents were not single value oriented, each policy may contain both or all the technical, eco-

nomical, occupational, social and cultural values, we should clarify these values before we commit on it. What we did in the past was exactly that we got the main value of a policy and abandoned the less important values.

Second, word frequency analysis and value evaluation in policy documents may find another way to find the common value or consensus in informatization affairs between different departments in China. Hundreds of thousands of policies have been published or issued these days, some may have conflicts. Value evaluation was the suggestion we gave to this problem.

What's more, this method is still not efficient enough, and can't satisfy all the demand in policy analysis despite of the social context. Politics and scholars prefer that, policy analysis should be set in social background and context, without that, data means nothing. Therefore, comprehensive using of different methods may be better.

Acknowledgements

This work was supported by the NSFC Fundamental Efficiently development and management schema of digital information resources based on lifecycle theory (Research Grant NSFC-70833005/G0314-2008) and Jiangsu Province of Philosophy and Social Science Fundamental Strategy, model and trace of information resources development in Jiangsu province according to Industrialization-Informatization Integration Theory (Research Grant JPSSF-2010ZDIXM021-2010).

References

[1] Ahmad Naqiyuddin Bakar. Towards a new mode of governance in Malaysia: policy, regulatory and institutional challenges of digital convergence. Unpublished doctoral thesis, University of Hull. 2008.

[2] Arnold, A.M. A situational analysis of national information policy, with special reference to South Africa. Unpublished doctoral thesis, University of South Africa. 2007, http://etd.unisa.ac.za/ETD-db/theses/available/etd-09172007-142628/ unrestricted/thesis.pdf.

[3] Arnold, A.M. Developing a national information policy considerations for developing countries. The International Information and Library Review, 2004,36: 199-201.

[4] Benoit, K., Laver, M., Estimating party policy positions: comparing expert surveys and hand-coded content analysis. Electoral Studies. 2007,26 (1), 90-107.

[5] Braman, S. Change of state: information, policy, and power. London: The MIT Press. 2006.

[6] Browne, M. The field of information policy: 1.Fundamental concepts. Journal of Information Science, 199723(4): 261-275.

[7] Browne, M. The field of information policy: 2.Redefining the boundaries and methodologies. Journal of Information Science, 1997,23(5): 339-351.

[8] Budge, I., & Pennings, P. (2007). Do they work? Validating computerised word frequency estimates against policy series. Electoral Studies, 26, 121-129.

[9] Burger R.H. Information Policy: A Framework for Evaluation and Policy Research .Ablex, Norwood, NJ. 1993.

[10] Christ, J. M. (1972). Concepts and Subject Headings: Their Relation in Information Retrieval and Library Science. Metuchen, N. J.: Scarecrow. Press.

[11] Creswell, J.W. Research design qualitative, quantitative, and mixed methods approaches. Ed. ke-3.Los Angeles: Sage. 2009.

[12] Debus, Marc. 2006 Parteienwettbewerb und Koalitionsbildung in den deutschen Bundesländern zwischen 1994 und 2006. S. 57-78 in: Uwe Jun, Oskar Niedermayer, Melanie Haas (Hrsg.).

[13] Du Jia. Study on the structure and system of National Information Policy and the Laws and Regulations ,Empirical study on NIPL database.Doctoral thesis in Wuhan University.2005.

[14] E.G. Guba, The alternative paradigm dialog. In: E.G.Guba (ed.), The Paradigm Dialog.Sage, Newbury Park,CA, 1990.

[15] E.G. Guba, The effect of definitions of policy on the nature and outcomes of policy analysis, Educational Leadership, 1984, 42(2): 63-72.

[16] Hatzivassiloglou and McKeown. Predicting the Semantic Orientation of Adjectives. In: Proceedings of ACL-97, 35th Annual Meeting of the Association for Computational Linguistics,1997:174-181.

[17] Hernon, P. and McClure, C.R. Federal information policy in the 1980s: Conflict and issues. Norwood:Ablex Publishing Corporation. 1987.

[18] Hernon, P. and Relyea, H.C. Information policy.In M.A. Drake (ed.), Encyclopedia of Library and Information Science. New York: Marcel Dekker, Inc., 2003:1300-1315.

[19] Hill, M.W. National information policies and strategies: Overview and bibliographic survey. West Sussex: Bowker Saur Limited. 1994.

[20] Jaeger, P.T. Information policy, information access, and democratic participation: The national and international implications of the Bush administration's information politics. Government Information Quarterly, 2007,24(4): 840-859.

[21] Kim, S.M. and Hovy, E. Extracting Opinions, Opinion Holders, and Topics Expressed in Online News Media Text. In Proceedings of ACL/COLING Workshop on Sentiment and Subjectivity in Text, Sydney, Australia,2005.

[22] Klemmensen, R., S. B. Hobolt, and M. E. Hansen (2007). Estimating policy positions using political texts: An evaluation of the wordscores approach. Electoral Studies 26(4), 746-755.

[23] Laver M., Kenneth B. and John G. Placing political parties in policy spaces. Unpublished paper, Trinity College Dublin. 2002.

[24] Laver M.and John G. Estimating policy positions from political texts. American Journal of Political Science.2000,44: 619-634

[25] M. Rein, Social Science and Public Policy.Penguin, Harmondsworth, 1976.

[26] Ma Haiqun, Qi Zengmei.On the Disciplinary Development of Information Policy Research and Its Basic Problems Analysis. Journal of the China society for scientific and technical information. 2007,26(1).

[27] McClure,Charles R. Information Policy: Libraries and Federal Information Policy. The Journal of Academic Librarianship. 1996

[28] McDonald, M. and I. Budge 2005 Elections, Parties, Democracy: Conferring the Media Mandate. Oxford: Oxford University Press.

[29] Moore N. Policy issues in the multimedia age, Journal of Information Science. 1996,22(3):213-18.

[30] Namenwirth J. Some Long and Short Term Trends in one American Political Value: A Computer Analysis of Concern with Wealth in Sixty-two Party Platforms,"Computer Studies in Humanities and Verbal Behavior1 .1968:126-133.

[31] Orna, E. Information policies: Yesterday, today, tomorrow. Journal of Information Science. 2008,34(4): 547-565.

[32] Overman, S.E. and Cahill, A.G. Information policies: A study of values in the policy process. Policy Studies Review, 1990,9(4): 803-818.

[33] Ritter H, Kohonen T. Self-organizing semantic maps. Biol Cybern.1989,61(4):241-54.

[34] Rowlands I. Understanding information policy: concepts, frameworks and research tools, Journal of Information Science:1996, 22(1) :13-25.

[35] S. Braman, Horizons of the state: information policy and power, Journal of Communication 1995,45(4) :4-24.

[36] Trauth E.M. An integrative approach to information policy research, Telecommunications Policy 1986,10(1): 41-50.

[37] Tu Duan-wu. The Control Structure of Higher Education Policy

978-1-935068-76-1 © 2011 SciRes.

and Its Evolution——An Empirical Study Based on Policy Text.Fudan education forum,2010,8(2).
[38] Yan Hui. Trends of Policies on Information Resources Exploiting and Usage for Public Benefits in China:An Research Based on Content Analysis .Library and information service.2009,53(14).
[39] Zawiyah M. Yusof, Mokmin Basri and Nor Azan M. Zin. Classification of issues underlying the development of information policy. Information Development .2010,26: 204-213.

Intelligent City-oriented to Innovation of Information Services

Peng CHENG, Huichao YAN, Jianxin SHENG, Wei LIU, Tingting YONG

Strategic Management Research Institute, Hubei Academy of Scientific and Technical Information, Wuhan, China

Abstract: Intelligent information services is currently an hot subject of research and the direction of development at information service field, and is the focus of the construction of intelligent city. From the perspective of the intelligent city development, this paper makes a new interpretation for information services, and proposes new information services system and the strategies for promoting.

Keywords: Intelligent city; information services; Internet of Things

Cities are an important part of human civilization, and city evolution process often closely links with the breakthroughs of technologies. Especially, in the context of the information society, every breakthrough of information technology, such as the transistor, the computer, the Internet and the wireless network, has driven the process of urban information, as well as changes of city lifestyle. Nowadays, humankind is in a new era, which a period of the emergence of new revolution of information society. As the representative of the new generation of intelligent information technology, the technologies of Internet of Things and Cloud Computing are in the ascendant, and these technologies correspond to the development model of "intelligent city", is becoming a strategic approach to future development of the cities world widely and leading a new direction for urban development.

The researches on intelligent city-oriented information services, originated from "Smarter Planet" concept that proposed by IBM in 2008, and it is one of most important aspects within the concept. The core idea of it is that intelligent city is based on a combination of Internet and Internet of Things, to achieve the purpose of instrumented, interconnected and intelligent. Furthermore, such as South Korea, Taiwan and Singapore have already been started relevant practices and explorations regarding how to build and develop intelligent city. At present, many cities in China are promoting the intelligent city construction actively; however, there exist the lack of systematic study regarding intelligent city-oriented information services. This paper attempts to explore this issue and makes further consideration about current intelligent city construction and relevant information services. We believe that, as one of basic function elements of city-building, information services will have new changes under the new technology background.

1 Intelligent City-oriented to Transformation of Information Services

1.1 The Changes from Digital Information Services to Intelligent Information Services

At present, as the basic direction of development of city information, digital city model has been recognized broadly, and model of urban information services was been formed to facing the digital services, that is, the groundwork contains computer technology, multimedia technology and spatial information technology, and broadband network as a link, to make multi-resolution, multi-scale, multi-dimensional space and various 3-D description for the city, by using technologies of remote sensing, GPS, GIS, telemetry and simulation-virtual, or namely the entire information concerning city's past, present and future would be displayed through digital and virtual information technology.

We believe that digital city is a form of initial development stage of the intelligent city; compared to digital city, intelligent city should require higher level of demands, fruitful contents and more ambitious goals. "Intelligent City" is not equal to the city informatization or a simple upgrade of "Digital City", but an embodiment of urban development in different historical stages. Intelligent city is not only the informatization of urban infrastructure and simple smart, but also to concern about city's sustainable development and actual needs of people's lives, and combined the advanced information technology with the scientific concept of urban development. On the basis of the digital city, intelligent city is emphasis on integration, coordination and interaction.

There exist some similarities and differences of information services between digital city and intelligent city. The major differences involve that Information services of digital city is limited informatization, local sharing, initial automation, and the essential part is digital virtual; on the other hand, information services for intelligent city is digital construction integrated, full sharing, high level of intelligence, and urban entity consider as the core part of it.

978-1-935068-76-1

Table 1. The Comparison of Physical Support System of Information Services for Both Digital and Intelligent City

	Digital City	Intelligent City	Key difference
Information Terminal	Computer	Intelligent Facility	Sensing Capability
Web Infrastructure	Internet	Internet of Things	Access Capability
Software Technology	Middleware and Proxy	Digital Nervous System	Collaboration Capability
Data Organization	Digitization	Integration	Reconstruction Capability
Information Process	Computing Center	Cloud Computing	Analysis Capability
Decision Support	Event-oriented	Social-oriented	Feedback Capability

1.2 Intelligent Information Services for Intelligent City

1) The content of information services

Under the background of intelligent city, during the construction process of city and the scope of city's environment, public services, local industry and all citizens, with the help of Internet, Internet of things and intelligent equipment those kinds of high-tech informatization methods, intelligent information services is been used to collect and dynamic monitoring comprehensive information regarding city's politics, economic, life and culture. By sufficient statistics, connection and sharing, these information were been perceived, analyzed, integrated and responded intelligently, after that, there are better decision support and dynamic control capabilities could be provided to the operation and development of the city, so urban management would become more intelligent in order to liberate as much as possible to maximize the use and promotion of people's own wisdom, to provide a more healthy, safe, harmony and happy living environment for the city residents.

2) The goal of information services

The goal of intelligent city information service is based on the conditions of city development. Through the establishment of an intelligent foundation support system that is comprehensive perceptive, interactive, sharing, to improve the application operating system, which is highly automated and intelligent, highly dynamic and predictable, highly enforceable and implementable, highly public nature and personalized, and highly safety and reliability. The purpose is to form an intelligent and harmonious social environment, to enhance the city's comprehensive development functions, achieve possible in step of happiness and urban development to all the participants under the intelligent environment.

3) The method of information services

During Information services in the traditional manner, the information always gain and use "passive", in most cases, only when people need to use some specific information, they would rely on search tools to inquire relevant resources from mass of database, in consequence, inefficient process have gained nothing frequently. However, intelligent information services more emphasis on "active", the essential part is will have high-quality information to complete fully sharing around whole services networks, and according to each person or origination's specific requirements, personalized information services are been delivered regularly to maintain interactive and dynamic updates intellectualized. This kind of information services simplifies the difficulty of information access and makes it more targeted in general.

2 Intelligent City-based Information Services Architecture

2.1 Intelligent City-oriented to Information Services System Infrastructures

Construction of the intelligent city needs intelligent information services and the building of basic platform is essential to intelligent information services. Generally, intelligent information services infrastructures include some platforms construction as follow:

1) Information Perception Platform

Information perception platform rely mainly on sensors and technology of Internet of Things, it is to consider related information of people, items and things as a collection of objects, and to make a reliable record based on their unique properties, at same time, to complete digital format according to unified standard code, achieves the targets of transmission, storage, sharing and application.

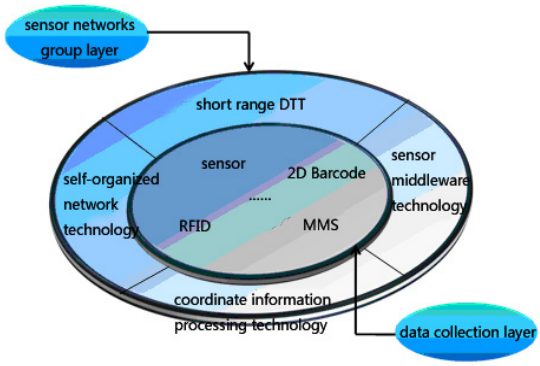


Figure 1. Diagram of Information Perception Platform

2) Information and Telecommunication Platform

Information and telecommunication platform is to transmit the data information, which from various applications and sub-networks, reliable and secure, by utilizing the technologies of tri-networks integration and next-generation Internet, to ensure the accuracy, real-time, security, integrity and effectiveness of the information transfer and to make sure the information flow of whole

intelligent city operate very well.

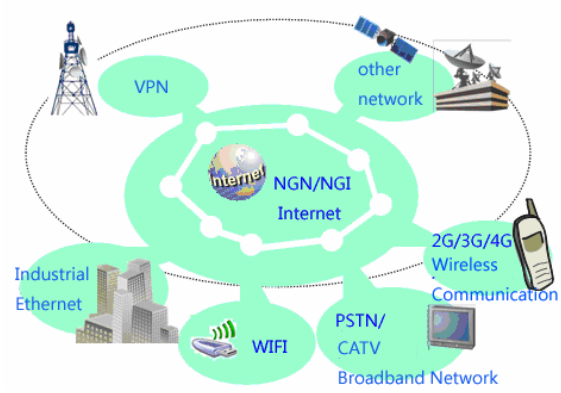


Figure 2. Diagram of Information and Telecommunication Platform

3) Information Processing Platform

By using Cloud Computing technology, Information processing platform analyze, process and store the data from different sources, according to relevant characteristics, while follow demands of different applications levels, take a specific algorithm to carry on large and secure calculation, and make intelligent discernment, selection and distribution, in order to complete the requirements of intelligent city dynamic monitoring and the dissemination of early warning information.

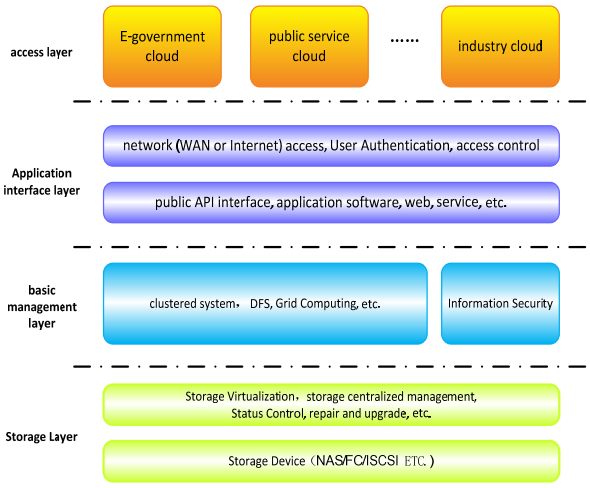


Figure 3. Diagram of Information Processing Platform

4) Information Sharing Platform

Information sharing platform focus on to set up the strategic data resources of intelligent city, through various information-based methods to share information to all residents. It is one of the core contents for supporting operation and development of intelligent city; it contains the city's politics, cultural, economy, transportation, technology, military, security and people's livelihood, and many other important information, and also play the role of information decision support for the whole city.

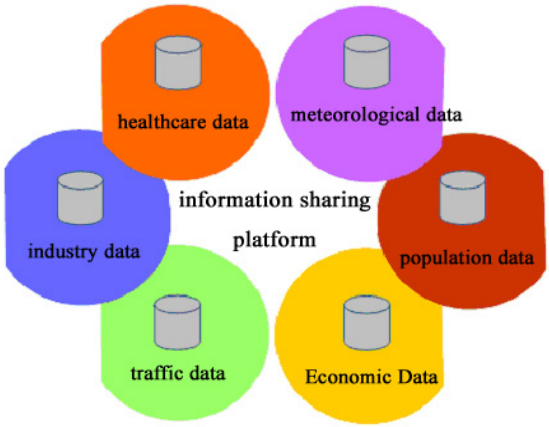


Figure 4. Diagram of Information Sharing Platform

2.2 Intelligent City-oriented to Application Architecture of Information Services

Information services of intelligent city is not to negate nor reconstruct the information services system on the original city, but closely related to the original system to strengthen, transform and upgrade it. In the strategic conceptual level of construct intelligent city, based on serving city itself, the range of application fields for intelligent information services could widely from public administration of government, industry dynamics to a small communities and individuals, to cover all aspects of city's management, development and people's life. From the perspective of specific applications, intelligent information services could be divided into three main areas, namely "public administration-oriented information services", "industrial development-oriented information services", "people's livelihood-oriented information services."

1) Public administration-oriented information services:

Based upon the content of public administration, it provides government decision support, interactive online office, public environmental monitoring, public safety monitoring and other varied information services of public affairs to help city managers to achieve more efficient, accurate and reliable public management.

2) Industrial development-oriented information services:

In this field, information services cover industry dynamic investigation, tracking industrial policy, market analysis and different kinds of consulting services for enterprise development. The main role of information services is reflected in two aspects: first, in terms of intelligent information analysis techniques, to promote

978-1-935068-76-1 © 2011 SciRes.

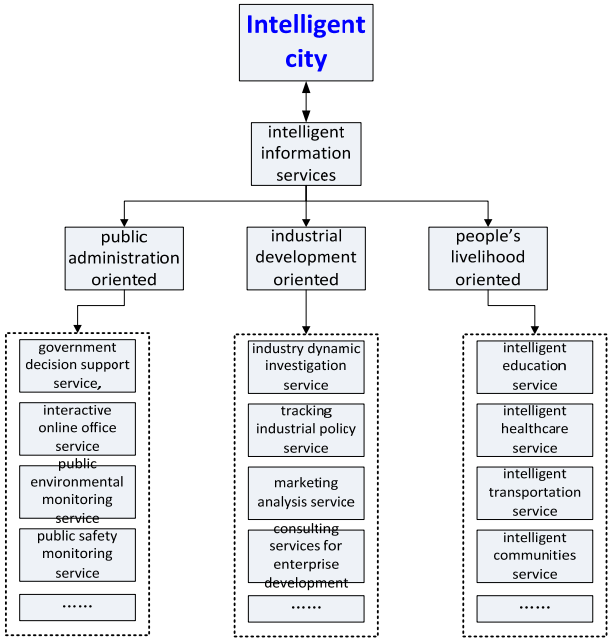


Figure 5. Application Fields of Intelligent Information Services

the cultivation of strategic new industries, as well as the transformation and upgrade of traditional industries; second, based on available data and macro consideration, to predict and foster some new industries that are potential and strategic in a global scale.

3) People's livelihood-oriented information services:

The content of people's livelihood is extensive, it closely related to people's life regarding food, medicine, housing, transportation, education, entertainment and other aspects. The typical representatives of such information services are intelligent healthcare, intelligent transportation, intelligent education and intelligent communities. And the high sense of happiness of urban residents in the intelligent environment was embodied by this kind of information services.

3 The Promotion Strategies of Information Services for Intelligent City

No matter the construction of intelligent city or building information services, both of them need to follow the objective laws of urban development and conform to the current international environment and background; for that reason, focus on scientific, rational, effective and feasibly natures, several fundamental principles for promoting strategies of information services are defined as below:

1) Outstanding Subjectivity

Intelligent city-oriented to construction of information services is a new engine that supports city's accelerated development. For the building objectives, the strategies should realize the vision of urban development and services, as well as the main demands of the city livelihood; to fully reflect the interaction with subject of urban construction and to encourage social subjects to participate on it, such as government, various non-governmental organizations and individuals.

2) Outstanding Integration

Intelligent city-oriented to construction of information services is a new starting point that drives the transformation of urban development patterns and to improve the development capabilities. Its integration should cover city's systems of society, physics and economy to make overall planning, and let the large city system more integrated, coordinated and propulsive.

3) Outstanding Openness

Intelligent city-oriented to construction of information services is the new process that emphasis on cooperation with cities and innovation development. Its openness should be reflected at all levels participation and effective opening for inside and outside of the city, province and country.

4) Outstanding Fundamental

Intelligent city-oriented to construction of information services is a fundamental "site project" that based on city's real development capacities. The fundamental must be embodied in a basis of realistic operation, to reflect the incremental construction and the reserve activation of intelligent city's entity, to reflect further development and maintenance of supporting infrastructures.

5) Outstanding Long-term Vision

Intelligent city-oriented to construction of information services is a long-term "top project" that leading the future urban development. Its long-term vision should reflects the respect of building process, correctly handle relationships between both input and output, fundamental support and application breakthrough; to reflect how grasp the nature of construction, the design planning, system architecture, standards development, technology selection, project construction and other key areas; to reflect the security of construction control; to maintain the building embodies sustainability.

4 Conclusion

Intelligent information services represent the "wisdom" of intelligent city directly. The paper analyzes the inevitability of transformation to intelligent information services from the revolution of information society and the rise of intelligent city, and makes description regarding system architecture of intelligent information services and promoting strategies. Nonetheless, we should also see that, whether the development of intelligent city or building information services system, aside from basic hard environment, suitable "soft environment" is necessary to make better protection, policies and regulations promote and restrict the construction of intelligent information services in both sides; meanwhile, to strengthen talent pool strategy and

the services innovation consciousness, so that, it can better realize the intelligence of city.

References

[1] K, Richard, B, Dominica, R, Joe, "Urban regeneration in the intelligent city", Proceedings of the 9th Internatinal Conference on Computers in Urban Planning and Urban Management, CASA, UCL, London, 2005.

[2] N. Komninos, "The Architecture of intelligent cities", Conference Proceedings Intelligent Environments 06, Institution of Engineering and Technology, 2006, pp.53-61.

[3] ZHANG Anzhen, On intelligent information service in the Network Environment[J], Information Studies: Theory & Application, 2004, 27(6).

[4] YAO Wanhua, Planning with intelligence for "intelligent city"[J], China Information Times, 2010, 137(1), p32.

[5] WANG Hui, WU Yue, ZHANG Jianqiang, et al. Intelligent Cities. Beijing: Tsinghua University Press, 2010.

978-1-935068-76-1

Accelerating Global Science Access

WorldWideScience.org's Combination of Search, Translations, and Multimedia Technologies

Walter L. Warnick[1], Brian A. Hitson[2], Lorrie Apple Johnson[2]

[1]*U.S. Department of Energy, Office of Scientific and Technical Information, Germantown, Maryland, USA*

[2]*U.S. Department of Energy, Office of Scientific and Technical Information, Oak Ridge, Tennessee, USA*

Email: Walter.Warnick@science.doe.gov, HitsonB@osti.gov, JohnsonL@osti.gov

Abstract: WorldWideScience.org (WWS.org) is a global science gateway governed by the WorldWideScience Alliance, with the U.S. Department of Energy Office of Scientific and Technical Information (OSTI) acting as the Operating Agent. WWS.org provides a simultaneous live search of more than 70 scientific and technical databases from government and government-sanctioned organizations representing 71 countries. From its inception in 2007, WWS.org has been at the forefront of transformational technologies, including, since June 2010, the first use of federated searching in conjunction with real-time translations capability for nine of the world's most widely spoken languages. This paper briefly reviews the history of the development of WWS.org and discusses several new features and technologies to be launched in June 2011, including the addition of Arabic to the WWS.org multilingual translations capability, expanded multimedia search functionality, and introduction of innovative mobile access. These ground-breaking technological advances will further increase WWS.org's ability to accelerate scientific discovery throughout the world.

Keywords: WorldWideScience.org; WorldWideScience Alliance; Federated Searching; Multilingual; Machine Translation; Multimedia Scientific Content; Mobile Phone Applications; International Collaboration

1 Introduction

As methods and practices for sharing and communicating scientific information rapidly evolve, WorldWideScience.org (WWS.org) is, likewise, rapidly adopting new technologies and strategies to accelerate scientific discovery. The history of WorldWideScience.org since its prototype development in 2007 is well documented. While this paper will briefly summarize that history, the primary purpose is to describe the WWS.org technical growth and diversification strategies implemented since 2009, including the specific new features and breakthroughs to be launched in June 2011. These include (a) transition of the beta version of Multilingual WWS.org to the default search for WWS.org; (b) addition of Arabic to WWS.org's multilingual translations capability; (c) integration of multimedia scientific content into WWS.org searches; and (d) development of a mobile web version of WWS.org (and why this is particularly notable). In addition to discussion of these particular technological advances, additional aspects of WWS.org growth and uniqueness will also be explored.

2 Brief History

2.1 The Model – Science.gov

WWS.org was conceived and developed using the model of Science.gov and its underlying federated search technology. Launched in December 2002, Science.gov provides a single search point to "over 45 databases and 200 million pages of science information" within 14 U.S. federal science agencies [1].

National governments frequently face the challenge of improving the visibility and accessibility of information and other database and website content. In many cases, citizens are indifferent as to which governmental agency conducted research in a particular area; they simply want to access *any* government information on a specific topic. It is quite common that multiple agencies address different aspects of a particular discipline and research topic. Before Science.gov, in the U.S. case, a person would need to be aware of each relevant agency's databases and websites in order to conduct a comprehensive search across the entire government. While having this awareness was possible (albeit unlikely), the follow-on temporal barriers (i.e., the time required to search individual sources sequentially and the time required to sort and order results by relevance) made such searching a practical impossibility for any busy researcher.

Science.gov solved all of these problems by providing users a single point to simultaneously search all of the U.S. government's scientific resources and retrieve relevance-ranked results. Science.gov has been cited by numerous public and private sector information advocates for increasing transparency to government-sponsored scientific information [2].

2.2 WorldWideScience.org – Launch, Governance, and Growth

As a principal figure in the development of Science.gov and subsequently its Operating Agent, Dr. Walter L. Warnick, Director of the U.S. Department of Energy's Office of Scientific and Technical (OSTI), in June 2006,

first proposed extending the Science.gov model on an international scale and invited other nations to partner in creating a "Science.world." The rationale behind the concept was fairly obvious and simple: if finding disparate and decentralized scientific databases in a single country is a barrier to information visibility and use, then finding information from geographically-dispersed databases around the world was an even greater barrier to scientific communication. The first "partner" to accept Dr. Warnick's invitation was the British Library, and in January 2007 this partnership was formalized, with other nations invited to join the effort. By June 2007, the "Science.world" concept was realized with the launch of the WorldWideScience.org prototype at the International Council for Scientific and Technical Information (ICSTI) annual conference in Nancy, France. The prototype initially searched 12 national scientific databases in 10 countries.

To reflect its multinational content, the initial partners in WWS.org established a multilateral governance structure called the WorldWideScience Alliance and elected officers from Africa, Asia, Europe, and North America, further evidence of its geographic diversity and representation. Current Executive Board Members include Richard Boulderstone, British Library (Chair), Pam Bjornson, Canada Institute for Scientific and Technical Information (Deputy Chair), Tae-sul Seo, Korea Institute of Science and Technology Information (Treasurer), Roberta Shaffer, International Council for Scientific and Technical Information (Ex-Officio Member), Walter Warnick, U.S. Department of Energy Office of Scientific and Technical Information, WorldWideScience.org Operating Agent (Ex-Officio Member), and Martie van Deventer, Council for Scientific and Industrial Research of South Africa (At-Large Delegate).

To avoid overlap and duplication of commercially-available scientific information resources, the WorldWideScience Alliance established criteria for the kinds of databases it would consider for WWS.org searches. The over-riding principle was that databases must be produced, sponsored, or endorsed by a national scientific body (for example, Science.gov (U.S.), the Institute of Scientific and Technical Information of China's databases, Japan Science and Technical Agency's J-STAGE and J-EAST databases). Using these primary criteria, WWS.org quickly grew beyond the initial 12 databases via one of two typical pathways: (a) as the WWS.org Operating Agent OSTI actively identified and contacted national databases seeking permission to include them in WWS.org searches or (b) vice versa, and increasingly commonly, national databases contacted WWS.org seeking to be included in WWS.org searches. As a result, growth in the content and geographic representation in WWS.org has been steady and impressive, going from 10 countries and 12 databases in 2007 to 71 countries and 77 databases by March 2011.

[1] The Deep Web (also called Deepnet, the invisible Web, DarkNet, Undernet or the hidden Web) refers to World Wide Web content that is not part of the Surface Web, which is indexed by standard search engines. (Wikipedia, 2011)

Figure 1. WorldWideScience.org Home page

3 Measures of Uniqueness and Value

3.1 Uniqueness of Records

A significant quantity of scientific information can be found by commercial and certain publisher-operated search engines. Historically, however, commercial search engines have not been able to perform real-time searching of deep web[1] databases, relying on automated crawls of static web pages to populate their enormous indexes. There are some exceptions to this general model, including, for example, arrangements such as Google Scholar's coverage of scholarly literature and the adoption of site-map protocols by database owners to expose their content to commercial search engines' crawlers.

A key feature and founding principle of WWS.org is that it searches national scientific databases and nationally-sponsored or –endorsed sources. To avoid overlap and redundancy with commercial and publisher sources, WWS.org does not directly search any commercial scientific database or website. To gauge whether this principle results in unique search results, the WWS.org Operating Agent performed two comparative analyses (September 2009 and February 2011) of search results from Google, Google Scholar, and WWS.org.

In the September 2009 analysis, search results were captured from each of the three search engines, using 33 queries across a wide range of scientific disciplines. A key area of interest was the amount of overlap in the results sets. Overlap was defined as exact title matches of the records. Across the 33 queries and accounting for all results, WWS.org results were uniquely different from Google and Google Scholar 96.5 percent of the time. Assuming that users typically focus on the first 50 results, there was a 9 percent overlap between the first 50

978-1-935068-76-1 © 2011 SciRes.

WWS.org hits and the Google and Google Scholar items. In other words, WWS.org results were unique 91 percent of the time within the first 50 hits.

The same analysis was repeated in February 2011. Using the same methodology with the same 33 queries, WWS.org results were 92.7 percent unique. Comparing the first 50 results from each search engine, the overlap had declined to 2.4 percent, or, stated differently, the uniqueness had increased to 97.6 percent.

The primary conclusion from the second analysis was that WWS.org uniqueness remains relatively high and especially high among the first 50 results. In other words, WWS.org is filling a niche by searching national scientific databases.

3.2 Usage Growth

A major challenge for any new search product, such as WWS.org, is building brand identity/recognition and user awareness. This is particularly true for a product that essentially has no advertising or marketing expenditures. One method by which WWS.org has addressed this challenge is by leveraging the referral power of major commercial search engines. Since the introduction of search engines' sitemap protocols, this technique has become easier by simply exposing deep web content (e.g., electronic full-text documents) to the sitemaps. The content then gets crawled by the search engines and can subsequently be found in, for example, a Google search, and linked to from the set of search results. This process works quite well for centralized databases, but it is not particularly effective for federated search engines such as WWS.org because WWS.org performs a search of decentralized databases. To benefit from Google's and others' referral power, OSTI developed a program that used results from live searches of WWS.org to create a list of over one million unique scientific query terms. Over five million search results were then searched and categorized by the scientific query terms to create a set of cached, topical search results pages (what OSTI termed "topic pages"). These topic pages were exposed to search engines through a sitemap.

Following the indexing of these topic pages, WWS.org web traffic roughly doubled during the first month after deployment. Traffic has continued to growth at a brisk rate. Current numbers show a ten-fold increase compared to the amount of traffic experienced prior to implementation of the topic pages. Once a user finds a WWS.org-generated topic page, he or she is then on the "premises" of WWS.org and has the option to repeat the search in real-time on WWS.org. Along with growing user awareness and repeat visits, the topic page-generated usage has also contributed to significant direct WWS.org queries. The average number of queries per month has tripled during 2011, as compared to the average number of queries per month during 2010.

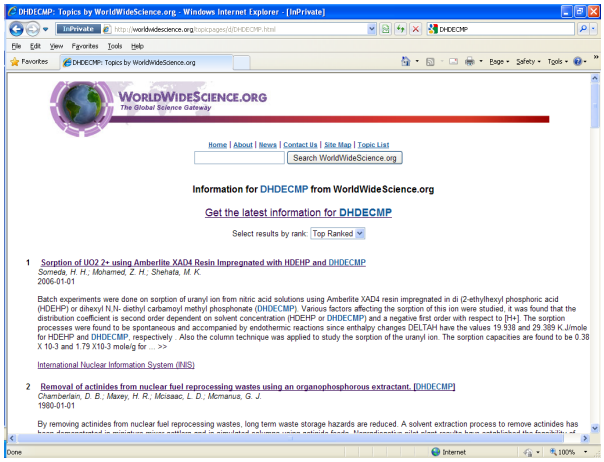


Figure 2. WWS.org Topic Page

4 Expanding Reach and Content Types

4.1 Multilingual Translations

When WWS.org was launched in 2007, it initially was limited to search and retrieval of English-language scientific databases. While English is traditionally the *lingua franca* for science, this limitation had the effect of (a) under-serving non-English-speaking populations and (b) excluding access to the expanding volume of non-English papers in countries such as Brazil, China, France, Germany, Japan, Russia, and many others.

For non-English-speaking populations, this meant that WWS.org had played little role in ameliorating the "digital divide," an issue addressed by developmental initiatives and organizations. Strategies for closing the digital divide have been stated and implemented perhaps most eloquently and effectively by the World Summit on the Information Society (WSIS). The justification for closing the divide was partially captured in this WSIS statement, "The ability for all to access and contribute information, ideas and knowledge is essential in an inclusive Information Society."[3]. Of course, more tangibly, such access provides the intellectual and physical foundations for health systems, scientific facilities, infrastructure, and technology-driven commerce. Considering the ubiquitous nature of globalization (including science), providing access to English-language science to non-English-speaking populations has the potential for accelerating the rate of scientific progress.

Conversely, there is an increasing volume of non-English scientific content, both conventional and non-conventional literature, being produced for national journals, institutional repositories, and regional data-bases. Addressing the issue of language barriers in science, Meneghini and Packer stated in 2007, "Is there Science beyond English? Initiatives to increase the quality and visibility of non-English publication might help to break down language barriers in scientific communication." [4].

Barany, in 2005, also stated, "As globalization increases, communication between linguistic communities could become a serious stumbling block." [5].

Indeed, of the world's "top 400" institutional repositories, 250, or 63 percent, have some or all non-English content [6]. Japan, France, Germany, Brazil, China, Russia, and other countries all produce vast amounts of science in their respective national languages. Practically all Russian scientific output is in Russian. In the case of China, prior to June 2010, WWS.org searched a small subset of English records from the Institute of Scientific and Technical Information of China (ISTIC). ISTIC holds a much larger collection of Chinese language content. According to Mr. Wu Yishan, Chief Engineer at ISTIC,

"In China today, we estimate that people who master sufficient English account for at most 5 percent of the urban labor force. On the other hand, more and more people in the world expect to understand China deeply, and the number of Chinese learners outside of China will reach 100 million by 2010, according to an expert estimate. For international community who are eager to keep abreast of the development of China's science and technology, understanding some Chinese is of particular importance. In 2008, while Chinese scholars published 110,000 papers in international journals recorded by Science Citation Index, they also published 470,000 papers in domestic Chinese journals. Without accessing these 470,000 papers, it is impossible to obtain a realistic feeling about the thrust of scientific and technological advancement in China. Therefore, the need for mutual translation between English and Chinese and for cross-language retrieval is increasingly urgent."[7].

The continuing growth of such non-English content creates its own digital divide in that it is not particularly accessible to English-speaking populations and, indeed, any population other than those speaking the language of a particular paper.

These challenges clamored for a multilingual translations solution, but automated/machine translations have historically lacked the precision necessary for scientific translation, and the application has generally been done on a one-to-one basis; that is, translating from a single language to another single language.

In a federated search environment, such as WWS.org, there was the potential to search databases with content in multiple languages simultaneously, but the WWS.org search engine needed to be able to translate a user's query into the various languages of constituent databases on the front end and then translate the search results into the user's language on the back end. Working with the translations team from Microsoft Research, the WWS.org search engine provider, Deep Web Technologies, successfully integrated such translations capability into the front and back end of the user experience. In June 2010, a beta version of Multilingual WWS.org was launched at the ICSTI annual conference in Helsinki.

The beta version provided translation capabilities for nine languages (Chinese, English, French, German, Japanese, Korean, Portuguese, Spanish, and Russian). From the search screen, users simply select their preferred language and enter the search terms, and the software translates the query as appropriate for each database. Users then receive the relevance-ranked results list, with the option to translate the results into their language as well. Upon viewing a specific record, users again have the option to translate the bibliographic record (title, abstract, etc.) into the language of their choice.

Since the beta multilingual version was released in June 2010, the system has proven quite stable, and the WWS Alliance agreed in February 2011 to incorporate multilingual translations into the main WWS.org product by June 2011. Users coming to the main WWS.org search page will be immediately offered the option of searching in one of ten languages. The original nine languages used for the beta version will be retained, and Arabic will be offered. Adding Arabic presented some

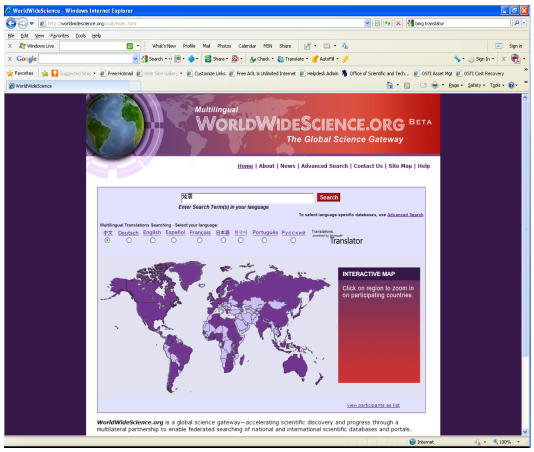


Figure 3. Multilingual WorldWideScience.org[Beta]

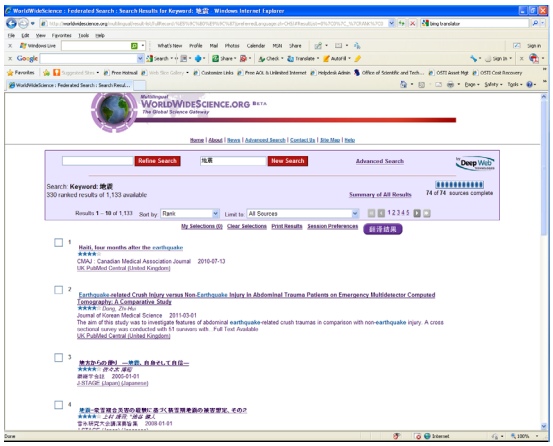


Figure 4. User's query has been translated and results returned. User clicks on Translate Results button to translate

978-1-935068-76-1 © 2011 SciRes.

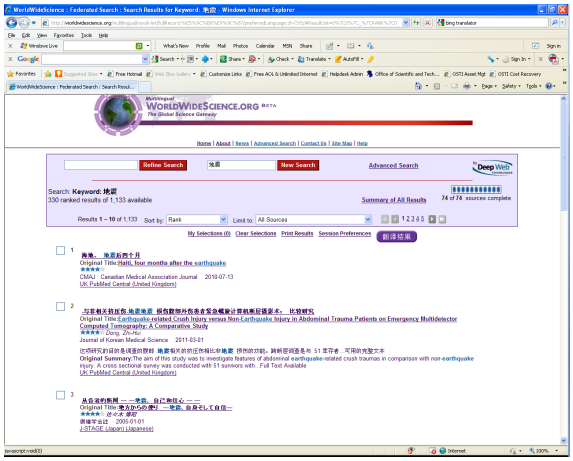


Figure 5. Results have been translated into user's native language

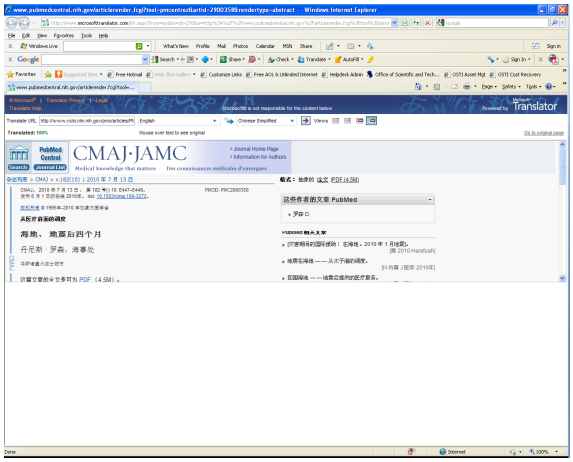


Figure 6. Users also have the option to translate the record into their preferred language

additional technical challenges, as the browser must accommodate right-to-left text. The WWS Alliance strongly endorsed adding Arabic, the fifth most widely-spoken language in the world, and to provide increased access to scientific and technical resources written only in Arabic.

Once multilingual translations are integrated into the production version of WWS.org, additional topic pages will be generated using terms and phrases from the ten languages currently available. This technique will continue to increase both accesses and queries, with the additional benefit of expanding potential users' awareness of WWS.org.

4.2 Access to Multimedia-based Science and Technology

New forms of scientific information, such as numeric data, multimedia, and social media, are emerging rapidly and becoming increasingly prevalent as a primary means of scientific communication. Many scientific and technical conferences and symposia, for instance, are now recorded, and presentations in video format are available for public access. Multimedia information introduces some special challenges, such as the lack of written transcripts, minimal metadata (no abstracts or keywords), and complex scientific/technical/medical terminology. Additionally, many of these videos are long, up to an hour or more in length. For a scientist interested in only one particular part of a video or experiment, locating it could represent a substantial time burden [8].

In February 2011, the WWS.org Operating Agent, OSTI, released a new product called ScienceCinema, which contains approximately 1,000 hours of research-based videos from the U.S. Department of Energy's National Laboratories. In a collaborative partnership with Microsoft Research, this project utilizes Microsoft Research Audio Video Indexing System (MAVIS). MAVIS is a set of software components that use speech recognition technology to enable searching of digitized spoken content [9].

Figure 7. ScienceCinema Home Page

Figure 8. ScienceCinema results list showing snippets where search term was spoken during the videos

The end result is that users can search for a precise term within the video and be directed to the exact point in the video where the term was spoken. Once a search is conducted, the results list contains "snippets" and a timeline with links showing the occurrence of the search term. The user simply clicks on the link, and the video will begin to play at the point where the search term was spoken.

In addition to OSTI's ScienceCinema product, several other WWS.org sources have multimedia content, including, most prominently, CERN and several U.S government agencies. Searches of multimedia content will be integrated into WWS.org in June 2011. Users will continue to have the option of viewing traditional text-based results, along with a separate results list of relevant multimedia items. In the case of ScienceCinema, the actual point in the video where the search terms occur will be identified in the WWS.org results list, and the user will be able to view the video by clicking on the "snippet" links.

Although Microsoft's MAVIS technology has been used in other applications besides ScienceCinema, the integration of ScienceCinema into WWS.org will represent the first use of this audio indexing technology in a federated search environment. It is anticipated that the MAVIS technology will be applied to other WWS.org multimedia sources in the near term. Beyond this, future goals include exploring multilingual translations capabilities for multimedia, in conjunction with the capabilities WWS.org now offers for text-based information.

4.3 Mobile WWS.org – Another First for Federated Search

Growth in mobile phone usage in the last decade has been exponential. A report from the United Nations released in March 2009 indicated that the number of mobile phone subscriptions throughout the world quadrupled, from 1 billion in 2002 to 4.1 billion by December 2008 [10]. The majority of this growth has emanated from developing countries. In fact, mobile phone usage is growing faster among African countries than anywhere else in the world [11]. Studies indicate that mobile phones allow developing countries to leapfrog old technologies. In the U.S., about 91 percent of the population subscribe to mobile phone services [12].

Beyond the virtual ubiquity of mobile phones, the devices have become increasingly "smart" and capable of rapid web transactions involving significant amounts of data. With the convergence of mobile device availability and capability, the WWS Alliance supported the creation of a mobile version of WWS.org. Although the current production version of WWS.org will work on a mobile phone, the graphics and other content are often difficult to read on smaller devices. A version of WWS.org developed specifically for mobile usage will be released in June 2011.

Mobile WWS.org features a streamlined search screen and results list. Users may view records directly at the original source database by clicking on the links or have the option to email the results list for later viewing on a PC or MAC. Mobile WWS.org works with major popular devices and operating systems, such as the iPhone, Android, and Blackberry.

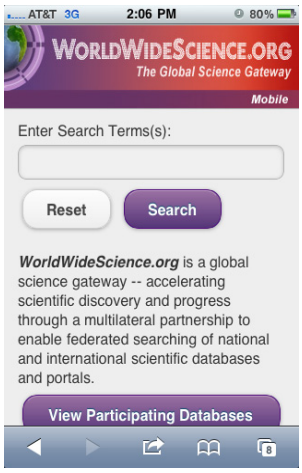


Figure 9. Mobile WWS.org on an iPhone

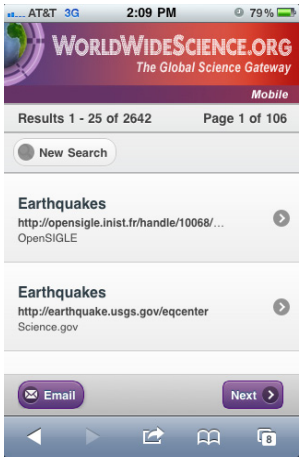


Figure 10. Mobile WWS.org results on an iPhone

While a mobile version of a web product is not new, Mobile WWS.org offers unique capabilities, as it enables federated searching of nearly 80 databases, many of which are not individually optimized for mobile web searching. Access to scientific and technical information provided by national and international entities, such as those represented in WWS.org, is important, particularly within developing countries. Because mobile phone usage within these areas is expanding rapidly, mobile applications may be the most successful means of reaching these users. After release, usage of Mobile WWS.org will be monitored and additional features will be added according to user feedback and demand.

978-1-935068-76-1 © 2011 SciRes.

5 Summary – A Unique Combination of Technologies

At its inception in 2007, WorldWideScience.org represented the unique extension of federated searching technology on an international scale. Its growth to searching nearly 80 databases across all inhabited continents represents a novel scaling of this technology. In 2010, federated search was combined with multilingual translations capability, another first in the blending of technologies to accelerate access to scientific information. And, in 2011, WWS.org builds on this innovation with the addition of Arabic to multilingual translations; the integration of speech-indexed multimedia search and retrieval; and a mobilized version of international federated search. This pattern of continuous growth in content and technological capability has resulted in significantly increased usage and unique search results. As a result, WWS.org is filling a niche in the scientific information landscape and playing a leading role in accelerating scientific progress.

References

[1] Science.gov website, "About Science.gov," http://www.science.gov/about.html, March 2011.

[2] Science.gov website, "Communications Archive Page," http://www.science.gov/communications/archive.html#photo, March 2011.

[3] Report of the Geneva Phase of the_World Summit on the Information Society," GENEVA-PALEXPO, 10-12 December 2003.

[4] R. Meneghini and A. L. Packer, "Is there science beyond English? Initiatives to increase the quality and visibility of non-English publications might help to break down language barriers in scientific communication," in EMBO reports (2007), vol. 8, pp. 112-116.

[5] M. J. Barany, "Science's Language Problem," in Bloomberg Business Week, 16 March 2005.

[6] "Ranking of World Repositories," Cybermetrics Lab http://repositories.webometrics.info/toprep.asp.

[7] Wu Yishan, Remarks at "From Information to Innovation," International Council for Scientific and Technical Information (ICSTI) Annual Conference, Helsinki, Finland, June 2010.

[8] B. Chitsaz and L. A. Johnson, "ScienceCinema: Multimedia search and retrieval in the sciences," Multimedia and Visualisation Innovations for Science, ICSTI Winter Workshop, Redmond, WA, USA, February 2011.

[9] Microsoft Research Audio Video Indexing System (MAVIS) website, http://research.microsoft.com/en-us/projects/mavis/.

[10] J. Fermoso, "U.N. world report picks up massive growth in mobile phone ownership," Wired.com, March 2009.

[11] "Mobile phone growth fastest in Africa," BBC News website, March 2005 http://news.bbc.co.uk/2/hi/business/4331863.stm.

[12] C. Foresman, "Wireless survey: 91% of Americans use cell phones," Ars Technica website, March 2010, http://arstechnica.com/telecom/news/2010/03/wireless-survey-91-of-americans-have-cell-phones.ars.

A Classification and Coding Solution for Information Resources of Science and Technology Talents

Xiaoli WU[1], Yunhong WANG[2]

Centre for Resource Sharing Promotion, Institute of Science and Technology Information of China, Beijing, China

Email: wuxiaoli611117@126.com, wangyh@istic.ac.cn

Abstract: This paper proposed a classification and codification solution to science and technology (S&T) talents information in order to optimize the existing classification and codification system. The solution includes the property characteristics of resources including talents' basic information, knowledge and competence, scientific activity, production as well as their credibility information. The solution could organize, relate and reveal S&T talents information, which would be very important to the database building and data sharing.

Keywords: classification; coding; S&T talents; Information resource of talents

In 2006 Chinese president Hu Jintao officially put forward the target of constructing China into an innovative country by 2020 on National science and technology conference. Innovation in science and technology should be the key engine driving China's development. In the past years the number of S&T talents has been increasing rapidly. Until 2008 the population of S&T talents has reached 49,600, 000, R&D talents to 1,965,000, scientists and engineers to 1,592,000[1]. The governments, universities, institutes, S&T companies and intermediary agencies have already started to digitalize and standardize S&T talents information.

Classification and codification of S&T talents information acts as the basis of data management, and interaction for database building. In China how to classify and code S&T talents information is still under exploration. Based on our survey to databases of S&T talents information built by government departments that responsible for scientific research management, we found that 44% of database classification system was based on "Classification and code disciplines", 12% on "Classification and code of disciplines of Ministry of Education", 9% on "Classification and code of occupation" [2].

The survey indicates that there is no uniform classification and codification solution to the S&T talents information. Most databases adopted the existing national or ministry standards, which could only indicate part of information and not a good solution to support S&T talents information sharing and management. Hence designing a solution to classification and codification of S&T talents information would be very important to information standardization.

This paper designed a classification and codification solution to S&T talents information according to property characteristics of information, which would be an optimization, integration, updating, and supplement to existing classification and codification methods.

In 2008, Institute of Scientific and Technical Information of China initiated a "China High-level S&T Talents Database", and all the information about the high level, and innovative S&T talents was included in the database.

1 The Contents and Characteristics of S&T Talents Information

1.1 The Contents of S&T Talents Information

S&T talents are the people who contribute to the development of S&T and society progress with their innovation capacity and spirit of scientific exploration. S&T talents have four characteristics: special knowledge and techniques; engaging in S&T work; high innovation capacity; great contribution to society [3].

S&T talents information is the data, characters, signs, and pictures representing S&T talents basic information, such as their personal information, knowledge and techniques, academic activities , scientific productions and credibility. In this paper S&T talents information only referred to every talent's individual information, management information is not included. The information includes the following contents:

1) Personal information, name, age, gender etc.
2) Knowledge and techniques, education background, training experience etc.
3) Academic activities, working experience, part time job experience, project reviewing, scientific research team information etc.
4) Research production: awards and honors, research papers, patents, projects, monograph etc.
5) Research credibility: academic ethics, research misconduct behavior, credibility information etc.

1.2 Characteristics of S&T Talents Information

1) Broad information, such as experience, achievements, research ethics, awards etc.;
2) Different types of information, different information needs different codification system;
3) Different classes, classifying the S&T talents information by property characteristics such as S&T tal-

978-1-935068-76-1

ents institutions, regions, subjects, fields, relationships etc..

2 Classification and Codification of S&T Talents Information

Since 1979 classification and codification work is developing rapidly in China. Different ministries establish the classification and codification standard one after another. The codification systems related to S&T talents information are summarized in the table 1.

Table 1. Codification systems and their contents related to S&T talents information in China

Standard levels	*Contents of the codification system*
National Standard	Gender(GB/T 2261.1-2003); Country (GB/T 2659-2000); Regions in China(GB/T 2260-2002); Nationality(GB/T 3304-1991); Industry (GB/T 4754-2002); Political identity(GB/T 4762-198)、Language(GB/T 4880.1-2005); Occupation(GB/T 6565-1999); Degree (GB/T 6864-2003); Position (GB/T 8561-2001); Types of Awards(GB/T 8563.1-2005); Types of Honors(GB/T 8563.2-2005), subjects(GB/T 13745-1992)
Ministry of Education Standard	Subject Classification of the Ministry of Education, Code for higher learning and research institutions
University/college standard	Supervisor, monograph, patent
Institutions of Higher learning standard	Technology field, S&T awards, Publications、project type, sources of projects
Chinese Library Classification committee	Chinese library classification
National Natural Science Foundation of China	National Natural Science Foundation of China Classification
Ministry of Science and Technology	National Torch Program preferential area of development classification

The establishment of classification and codification system mentioned in table 1 is according to the industry, subject, literature, region, occupation, position, title, research project, National Science and Technology Planning Project. They can't be directly applied to the classification and codification of S&T talents information because of the following reasons:

1) Lack of top planning. In the existing classification and codification system only some could be used but can't meet all the codification demands.
2) Some Classification is out of date and needs to be updated, such as awards /honors, institutions, project types, source of projects.
3) Some information is absent. There is no code for awards/honors, institution types, scientific credibility etc.

3 Principles of Classification and Codification of S&T Talents Information

3.1 General Principles

1) Practicability principle: connecting the demands and objectives of S&T talents information in order to make classification easy to use;
2) Compatibility principle: according to the process of "International /National standard—industry standard—Ministry standard—update—initiate", preferring to use the existing standard or updating it in order make it popularize;
3) Stability principle: choosing the most stable property of S&T talents information as the basis of classification;
4) System principle: categorizing according to the hierarchy of concepts in order to avoid the overlaps, repetition, omission of different hierarchies, forming a complete and consistent classification system;
5) Extending principle: Sparing enough codes in order to add more new items and concepts to ensure extension and segmentation the concepts;
6) Flexibility principle: allowing adding prefix or suffix to the code in order to meet the demands.

3.2 Special Principles

According to the characteristics of S&T talents information, its classification and codification should accord with the following principles.

1) Multi-dimension principle, selecting important properties to be signs of classification and codification in order to locate the amount of S&T talents information;
2) Integration of similar items principle, merger the same property of the information and in order to establish a consistent system;
3) Differentiation principle, establish different sub system of classification according to the differentiation of S&T talents information.

4 Design of Classification and Codification of S&T Talents Information

4.1 Design of Solution Framework

Establish the classification and codification solution framework according to the principles of S&T talents information (Table 2).

Table 2. Framework of S&T talents information classification and codification solution

Objective	*Name of code*	*Function*	*Source*
1. Personal information	Gender code	auxiliary sign	National standard
	China nationality code	auxiliary sign	National standard
	Political identity code	auxiliary sign	National standard
	Language code	auxiliary sign	National standard

2. Institution	Code for higher learning and research institution	Retrieval Level 1	Ministry of Education standard and updating
	Code for institution type	auxiliary sign	Self-build
3. Region code	Codes for the administrative divisions of china	Retrieval Level 2	National standard
	Codes for the representation of names of countries and region	auxiliary sign	National standard
4. Subject/ degree	Classification and code disciplines	Retrieval Level 3	National standard
	National Natural Science Foundation Application Code	auxiliary sign	National Natural Science Foundation
	Degree code	auxiliary sign	National standard
5. Technology	S&T Planning Program preferential area of development classification	Retrieval Level 4	Ministry of Science and Technology
6. Awards/honor	Awards/honor code	Retrieval Level 5	Self-build
	Types of award/ honor code	auxiliary sign	National standard
7. Industry/ occupation	Industrial classification and codes for national economic activities	auxiliary sign	National standard
	Classification and codes of occupations	auxiliary sign	National standard
8. Title/ position	Code of professional technical position	auxiliary sign	National standard
9. Monograph	Code of monograph types	auxiliary sign	Industry standard
10. Patent	International Patent Classification Code	auxiliary sign	International Patent classification Agreement
	Patent code(types, legal status, approval)	auxiliary sign	Industry standard
11. Journal paper in China	Item code on publication	auxiliary sign	Institutions of higher learning standard
	Chinese library classification	auxiliary sign	Chinese library classification committee
12. SCI papers	Subject type	auxiliary sign	SCI
13. Project	Project type code	auxiliary sign	Self-build
	Project source code	auxiliary sign	Self-build
14. Scientific credibility	Classification code of scientific misconduct	auxiliary sign	Self-build

The whole classification and codification solution is made up of 14 objectives and 26 sets of code systems.

The top level of the framework is divided into 14 objectives, including personal information, institutions, region code, subject/degree, technology field, award/honor, industry/occupation, position /title, monograph, patent, Chinese journal paper, SCI paper, project, scientific credibility.

26 sets of code systems could describe S&T talents information in a specific dimension; every code includes a concept and represents the implication and extension of the concept.

The 26 sets of classification systems have three parts, one is the existing national, ministry, institutional code, and the other part is the supplement code system, and another part is self -built code.

1) Update and supplement out-of-date Ministry of Education code "Institutions of higher learning and scientific research code";
2) Design and establish the awards/honors, project types, project source code that changed dramatically.
3) Self-built the code for types of institutions, research credibility that is absent from the existing code system.

In addition, according to the retrieval function of the classification and codification solution, it could be divided into two categories,

1) Retrieval sign: selecting five sets of code such as institution, region, subject, technology field, award/ honor to locate and retrieve the S&T talents information.
2) Auxiliary sign: the rest of 21 sets of code could be the auxiliary sign to the S&T talents information.

All these codes are relate and rely on each other in order to meet the demands of S&T talents information resources planning, signs of data sets, and information system construction etc.

4.2 Classification and Codification Design

The classification and codification system includes three parts, identification of objectives, information classification, and information codification. Taking the award/honor classification system for instance, the design process are as the following.

1) Identification of objectives: the objectives are the S&T talents with the awards and honors because of their contribution to scientific discovery, technology inventions or other scientific productions [4].
2) Hierarchies of information: according to the hierarchies of awarding institutions, the awards/honors could be divided into five types, such as national awards, industry/ministry awards, provincial awards, non-government awards, and important overseas awards. The awards type would be expanded if necessary.
3) Information codification: information code takes the structure of "type code + region code + sequence code" and has 9 numbers.
   - Type code. The first 2 letters represent the types of awards, PA means national awards, PB means industry/ministry awards, PC means provincial

978-1-935068-76-1 © 2011 SciRes.

awards, PD means non-government awards, PF means important overseas awards.

- Region code. The 3 numbers in the middle mean region. For China's awards, according to the Codes for the administrative divisions of the people's republic of china selecting the first 3 numbers of each code to be the region code, for overseas awards selecting 3 corresponding numbers from Codes for the representation of names of countries and regions.
- Sequence code. The last 4 numbers mean the sequence code. Giving a sequence of the awards in the same region and a sequent number for each award.

× ×	× × ×	× × × ×
Award Type Code	Region Code	Sequence Code

5 Conclusion

The classification and codification solution to S&T talents information represents characteristics of the resources, could organize, reveal the information of S&T talents resources well, could reduce the redundancy of the data and meet the demands of database building. The solution is proved to work well after its application in building of "China High level S&T talents database".

A code solution design team has been founded to ensure the work working well. This team took in charge of the planning, developing, testing, examining and supervising the implementation of the coding work. Once the classification and codification solution is established, it will be the standards that all the departments need to obey.

References

[1] National Bureau of Statistics, China statistical Yearbook on Science and Technology (2009), Beijing: China Statistics Press, 2010.

[2] Research on integration mechanism of S&T talents information, Supported by National Soft Science Research Program Report (Number: 2008GXS4K062),2010.

[3] MC. Liu, CR Duan, Q. Wang, and JM, Zhang, Dictionary of Talent. Chengdu: Sichuan Academy of Social Sciences Press, 1987.

[4] GD. Ning, and TW. Zhang, "On the Legal System of Technology Award in China", Technology and Innovation Management, vol. (3), pp.235-237, 2008.

Scientific Research

Construction and Application of Academic Identifier for Science & Technology Talent

Jie PENG[1], Baoqiang QU[1], Haiyun CAO[2], Yunhong WANG[1]

[1]*Center for Resource Sharing and Promotion, Institute of Scientific and Technological Information of China, Beijing, China*

[2]*York University Libraries, Toronto, Canada*

Email: pengj@istic.ac.cn

Abstract: The paper starts off the common problem of name misidentification in constructing of science & technology talent integration. It puts forward the science & technology talent academic identifier as the unique identifier of science & technology talent in engaging scientific research and studying activities, conceives the framework of AID, and then analyzes its functionality in managing science & technology activities, appraising science & technology talent as well as building the integrity system of science & technology talent. At last it gives the indispensable conditions in popularizing AID.

Keywords: science & technology talent; academic identifier; information integration; information correlation

1 Introduction

The development of advanced science & technology has been the key factor in changing the way of economic development in the 21st century. Human resource in science & technology, especially science & technology talent, has been the most active resource in scientific activities. They play an essential role in technology renovation, dissemination of advanced culture and social development. The information service of science & technology talent became particularly important in such a context. It is necessary to develop a systematic solution for science & technology talent information integration to meet the multi-angle information need on science & technology talent so as to facilitate the science & technology administration, scientific institutions and scientific enterprise as well as science & technology talent themselves in career development, fund application and evaluation, communication and collaboration. Such a mechanism is to be built with science & technology talent as the center and substantial data as the knowledge base in the purpose of implementing correct linking of science & technology talent to their publications, patents, research outputs as well as other scientific activities. Nevertheless, disambiguating the identity of science & technology talent and attribution has long been a persistent, crucial problem in information integration and scholarly communication.

In recent years, some international researchers started to pay attention to this issue. "Earning a name for yourself in science can be a struggle" began a recent Random Samples article in Science[1]. Vu, Quang Minh[2] proposed to deal with personal name disambiguation in online information indexing by using web directories. Enserink, Martin and Cals, Jochen[3-5] discussed this issue at the Science Magazine and the Lancet. They pointed that it would be much easier if each scientist had his own unique identifier. The key point was developing an infrastructure for such a unique identifier system. Jeffrey Beall[6], Wolinsky, Howard[7], Ptolemy, AS[8], Wang, L[9]also explored the possibility of researcher identifier system. Another similar research focused on Chinese name indexing in medical publications[10]. Overall, there already have enough international researches focusing on unique identifier for scientist and researchers. However, there is no specialized statewide identifier system in China for science & technology talent to participate in research activities except identity card number (personal identity information is confidential so not suitable for exposed in information system of public service). This paper conceives the framework of national science & technology talent unique identifier and look forward to the application prospect in scientific management, science & technology talent appreciation and research integrity.

2 Science & Technology Talent Academic Identifier and Coding

2.1 Existing Unique Identifier Systems

There are several mature identifier systems in research activities and scientific products including academic publication, patent, research project, dissertation and research report, such as ISBN (International Standard Book Number) for monograph publication, ISSN (International Standard Serial Number) for academic journals, patent number, research project numbers (State natural sciences fund, 863, 973, Science and technology support program), etc. Dissertations, scientific reports and research findings also have corresponding identifier systems. DOI (Digital object unique identifier) was developed for disambiguating digital objects. Identity card number is used for identifying personal identity.

At the international level, ORCID is a non-profit initiative aiming at establishing an open independent registry for researcher name disambiguation. This effort was initi-

978-1-935068-76-1

Table 1. Existing unique identifier systems

Identity	Unique identifier	Identifier registration	Coding rule
Monograph	ISBN	International ISBN Agency	Consists of "ISBN" and 10-13 digits. The digits represent group number, publication number, publisher number and check code.
Academic journal	ISSN	International ISBN Agency	Consists of 8 digits in two sections with 4 digits in each section and connected with "-". The last digit is check code
Patent	Patent number	State Intellectual Property Office	Consists of application number and document number
Science and Technology plan project	Science and Technology plan project number	Plan project administration	Separate numbering system for each plan
Digital object	DOI	International DOI Foundation	<DIR>.<REG>/<DSS>. <DIR> is designated code for DOI with value of 10. <REG> is the 4 digit registration agency. <DSS> is assigned by DOI registration agency academic publisher.
Identity card	Identity card number	Public Security	Consists of code of province, region/city, county/district, birth date, serial number and check code (18 digits).
Researcher	ORCID	ORCID	
Name	(ISNI) International Standard Name Identifier	ISNI International Agency	Consists of 16 digits with 4 sections. Each section has 4 digits. The last digit is check code. E.g. ISNI 1422 4586 3573 0476

ated by a group of most influential stakeholders in scholarly communication including universities, funding organizations, societies, publishers and corporations from around the globe, in the purpose of establishing researcher identifier standard and registry as well as ensuring open access, global communication and researcher privacy. The standard and registry will provide substantial support for the transition from science to e-science; improve the accuracy of information retrieval and develop new information services. Thomson Reuters and Nature Publishing Group convened the first Name Identifier Summit in Cambridge, MA in November 2009. One of the major topics was name ambiguity in scholarly communication and its possible solution. The summit decided to launch ORCID initiative. The ORCID initiative officially launched as a non-profit organization in August 2010 and has been moving ahead with broad stakeholder participation[11]. As part of the National Knowledge Infrastructure in Holland, the International Standard Name Identifier (ISNI) is a draft ISO Standard as a global identification system of public identities of parties, which is involved throughout the publication and media content industries in the creation, production, management, and content distribution chains. The draft of ISNI (DIS) was approved with 100% support in March 2010. The ISNI International Agency was officially incorporated as a not-for-profit organization on Dec. 2010 and the Agency is establishing an infrastructure that includes a central registry for assigning and managing the identifiers[12]. In January 2008, Thomson Reuters released ResearcherID, a researcher identification system for disambiguation of same names, different name abbreviations, same initials and typographical errors in names. Researchers are suggested to add their researcherID to their articles, webpages and blogs. Integrated application of DOI and ResearcherID will create a unique connection between the author and his/her scholarly publications [13].

2.2 Academic Identifier and Coding Rule

Some existing academic identifier systems are commercial products, such as ResearcherID; others are only applicable in certain geographical areas, such as DAI. Even though ISNI, the international standard, will be officially released in the near future, it covers not only academic activities but a wider scope involving media contents industries. According to the requirements of science & technology talent information integration and information retrieval in science and technology management in our country, this paper proposes to establish permanent unique identifiers, that is, Academic Identifier for Science & Technology Talent (AID) to facilitate their scientific research, scholarly communication and collaboration and science & technology talent management. AID aims at identifying science & technology talent's identity and correctly linking to relevant information. Once a researcher's AID is confirmed, it will become the only unique identifier for his/her scientific activities. AID must be registered, approved and managed by official science & technology administration (organized by the Ministry of Science & Technology and implemented by the Center of Communion and Exploitation of Science & Technology talent). The science & technology administration will also be in charge of developing the interface of AID registration, data infrastructure, disambiguation algorithm, as well as the design, development and management of the database distribution system. The development will integrate existing research output and progress, as well as international standards. This paper will mainly discuss the coding rule of AID.

The coding rule of AID is developed according to the existing coding rule of unique identifier and based on the constitutional system of personal identity and scholarly activities of science & technology talent. AID consists of the following components: country code (3 digits) + region code (2 digits) + discipline code (3 digits) + birth year (4 digits) + gender code (1 digit) + sequence code (2 digits) + extension code (2 digits) + check code (1 digit), totally 18 digits.

In Figure 1, country code represents the science & technology talent's nationality using International standard country code (ISO 3166). Currently 244 countries

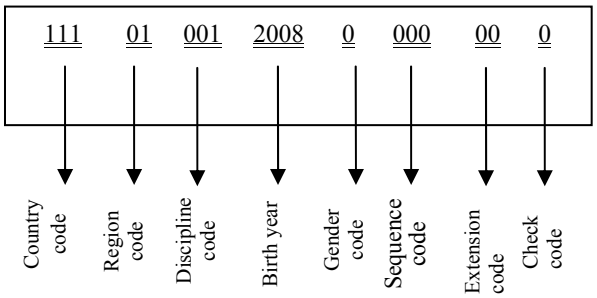


Figure 1. AID coding rule

have been listed in ISO 3166-1 code table. Each country is assigned 3 types of code, 2 digit alphabetic code, 3 digit alphabetic code and 3 digit numeric code created by the United Nations Statistical Office. The 3 digit numeric code is applied in this paper. Region code represents the province and city where the science & technology talent affiliates using the administrative division code of the People's Republic of China, GB/T 2260-1999. This standard was released by the State Statistics Bureau, which lists the administrative division code at and above the county level. Each province (autonomous region, municipalities and special administrative region) is assigned 2 digit numeric code or 2 digit alphabetic code (pinyin abbreviation). The 2 digit numeric code is applied in this paper. Discipline code represents the discipline that the science & technology talent's scholarly area using the national standard of classification and code of first level disciplines (GB/T 13745-92). First level discipline code consists of 3 numeric digits with totally 58 first level disciplines. Birth year code uses 4 digits. Gender code uses 1 for male and 0 for female. Sequence code starts from 001 to 999. Extension code has 2 digits with 00 as default value. Extension code will be used when sequence code is larger than 999 or extra information is needed for further identification. The last digit of AID is random check code.

Similar to other identifier, the extension code of AID can be adjusted for further upgrade according to the application situation. For example, Prof. Zhang, Beijing University, Faculty of Mathematics, male, born in 1960, research interests in probability theory. His AID will be 1561111019601001001. 156 is China; 11 is Beijing; 110 is discipline of mathematics; 001 is sequence code; 00 is extension code; 1 is check code. In case sequence code is larger than 999, e.g. 1008, then his AID will be 1561111019601100801.

2.3 Function of AID

AID works as the unique identifier in the information infrastructure of science & technology talent management. This unique identifier will improve the information integration of science & technology talent, provide accurate and reliable connection and offer an indexing and retrieving solution for science & technology talent information. The implementation of AID can be achieved by adding AID to existing varieties of science & technology talent databases. This paper divides science & technology talent information into 3 categories: basic information (e.g. different level of scientific administration and human resources and basic personal information of science & technology talent in enterprises and public institutions), research achievement information (academic journals, dissertations, conference proceedings, monographs, research reports, research outputs, patents and standards, etc.), Current scholarly activities (recruitment, communication and collaboration, etc.). Varieties of databases can be connected and integrated together by AID including different levels of scientific administration and human resources, isolated and scattered databases with basic personal information of science & technology talent in enterprises and institutions, Chinese e-journal database, Chinese dissertation database, Chinese conference proceeding database, Chinese science & technology achievement database, Chinese patent and national standard database, Chinese scientific & technological reward database, Chinese science & technology monograph database and Chinese science & technology report database, etc. [14]. Linking AID to the achievement and scholarly activities of science & technology talent during the integration of science & technology talent information will resolve the problem of name misidentification and misconnection, therefore, improve the efficiency and accuracy of information integration.

AID will facilitate retrieving someone's complete information and improve the efficiency of information integration and application in storing, integrating, connecting and retrieving science & technology talent information, as well as prevent name misidentification and confusion. For example, one can obtain author's AID by searching author's name and then accurately search all relevant databases with author's AID to retrieve all the information including basic information, published journal articles and monographs, research rewards and research projects, etc. for integration and application. Vice versa, searching research achievements and scholarly activities can also obtain AID of researchers, therefore, retrieve complete scholarly activities of the researchers and compare with other researchers. By doing so, the connection either from researchers to their activities or from their activities to researchers is achieved. The purpose of such a mechanism is pursuing highly efficient information integration and collaboration and reducing repeated information creation.

3 Using the Template

AID is the unique identifier for science & technology talent in their scholarly activities. It is expected to have a broad prospect in scientific research management, evaluation of science & technology talent and research integrity.

978-1-935068-76-1 © 2011 SciRes.

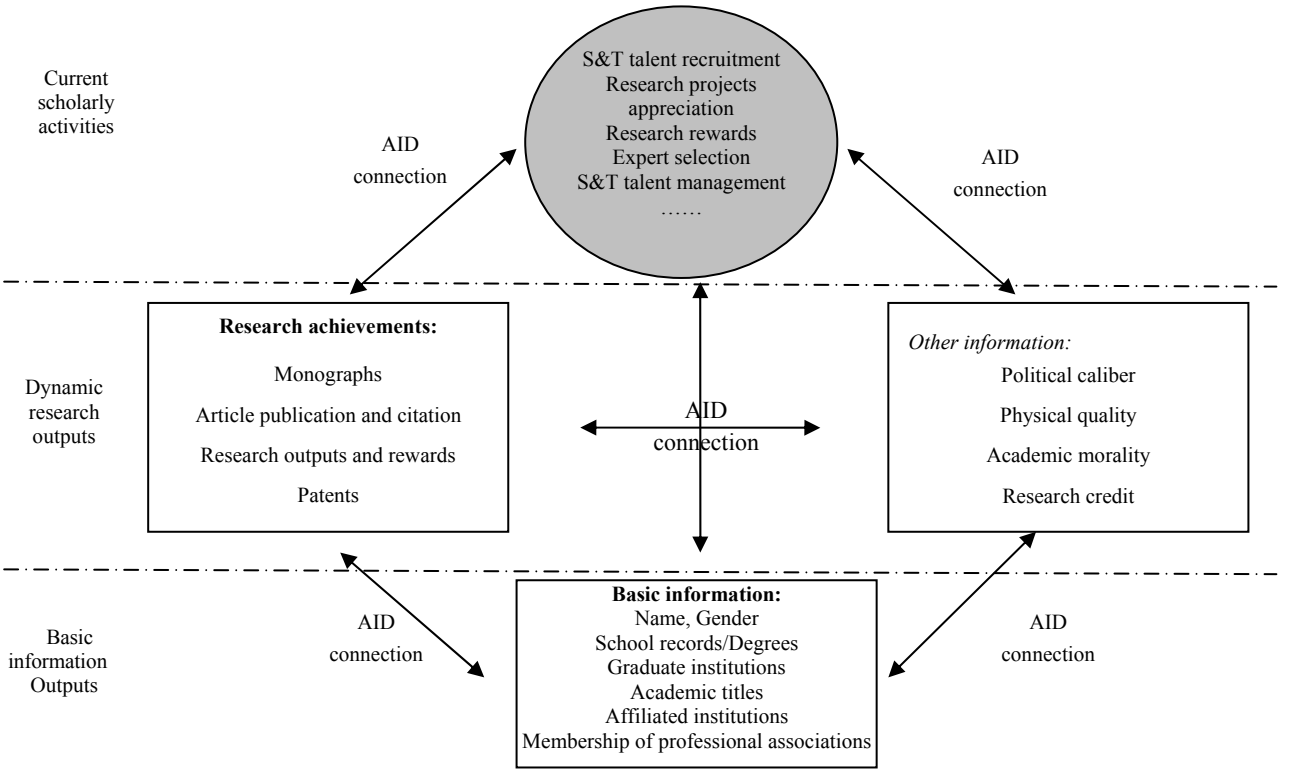


Figure 2. Connection and integration of science & technology talent information

3.1 Scientific Research Management

In the scientific research community, scientific research administration requires and uses significant amount of science & technology talent information. The administration activities in government agencies, enterprises and research institutions include recruitment and selection of science & technology talent, research project appreciation, research reward management, science & technology expert selection and science & technology talent management. The application of AID is expected to provide up-to-date, accurate reference for decision-making process, as well as optimize management workflow and enhance the efficiency of science & technology plan management[15].

With the application of AID, different levels and types of scholarly information and their relations can be retrieved, analyzed and processed. On one hand, The management of science & technology talent information in science & technology administration can be standardized and modularized so that more effort can be concentrated on processing of current scientific research information in order to track down the scholarly activities of science & technology talent in a simultaneous and dynamical attitude and implement scientification and accuracy of scientific research management. On the other hand, AID will help to reduce the trivial routine in application, evaluation and assessment of science & technology talent's academic activities including applying for research grants and rewards and career changing. Relevant human resource databases, publication databases and research output and reward databases will be automatically connected through AID resulting in advanced working efficiency.

3.2 Evaluation of Science and Technology Talent

The evaluation of science & technology talent is usually assigned to the trustee based on the client's interests and following prescriptive principle, process and standards. Productivity, academic ethics, expertise and qualification of science & technology talent will be evaluated[16]. Currently the evaluation mainly considers academic degrees, academic titles, publications, research grants and rewards of science & technology talent. Complete and accurate academic profiles play a crucial role in the evaluation process. The application of AID will be helpful in establishing complete and accurate academic profiles for science & technology talent and ensure the consistency of profile contents including personal information, scientific achievement and scholarly activities, peer-reviewing and self-reviewing, etc.,[17] therefore, guarantee the quality and accuracy of the evaluation (annual scholarly output and academic career). With the presence of AID, evaluation criteria and process can also be recorded in the academic profile and become important evidence of academic integrity, thus, improve the accountability and authenticity of the evaluation. From this

point of view, collecting, organizing and integrating evaluation information of science & technology talent by AID may enhance the current evaluation criteria and methodology that mainly focus on research productivity but neglect the subjective feedback from peers and experts in the same academic field or same discipline.

3.3 Building Research Integrity

Promoting academic integrity and preventing academic dishonesty is a global issue. The Ministry of Science and Technology established the Office of Research Integrity and released "Policy of preventing and handling academic dishonesty" [18]. Science foundations, scientific research institutions and universities also established corresponding management policies. Assigning AID to science & technology talent will not only facilitate but also provide technical support to integrity management of science and technology administration and science & technology talent.

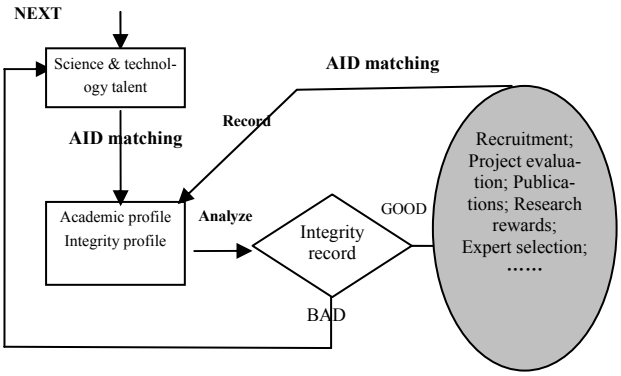


Figure 3. AID application in building research integrity

The strength of AID application in research fund application is obvious. Firstly, the complete research activities of science & technology talent can be tracked with AID, but actual personal name, when he/she applies for research fund. Through a pertinent key, all his/her basic information and previous research outputs is correlated, therefore, only the information of current project need to be included in his/her application and background information is not needed, which will simplify the workflow of research plan and research fund application, as well as implement anonymous application of research fund in order to avoid the influence of human intervention. Secondly, the application of AID can promote the expert evaluation system. Recording the evaluation process into the evaluator's research profile and integrating into research integrity assessment will guarantee the authenticity of evaluation. Thirdly, AID can help to analyze science & technology talent's research profile and discern the truthfulness of his/her application. Fourthly, AID can help to quickly identify illegal practices, such as one applicant submitting multiple applications or one applicant participating in multiple projects[19]. Finally, the application of AID may help to identify, from another angle, whether the author's expertise fall into this research area by analyzing the correlated academic background information of the author so as to prevent academic dishonesty behaviors, such as false author, plagiarism and false application, etc.

4 Implementation of AID

The application of AID for Science & technology talent bears a broad prospect. The implementation can be achieved from the following aspects. Firstly, Promoting AID to science & technology talent and research institutions and helping them realize the benefit and importance of AID, so that they are willing to voluntarily update and maintain their profile and relevant information through their AID. By doing so, science & technology talent will establish their own academic profile. In research institutions and scientific planning administration, at the same time, workflows will be optimized and operations will be more efficient. The key point of AID implementation rests within the acknowledgement of stakeholders. Secondly, Establishing AID standard system. The AID standard and associated analytical system should keep consistent in all databases of science & technology talent information. Thirdly, Making policy on AID implementation, application and management. Designated AID Administration should be in charge of AID promotion and application (e.g. the Center of Communion and Exploitation of Science & Technology talent of the Ministry of Science & Technology). The AID application can be included in the routine of research management, such as listing AID as a mandatory condition in application for science & technology research projects and research fund, or establishing compensation mechanism for science & technology talent, research planning administration and foundations to implement and promote AID, as a result, they will have the motive to take AID into account. Finally, Improving the information integration service for science & technology talent, establishing the coordinated mechanism among information holding institutions, publishers, scientific administration and science & technology talent, and forming the infrastructure of AID coding, linking, management and revision.

AID offers an effective solution for name disambiguation, data integration and application of science & technology talent information. Its implementation, application and optimization will require long-term collaborated effort of science & technology talent, research institutions and scientific administration.

Acknowledgment

This paper is supported by The National Soft Science Research Program, Science & Technology Talent Information Integration System (No. 2008GXS4K062).

References

[1] Burton, K.W., "One wei or another [in Random Samples]". Sci-

ence, 2008, vol. 319, pp.881.
[2] Vu QM, Takasu A, Adachi J. "Improving the performance of personal name disambiguation using web directories." Information Processing & Management, 2009, vol. 44, pp.1546-1561.
[3] Enserink, Martin. "Scientific publishing: are you ready to become a number?" Science, 2009, vol. 323, pp.1662–1664.
[4] Cals, J.W.L. & Kotz, D. "An author identification system for vast coverage of researchers." Science, 2009, vol. 323, pp.1662–1664.
[5] Cals, Jochen WL, Daniel Kotz. "Researcher identification: the right needle in the haystack." The Lancet, 2008, vol. 371, pp.2152-2153.
[6] Jeffrey Beall. "Cataloguing names the old-fashioned way."Science, 2009, vol. 19, pp.1514-1515.
[7] Wolinsky, Howard. "What's in a name?" EMBO reports, 2008, vol. 9, pp.1171–1174.
[8] Ptolemy, AS. "Expanding and improving research attribution with the open researcher and contributor identification initiative." Clinical Chemistry, 2010, vol. 56, pp.876-877.
[9] Wang, L. (2008). "Lost in transliteration." Science, vol. 320, pp.745.
[10] Cheng, T.O. "Chinese personal name variation in medical publications: Implications for information retrieval." Int. J. Cardiol., 2010, vol. 139, pp.211-212.
[11] http://www.orcid.org.
[12] http://www.ISNI.org.
[13] http://www.researcherid.com.
[14] He, Defang. "A Study on the Dynamic Evaluation Model of Scientific and Technical Talents based on Knowledge Network." China Soft Science, 2005(6), pp.47-53.
[15] Yan, Tinglan. "The application of advanced scientific & technological information in scientific research management." Journal of Shandong Youth Administrative Cadres' College: Youth work forum, 2010(5), pp.89-90.
[16] You, Rongguang. "The urgent need of establishing national science & technology evaluation system." Management and Review of Social Sciences, 2002(1), pp.35-39.
[17] Li, Ying. "The current situation and reformation thoughts on talent profiles in implementation of talent strategy." Modern Management Science, 2003(10), pp.93-94.
[18] "The Ministry of Science & Technology analyzed the source of academic dishonesty and proposed 5 prevention solutions." http://tech.qq.com/a/20060620/000020.htm [2008/7/20].
[19] "Research integrity and research profile." Chinese University Technology Transfer, 2007(8) , pp.48-49.

Value Relationships in Sharing of Scientific and Technical Information Resources

Wei ZHAO, Baoqiang QU

Center for Resource Sharing and Promotion, Institute of Scientific and Technological Information of China, Beijing, China

Email: zhaowei@istic.ac.cn, Qubq@istic.ac.cn

Abstract: Scientific and technical information resources (STIR) are important support and basic security for development and innovation of national science and technology. Complex value relationships are formed during various processes of STIR sharing. This study investigated the relationships between the value subjects (including the producers and users of value), and the relationships between value subject and value object based on the analysis of the features and compositions of value subject and object in STIR sharing. Furthermore, suggestions on coordinating the value relationships during STIR sharing were presented.

Keywords: Scientific and technical information resources (STIR); value relationship; subject; object

In the processes of production, conservation and use of STIR, the flowing and delivering of resource value are accompanied with all the time. It is just because of the value of STIR with properties, and the value realization during transmission, which makes the construction of STIR and sharing service possible[1]. Meanwhile, complex value relationships are formed among various elements in sharing, including the relationships between STIR subject and object, subject and subject, object and object. On the different stages of sharing, the situations of subjects and objects are not the same, and the value relationships will undergo changes. Clear value relationships in STIR sharing are of important guiding significance for further discovering the features and basic laws of value transmission and increment, as well as for exploring the driving force and mode selection of STIR sharing service[2].

1 Identification of Value Subject and Object in STIR Sharing

Based on Marx's value theory of labor, labor is the only source of value[3]. Since the production of STIRs is the result of tireless labor of scientists, STIR is valuable. The value of STIR actually refers to the role or potential of specific STIR that can meet the requirement of scientific research, innovation and social development.

Value flowing of STIR exists in the entire process of resource production, conservation, transfer and use, and plays important role in various forms[4]. The track of value flowing in STIR management and sharing is presented in Figure 1.

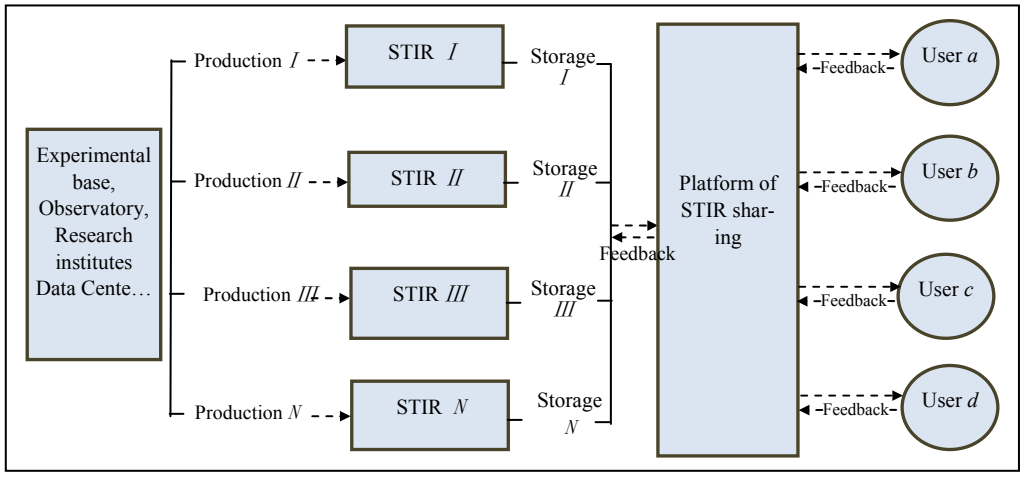


Figure 1. The track of value flowing in STIR management and sharing

The value flowing process in STIR construction and usage includes the processes of production, analysis, conservation, sharing and service. Through the sharing platform of STIR, the clients (including parts of resource producers and users) can more effectively utilize the STIR, and they can also feedback the utilization situation and further requirement of STIR to the platform, even directly feedback to the construction process of STIR, so as to

978-1-935068-76-1

provide guidance for the production and conservation of STIR.

It can be seen that in the process of STIR sharing, the value elements are composed of producers, managers, transmitters, users and information resources themselves.

1.1 Value Subject

In the sharing process of STIR, the value subjects refer to the STIR clients (users of various STIRs), resource managers, information transmitters and producers (the original subject of various STIRs). The different STIR subjects have different information requirements.

Producers and transmitters of information resources should be familiar with and master modern knowledge of Library and Information Science, have the development capabilities for STIR and a strong grasp of modern information technologies and modern equipment operating skills.

The user's requirement is the fundamental driving force for the service organization to carry out STIR sharing and promote the value increment[5]. In view of the innovation process, the client requirement can be divided into information needs before work, at work and after work. And the contents and scope of information needs include object information needs, information needs on the progress of industry research, collaboration and exchange information needs, equipment and materials information needs, policy and law information needs, and talents information needs, etc. The information needs from clients have following characteristics: (1) Advancement, means that the information should reflect the frontiers and hot spots of the discipline; (2) Timeliness, means that the information should be collected and refined promptly and transferred conveniently, so that the researchers can get the latest information, which is very important for scientific innovation; (3) Novelty, means that the innovative information should be provided, and the STIR producers must have a strong sense of competition and information awareness, in the meantime with the capacities of judging, analyzing, identifying, collating, disseminating and processing information. In addition, the clients may have particular interest in the newly published papers, monographs, conference materials, academic trends, scientific data, emerging cross-disciplinary and other information, and they emphasize the accuracy, completeness and continuation. Especially, they have special needs for internal reporting, special reports, statistics and dynamic analysis, and demand high precision for information retrieval.

1.2 Value Object

Value object is the object of value to be used, that is the STIRs themselves. For information resource, the academic communities at home and abroad have different understandings, which can be divided into the narrow and the broad claims. Narrow interpretation: information resource is a collection of a large number of useful information accumulated through the processing and ordering in the development of human society. Broad interpretation: information resource is the collection of the accumulated information, information producers, information technology and other elements of information activities in the development of human society[6,7]. Here we adopt the narrow interpretation and the concept of information resource is confined in the field of science and technology.

For the definition of STIR there are some different understandings. Some experts believe that STIR is the general term of all the information related to science and technology obtained from human activities. In the long-term national science and technology development plan, STIR is defined as "basic scientific and technical data, information, and various scientific data products and a variety of vector science and technology books, journals, reports, papers, patents and other scientific literature that produced facing different needs in the human scientific and technological activities" [8]. This paper argues that STIR usually refers to a variety of information resources that produced during basic scientific research and technology development and those needed in the course of the various types of science and technology activities. STIRs mainly includes scientific literature information resources, database resources and network resources, etc.

2 Relationship between Subjects of STIRs

Because of the different acquirement and cognitive ability of the subjects, during the flowing process of STIRs, the relationships between the subjects (e.g. the users and the users, the producers and the producers, the producers and the users of STIRs) are complex with both cooperation and competition exist, as well as the positive and negative feedbacks (see figure 2).

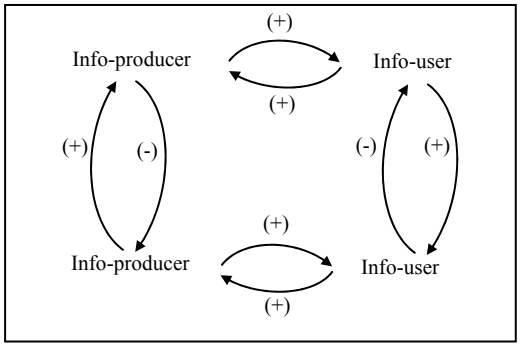


Figure 2. Positive and negative feedback relationships between subjects of STIRs

2.1. Relationship between producers

The relationship between producers of STIRs is generally in the form of competition with negative feedback. In order to develop and use information resources in the most efficient manners, there is bound to competition in

the development, organization and management of information resources. On the other hand, as STIRs are not only highly complementary, but also time-sensitive, the producers need to renew the STIRs continuously. So the producers can cooperate in the information resource production and establish positive feedback relationship, resulting in higher efficiency.

2.2 Relationship between Producer and User

When the STIR is just produced, the producer has its full value (including the value and use value), but with the transfer of STIR, the value flows towards the users. At this time, the flow of the value is only the flow of initial value, and it is also the initial performance of value form. When the STIR flows among various users and fully sharing is achieved, the users can utilize STIR from different angles and in different ways, and the value of STIR can be well reflected. The general form of value flow is like this: producer → user → user. The value of STIR can be divided into hidden value and dominant value. One part of dominant value is directly shown, and the other part is transformed from hidden value during the usage of STIR by various users.

Influence of producer to users——the workers engaged in STIRs collection, processing, transmission, integration, external service and other value added processes make the information transmission among producers, users, and between producers and users, so that STIR can be timely applied to various industries and fields, and provide basic conditions for science and technology development of the society. The producers determine the applicability of information type to the user needs, as well as preciseness, convenience and cost of information. Therefore, the influence of producer to information usage of clients is positive feedback.

Feedback from users to information producers——on the one hand, the produced STIRs can be utilized from different angles and ways by different users, and the value of STIR can be enriched continuously. However, at present Chinese clients generally have low capacity in acquiring and utilizing information, the development of information resources is difficult to work. For the network information consumers, there is still serious imbalance in the geographical distribution across the state, which directly constraints the amount of accesses to information from the Web. Influenced by the differences in user's own knowledge, intelligence and ability to express the information, the information acquirement relying mainly on manual information retrieval services of management departments to value access will also be subject to different restrictions. On the other hand, the users can also feedback the opinions on requirement of information to the producers and promote the producers to improve the information products continuously during the course of information processing. Therefore, from this perspective, the users have positive feedback to the producers.

The producers and users of STIRs can transform into each other. The producer himself of STIR is often a user. Furthermore, the new STIR may be produced based on original STIR, namely the user becomes a producer. This is the most complex relationship in STIR flowing, in which any producers or users exist relative to the STIRs they possess. It also illustrates from another aspect that during the flow and transmission of STIRs, the value of resource plays a promotion role, because in the flow and transmission of STIRs, the producers and the users are utilizing the original as well as the derived values of STIRs. The value of STIR includes two aspects, one is the added value on the original, and the other is a changed value with more abundant content.

2.3 Relationship between Users

The flow and transmission of STIRs between users may transform hidden values into dominant values, which make the value expression become clearer (see figure 3). Competition between users originates from scarcity. At the early stage of information production, restrained by the manners and technologies of information transmission, the lack of information leads to a competitive and negative feedback relationship between producers. Timeliness and the imbalance distribution of STIRs in space and time exacerbated the pattern of competition between users. With the development of technologies and enlargement of user scope, the relationship between users may gradually transform from competitive to cooperative and positive feedback.

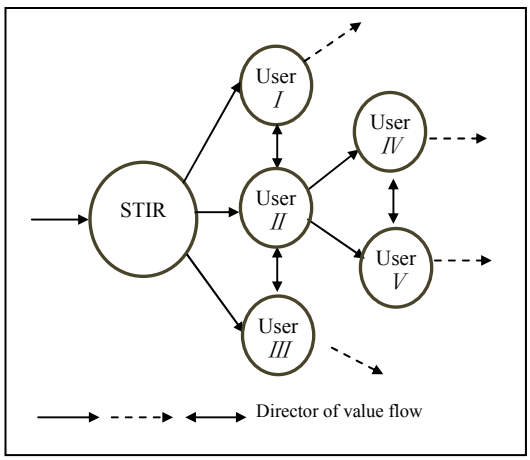


Figure 3. Flows of STIRs among users

3 Relationship between Subjects and Objects of STIRs

The relationship between STIR subject and object means the relationship between producers/users and STIRs.

978-1-935068-76-1 © 2011 SciRes.

3.1 Impact of Information on Resource Subject

STIRs will help improve the current level of science and technology and enhance the accumulation of knowledge and knowledge discovery capabilities of resource subjects. The capabilities of resource subjects are composed of the capacities of production, organization, external service and user access, as well as information mining.

During this sharing course of STIRs, the subject is constantly changing, and the content of object is becoming more abundant. From production, processing, storage to the sharing service stage, STIR is constantly changing. If the producers and the users can effectively utilize the change of STIR, its value can be better displayed, and this is beneficial for the development of science and technology. Moreover, one STIR can be used in different periods, not only for this generation, but also for next generations. The development and utilization of information is also a process of study and accumulation, which make it possible to break through the limits of physical resources, and realize sustainable development.

3.2 Impact of Resource Subject on Information

The users have different ability and they use STIRs from different way angles by different manners, so that affecting the effectiveness of resource value (including the quality and quantity). For information producers, the capacities in production, processing, storage and external service of resources directly affect the quantity and quality of information. Thus, in the flow and transmission of STIRs, the value of resource will be increased or decreased.

4 Coordination of Value Relationships in STIR Sharing

In the process of STIR sharing, there exist complex relationships. From the microcosmic angle, we need to establish a tracking service model for information integration, organization and sharing, and strengthen the communication and coordination of various relevant elements at key points. The service for information should also be consolidated, including the research service for technology innovation, effective information transfer service and network intelligence service[9].

On the other hand, from the macro- and micro- angles, we need to carry out effective STIRs distribution, adjusting the current distribution and the distribution perspective according the information requirement, distribution efficiency and effectiveness. In addition, a favorable circumstance for information exchange should be created.

Acknowledgment

This paper is supported by the National Social Science Fund (No. 09CTQ008).

References

[1] Zhao Wei, Zhao Kuitao, Peng Jie. Analysis on the value and the value manifestation of science and technology resources[J]. Studies in Science of Science, 2008, vol.3, pp. 461-465.
[2] Tan Jing, Zhou Quanxi. Dialectical Analysis of the Information Resource Value Transfer[J]. JOURNAL OF INFORMATION, 2004,vol.11, pp. 62-63.
[3] Marx and Engels. Marx and Engels Anthology (Vol. I) [M]. Beijing: People's Publishing House, 1995.
[4] Zhao Wei, Zhao Kuitao, Wang Yunhong, etc. Study on the Value Transfer in Sharing and Serving of Scientific and Technical Information Resource[J]. Science & Technology Progress and Policy,2009,vol.15, pp. 8-11.
[5] Lai Maosheng, Wang Fang. Information Economics [M]. Beijing: Peking University Press, 2006.
[6] Kudyba S, Diwan R. Research report: increasing returns to information technology[R]. Information Systems Research, 2002, vol.1, pp. 104-113.
[7] Ding Houde. The new challenges on S&T resource allocation and strategies analysis[J].Studies in Science of Science,2005,vol.4, pp. 474-480.
[8] Zhao Wei, Peng Jie, Huang Dingcheng, etc. Construction of the Index Systems for the Performance Assessment of the Nation's Science and Technology Infrastructure[J]. Science & Technology Progress and Policy,2007, vol.24, pp.131-134.
[9] Jing Jipeng, Zhang Xiangxian, Li Beiwei. Information Economics (second edition) [M]. Beijing: Science Press, 2007.

Scientific Research

Critical Factors Impacting Knowledge Sharing among R&D Organizations

Latif AL-HAKIM

Faculty of Business, University of Southern Queensland, Australia, 4350

Email: hakim@usq.edu.au

Abstract: This research explores factors affecting the level of trust impacting knowledge sharing (service) among R&D organisations. These factors are investigated in order to gauge the gap between expected importance and perceived performance rating of trust factors impacting knowledge sharing. The research presents results of a survey comprising eighteen R&D organisations. The research identifies the critical gap using three types of tests; the paired-samples t-test, the weighted mean gap analysis method and the unweighted importance-performance analysis method. Data analysis recognises six factors with critical gaps affecting knowledge sharing which can be interpreted as follows: R&D organisations perceived that their partners were slow in response to requests and orders, reluctant in customising the products and services, unpredictable in their behaviour, below the expectation in regards to performance level and, consequently, and they failed to provide products/service in timely fashion. There was also need to have well defined agreements.

Keywords: R&D organisation; importance; performance; trust; knowledge sharing

1 Introduction

Over the past few decades, the industrialised economy has fast moved from production-based economy (being based on natural resources) and information-based economy to knowledge-based economy relying on intellectual assets and knowledge service [1]. The explosion of the Internet and other telecommunication technology has change the business environment and has made real-time and on-line communication throughout the entire supply chain a reality. Currently we are experiencing the knowledge-based economy in which inter-organisational knowledge sharing (service) become necessity for business competitive advantages [2]. Knowledge service is very distinct from earlier information service stage. The distinction between the two services can be viewed from the definition of information and knowledge [3]:

Information: Information comprises data that have been organised in a manner that gives meaning for the recipients.

Knowledge: Knowledge consists of data items and /or information organised and processed to convey understanding, experience, accumulated learning, and expertise that are applicable to current problem and activity. Knowledge can be the application of data and information in making the decision.

It follows that from above definitions that having or serving knowledge implies it can be exercised to solve a problem, whereas having information does not carry the same connotation [4]. Knowledge is unique as an organisational resource in that it is intangible and most other resources tend to diminish with use, while knowledge increases with use [5] as "idea breed other ideas, and knowledge management processes stays with the giver while it enriches the receiver" [6]. With knowledge, a human being can perceive the world and interpret and response to external stimuli. Hence, knowledge often determines action. It is information that is contextual and actionable [4]. Polanyi [7] first conceptualised two kind of knowledge; explicit knowledge and tacit knowledge. This conceptualisation of knowledge is used up to know:

Explicit knowledge deals with more objectives, rational, and technical knowledge. It is a knowledge that can be codifies such as procedures, policies, reports, strategies or alike.

Tacit knowledge is usually in the domain of subjective, cognitive, and experimental learning; it is highly personal and difficult to formulise or documented. It is cumulative stores of experiences, expertise and know-how. It can be referred to as embedded knowledge [8].

Knowledge is the most valuable intangible and intellectual asset of an organisation and becomes fundamental basis for competition [9] and, particularly tacit knowledge because it is unique, imperfectly mobile, imperfectly imitable and non-sustainable [10]. Tacit knowledge may be unstructured and is difficult to document. Some tacit knowledge could be documented using advance techniques such as data mining and expert systems but has remained tacit if the other receiver may not recognise its potential value.

Inter-organisational knowledge sharing has potential to generate ideas and develop new business opportunities [11]. Recent advances in information and communication technology (ICT) have brought with it ubiquitous, real-time connectivity. These advances, coupled with globalisation, have created new form of electronic mar-

978-1-935068-76-1

kets and fuelled the adoption and proliferation of online inter-organisational knowledge sharing [5,12]. However, the availability of ICT does not automatically induce a willingness to share knowledge because knowledge could misuse or disclose to competitors. In addition, successful sharing of knowledge requires a certain context or situation [13] and tacit knowledge sharing takes place through association, internships and social and interpersonal interaction. The literature emphasises that, in addition to ICT, effective knowledge sharing requires a foundation of trust and commitment and such trust cannot be built without strategic collaboration. The consensus is that trust can contribute significantly to the long-term stability of an organization [14] and provides the required association and interpersonal interaction.

Morgan and Hunt [15] argue that when both commitment and trust are present, collaborated organisations produce outcomes that promote efficiency, productivity and effectiveness. Practices show that many firms collaborate or form strategic alliances to better their competitive position.. Lack of trust in a supply chain is claimed by Poirier [16] as "the single biggest obstacle to advancing supply chain improvement". It has also been reported that the biggest stumbling block to the success of strategic alliance formation is the lack of trust, and subsequently trust is perceived as a cornerstone of strategic partnership and knowledge sharing. Ryu et al. [17] emphasise that trust is the key to the expansion of inter-organisational relationships. The study of Ryu et al. [17], however, shows that trust in one exchange party is formed after the other party proves its abilities to offer solutions, and demonstrates its willingness to share knowledge and information. Further, building trust relies on the parties' willingness to relinquish some independence and developing mutual dependence [18].

Sharing scientific and technical information by a party may have limited benefits to the other collaborated party because of the lack of sharing expertise of how to deal with the shared information. As obvious example include disastrous during earthquakes, epidemic situations and outbreak of animal diseases such as bovine spongiform encephalopathy (BSE) and foot-and-mouth disease (FMD). This research deals with knowledge sharing between R&D organisations. In order to improve our understanding of the effect of trust on knowledge sharing between collaborated R&D organisations, this study employs exploratory quantitative approach in which eighteen (18) R&D organisations from Australia were selected. The reminder of this paper is structured as follows: First, the paper deals with trust dimensions and identifies factors impacting trust followed by section explaining the research methodology used. This leads to data analysis. This final section concludes the research.

2 Dimensions of Trust

Following Sako [19], this research distinguishes three types of trust, namely contractual trust, competence trust and goodwill trust. In addition, this research also considers benevolence, as there is a marked psychological difference between goodwill and benevolence, which for some is also a dimension of trust. Contractual trust is the belief that both parties in a relationship will adhere to universalistic ethical standards, such as honouring contracts, being honest, keeping promises made [20], and carrying out their duties as agreed [21].

Competence trust refers to faith in the abilities of the other partner to perform their role in the project [21]. It addresses the question of whether the other party is seen to be capable of doing what it says it will do. Competence trust requires a shared understanding of standards of professional conduct and technical and managerial standards [20].

Goodwill trust embodies the belief that both parties in a relationship will consider the interests of the other, regardless of formal agreements, and will avoid opportunism thereby minimizing the threat of moral hazard [21]. Goodwill trust requires consensus on what is 'fair' between the parties [20].

Boersma et al. [22] develop a process model of trust building on international joint venture relationships and find that in the early stage of an international joint venture, promissory-based trust predominates. As the joint venture progresses, competence-based trust emerges. Goodwill-based trust is important throughout the process of a joint venture.

Benevolence is the extent to which an organisation is believed to want to do good to its partner, aside from an egocentric profit motive [23]. Benevolence is the assessment that the organisation is concerned enough about its partner's welfare to either advance interests, or at the minimum not to impede them. It is understood to be of a more inter-personal nature in terms of a specific attachment between the organisation and its partners [5,21]. This study extracts 19 factors from the four trust dimensions. Table 1 lists the trust factors used in this study.

3 Research Methodology

The research employed a quantitative methodology which required the design of a questionnaire and selection of sample organisations to answer the questionnaire. The questionnaire was designed using the approach of Watson and Frolick [24] to measure the expected importance and perceived performance of trust factors. This research adopted a seven-point Likert scale to measure performance gaps. There are two reasons why a seven-point scale was chosen instead of the normal five-point scale: (a) it provides a more accurate comparison between respondents; and (b) it provides respondents with a choice for selecting an impartial answer should they become dubious of the "right" or appropriate answer. The questionnaire consisted of four sections (general and demographic, importance-performance analysis of trust,

Table 1. Trust dimensions and factors.

Trust Dimension	Code	Trust Factor	Description of Trust Questions
Contractual trust	V1	Well-detailed agreements	Well-detailed agreements with our partners, reflecting our aims from the agreements.
	V2	Timely products/services	Confidence in our partners to carry out work/provide services at an agreed time.
	V3	Standards and performance levels	Confidence in our partners to carry out work/provide services with the standards and performance as agreed.
	V4	Skills and expertise knowledge	Confidence in our partners' skills and knowledge in their area of expertise as set out in our agreements.
	V5	Products/services as per agreement	Confidence in our partners to carry out work/provide services as promised.
Competence trust	V6	Safety and quality standards	Confidence in our partners to follow safety and quality standard requirements precisely.
	V7	Need for monitoring	The need to monitor our partners' work closely to ensure conformance to safety and quality standard requirements.
	V8	Inform of any potential problems	Attempts by our partners to inform our organization of any potential problems.
	V9	Reliability of advice	Confidence on the reliability of our partners' advice.
	V10	Satisfy needs and requirements	Needs and requirements are satisfied by our partners.
Goodwill trust	V11	Actions beyond the norms	Actions beyond the norm by our partners in an attempt to help.
	V12	Business relationship development	Commitment from our partners to maintain and develop business relationships.
	V13	Products/services customization	Customization of products and services for our organization.
	V14	Dedicated resources	Investments in resources dedicated to consolidate relationships with our organization.
	V15	Sincerity and honesty	Sincerity and honesty of our partners.
Benevolence	V16	Uphold formal/informal agreements	Formal/informal agreements are maintained despite potential benefits to be reaped.
	V17	Truthful exchange of needs/facts	No exaggeration of needs or alteration of facts by our partners for self-beneficial gains.
	V18	Predictable behaviour	Performance of our partners can be accurately predicted.
	V19	Level of responsiveness	Responsiveness of our partners to changes in specification at short notice.

general questions and comments). The survey questionnaire was sent to the sample population by way of mail, facsimile or electronic mail.

As there is currently no preferred or correct method for selecting factors with critical gaps, this research used a combination of three methods of analysis to determine the existence of statistically significant differences: paired-samples t-test, weighted mean gap analysis, and unweighted importance-performance analysis (IPA).

The paired samples t-test compares the means of two variables and tests to see if the average difference (p-value) is below threshold chosen for statistical significance (usually 0.01 or 0.05 level). The weighted mean gap analysis calculates the weighted mean gap value by multiplying the importance rating of a factor by its gap value. The factors are ranked in a descending order according to their respective weighted gap value. For the unweighted IPA method, factors with gaps fell into four quadrants, labelled "Critical", "Significant", "Important" and "Necessary". Factors designated in the "Critical" quadrant require the most improvement effort, while those located in the "Necessary" quadrant require the least amount of attention (Figure 1).

For this research, a factor is critical if the performance gap falls within the following criteria; (a) a factor should obtain a value less than the 0.05 significance level required for paired-samples t-test; (b) a factor must fall

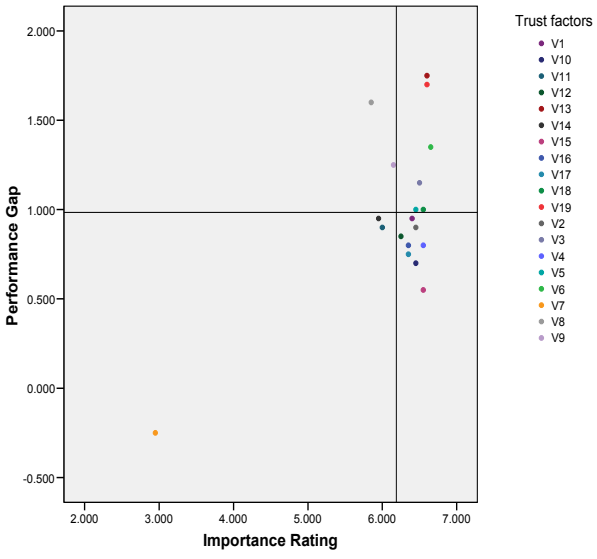


Figure 1. IPA analysis for trust factors

within the range of factors with the highest weighted mean gap values for the weighted mean gap analysis method; and (c) must fall within the "Critical" improvement quadrant for the unweighted IPA method. Only upon satisfying these three criteria, will the factor be considered for selection. The paired sample t-test is first

978-1-935068-76-1 © 2011 SciRes.

Table 2. Results of the survey.

Code	Trust Factor	Importance (EI)	Performance (PP)	Mean Gap (PG)
V1	Well-detailed agreements	6.445	5.695	0.75
V2	Timely products/services	6.580	5.53	1.05
V3	Standards and performance levels	6.500	5.65	0.85
V4	Skills and expertise knowledge	6.540	6.02	0.52
V5	Products/services as per agreement	6.410	5.77	0.64
V6	Safety and quality standards	6.450	5.86	0.59
V7	Need for monitoring	2.830	3.38	-0.55
V8	Inform of any potential problems	5.750	4.45	1.30
V9	Reliability of advice	6.250	5.30	0.95
V10	Satisfy needs and requirements	6.450	6.04	0.41
V11	Actions beyond the norms	6.000	5.40	0.6
V12	Business relationship development	6.250	5.70	0.55
V13	Products/services customization	6.590	5.20	1.39
V14	Dedicated resources	5.950	5.30	0.65
V15	Sincerity and honesty	6.530	6.35	0.18
V16	Uphold formal/informal agreements	6.350	5.85	0.50
V17	Truthful exchange of needs/facts	6.350	5.90	0.45
V18	Predictable behaviour	6.500	5.80	0.70
V19	Level of responsiveness	6.600	5.20	1.40
	Mean	6.175	5.494	0.680

applied and critical factors are identified. The weighted mean gap analysis is then used to rank the t-test determined critical factors in a descending order according to their respective value. The unweighted IPA method is then applied and only factors within the critical quadrant are selected. This methodology ensures that the selection of factors with critical gaps will be more objective. It also ranks the selected factors according to their criticality.

4 Data Analysis

Table 2 identifies the mean ratings of trust factors for expected importance (EI), perceived performance (PP) and performance gap (PG) rating, which shows the total importance rating was higher than its performance rating (EI = 6.175, PP = 5.195) with average performance gap of 0.980. The "level of responses" and "product/service customisation" shared the highest importance rating (EI about 6.600) with the highest performance gap (approximately 1.4) indicating low performance exercised by collaborated organisations. The factor "Sincerely and honesty" receive the highest performance rating of 6.35 indicating no question had been raised about sincerity and honesty of the partners. The factor "need for monitoring" has higher performance than its importance. This factor has no criticality and can be excluded without further calculation.

Table 3 shows the combined results from all three methods, denoted by Test 1, Test 2 and Test 3 respectively. Test 1 refers to the significance value obtained from the paired-samples t-test. Test 2 refers to the weighted mean gap analysis method and highlights the top ten factors with the highest weighted mean gap values. Finally, Test 3 refers to the unweighted IPA method and brings attention to the factors listed within the "Critical" improvement area (Figure 1).

Based on the guidelines proposed, Table 3 shows that only six trust factors with critical gaps: "Product/service customisation", "Level of responsiveness", "Timely products/services", "Standard and performance level", "predictable behaviour" and "Well detailed agreements".

5 Conclusion

This research investigates trust factors affecting knowledge sharing among R&D organisations.The research identifies nineteen trust factors and use methodology with three types of tests; the paired-samples t-test, the weighted mean gap analysis method and the unweighted importance-performance analysis method. Data analysis recognises six factors with critical gaps affecting knowledge sharing which can be interpreted as follows: R&D organisations perceived that their partners were slow in response to requests and orders, reluctant in customising the products and services, unpredictable in their behaviour, below the expectation in regards to performance level and, consequently, and they failed to provide products/service in timely fashion. There was also need to have well defined agreements. The results

Table 3. Results of the combined methodology.

	Trust Factor	Test 1	Test 2	Test 3	Selected factor
V13	Products/services customization	Significant	1	Critical	√
V19	Level of responsiveness	Significant	2	Critical	√
V8	Inform of any potential problems	Significant	3	Significant	
V2	Timely products/services	Significant	4	Critical	√
V9	Reliability of advice	Significant	5	Significant	
V3	Standards and performance levels	Significant	6	Critical	√
V18	Predictable behaviour	Significant	7	Critical	√
V1	Well-detailed agreements	Significant	8	Critical	√
V5	Products/services as per agreement	Significant	9	Important	
V6	Safety and quality standards	Significant	10	Important	
V14	Dedicated resources	Significant	11	Necessary	
V11	Actions beyond the norms	Not Significant	12	Necessary	
V12	Business relationship development	Significant	13	Important	
V4	Skills and expertise knowledge	Significant	14	Important	
V16	Uphold formal/informal agreements	Significant	15	Important	
V17	Truthful exchange of needs/facts	Significant	16	Important	
V10	Satisfy needs and requirements	Significant	17	Important	
V15	Sincerity and honesty	Significant	18	Important	
V7	Need for monitoring	Not significant	19	Necessary	

also indicate that the R&D organisations reasonably relied on the advice of each others, has good business development, truthfully exchange of needs and facts and satisfy needs and requirements.

References

[1] C. Tseng, and J. Goo, J. "Intellectual capital and corporate value in an emerging economy: empirical study of Taiwanese manufacturers". *R&D Management*, vol. 35, no. 2, 2005, pp. 203-215.

[2] B. Godin, B. "Knowledge-based economy: conceptual framework or buzzword". *The Journal of Technology Transfer*, vol. 31, no. 1, 2008, pp.17-30.

[3] E. Turban, J.E. Aronson, and T.P Liang, "*Decision Support Systems and Intelligent Systems",* (7th ed.). 2005, Uper Saddle River, NJ: Prentice Hall.

[4] E. Turban, J.E. Aronson, T.P. Liang, and R. Sharda, "*Decision Support Systems and Business Intelligence Systems",* (8th ed.). 2008, Pearson Education International, NJ: Prentice Hall.

[5] A.Usoro, M.W. Sharratt, E. Tsui, and S. Shekhar, "Trust as an antecedent to knowledge sharing in virtual communities of practice". *Knowledge Management research and Practice*, 5, 2007, pp.199-212.

[6] T.H. Davenport, and L. Prusak, "Working Knowledge. 1998, Boston, MA: Harvard Business School Press.

[7] M. Polanyi, "*Personal Knowledge"*. 1958, Chicago: University of Chicago Press.

[8] E.D. Tuggle, and W.E. Goldfinger, "A methodology for mining embedded knowledge from process maps. Human System Management, vol. 23, no. 1, 2004, pp. 1-13.

[9] M. Zack, "Developing a knowledge strategy". *California Management Review*, vol. 41, no. 3, 1999, pp. 125-145.

[10] C. Lopez-Nicolas, and A.L. Merono-Cerdan, "The impact of organisational culture on the use of ICT for knowledge management". *Electronic Markets*, vol. 19, 2009, pp. 211-219.

[11] H.F. Lin, "Impact of organisational support on organisational intention to facilitate knowledge". *Knowledge Management research and Practice*, vol. 4, 2006, pp. 26-35.

[12] G. Peng, and P. Woodlock, P. "The impact of network and recently effects on the adoption of e-collaboration technologies in online communities". *Electronic Markets*, vol. 19, 2011, pp. 201-210.

[13] S.G. Shariq, and M.T. Vendelo, M.T. "Tacit knowledge sharing". In D.G. Schwartz (ed.). *Encyclopaedia of Knowledge Management*. 2006, Hershey, PA: Idea Group reference.

[14] S. Ba, and P.A. Pavlou, "Evidence of the effect of trust building technology in electronic markets: Price premium and buyer behavior. MIS Quarterly, vol. 26, no. 3, 2002, pp. 243 268.

[15] R. Morgan, and S. Hunt, "The commitment-trust theory of relationship marketing". *Journal of Marketing*, vol. 58, 1994, pp. 20-38.

[16] C. Poirier, C. "*Advanced Supply Chain Management"*. 1999, San Francisco: Berrett-Koehler Publishers, Inc.

[17] S. Ryu, J.E. Park, and S. Min, "Factors of determining long-term orientation in interfirm relationships" *Journal of Business Research*, vol. 60, 2007, pp. 1225-1233.

[18] M. O'Keeffe, "Establishing supply chain partnerships: lessons from Australian agribusiness". *Supply Chain Management*, vol. 3, no. 1, 1998, pp. 5–9.

[19] M. Sako, "*Prices, Quality and Trust: Inter-firm Relations in Britain and Japan"*. 1992, Cambridge: Cambridge University Press.

[20] PMMS, "Trust: the academic's perspective, PMMS' perspective". *PMMS Asia* Pacific, 2004 Available at: http://www.pmms.com.au/whitepapers/Trust.pdf.

[21] P. Ryan, M. Giblin, and E. Walshe, "From sub-contractual R&D to joint collaboration: the role of trust in facilitating this process". *International Journal of Innovation and Technology Management*, vol. 1, no. 2, 2004, pp. 205-231.

[22] M.F. Boersma, P.J. Buckley, and P.N. Ghauri, "Trust in international joint venture relationships". *Journal of Business Research*, vol. 56, 2003, pp. 1031-1042.

[23] R. Mayer, J. Davis, and F. Schoorman, F. "An integrative model of organizational trust". *Academy of Management Review*, vol. 20, no. 3, 1995, pp. 709-734.

978-1-935068-76-1 © 2011 SciRes.

[24] H.J. Watson, and M.N. Frolick, "Executive Information Systems: Determining Information Requirements," Information Systems Management, vol 8, no. 2, 1992, pp. 37-43.

Study on the Value Attributes of Network Public Knowledge Platform and Digital Fair

Lvyin LIU

Department of Informatization Research, National Information Center, NIC, Beijing, China

Email: liuluyin@163.com

Abstract: The public knowledge platform has been widely established in China, but there are still many problems on its efficiency and effectiveness, it is because the public knowledge platform is still in the construction phase, not yet developed to the stage of value realization. Based on the research of the value judgments and value reference of public knowledge platform, this paper brings out the corresponding policy recommendations.

Keywords: public knowledge platform; information value; digital fair

1 Preface

The accumulation and communication of knowledge is an important driving force for human development. In the Internet age, because of its openness, interactiveness, collaborativeness, the network has become the main carrier of knowledge. Particularly, the dissemination of public knowledge by means of networks is becoming an important indicator of the information society.

With the broad construction of public knowledge platform (hereinafter referred to as the platform), the platform's running quality, efficiency and the degree of value realization has received attentions. From the point of the platform establishment practice, the current concern of information and knowledge value has mainly focused on the microcosmic information value, and has strengthened the quality of the collected information. But the value of the entire public knowledge platform as well as the knowledge network linked by a number of platforms has not been given due attention, which has greatly influenced the spreading scope and efficiency of the platform, thereby reduced the overall value of the platform.

This problem has begun to arouse the public concern. When people are researching the problems, such as information benefits, resource allocation, digital divide, and information accessibility etc, they are urgently aware that the value realization of public knowledge platform can greatly promote the digital equity and social development.

On the perspective of macro-management and development, this paper analyzed the value attributes of the platform, further, the overall effect which the platform truly need to achieve to promote digital equity and social development.

2 Concept of Public Knowledge Platform and Its Value Judgment

The knowledge system of human beings has been built up layer by layer through thousands of years. From the experiential knowledge to theoretical knowledge, and further to philosophy, it gradually has formed a relatively knowledge. In the process of completing human knowledge, due to the limitations of printing tools and communication costs, knowledge sharing, in the very long period, is a activity of the minority, with quite limited audience. However, with the development of information technology and the Internet, - the dissemination cost of knowledge has become low, and knowledge sharing has become increasingly open. Utilizing technology to gather knowledge together in a unified interface to retrieve and use by people has become an important means of disseminating knowledge. In practice, the public library network, the platform for sharing of technological resources, and the system of sharing and supporting of scientific documents all provide users with a variety of knowledge sets, and can be called the public knowledge platform.

Briefly speaking, public knowledge platform is the knowledge s integrated by the use of information technology, each of the knowledge collection provide all kinds of (user groups, individuals) with the retrieval, utilization and analysis of information and knowledge resources, and several knowledge collections can form a knowledge network give service to certain service objects. A number of knowledge collections aggregate knowledge system where people can use under the network environment.

The value judgments of public knowledge platform are usually the same with the value judgments in common sense. That is, the value of an object is determined by whether it meets certain needs of s and society or not; the size of its value is determined by the degree it meets the individual and social needs and the importance of those needs themselves.

Public knowledge platform is featured by system openness, content availability, and spread extensiveness, its values mainly reflect on providing a relatively perfect knowledge environment for people, enhancing the innovation level of all types of users, realizing digital equity, and ultimately promote the overall development and pro-

978-1-935068-76-1

gress of society. Public knowledge platform has become the foundation of knowledge innovation, dissemination and utilization and an important vector of knowledge in this modern society.

3 Value Orientation of Public Knowledge Platform

Based on the above described characteristics of public knowledge platform, its value orientation will expand successively. The basic value of public knowledge is providing people with a comparatively comprehensive knowledge environment, satisfying the information needs of the main types of innovation, and realizing digital fair in the whole range of society, thus contributing to social development. These four aspects of public knowledge platform's value orientation have phrasal distinctions. However, if the community has established a comprehensive system of public knowledge platform, these four aspects could exist at the same time.

3.1 Knowledge Environment

This is the most basic value orientation of public knowledge platform--building a knowledge requiring environment for users through the construction of knowledge platform.

At present, every country in the world has paid great attention to the construction of the knowledge environment. Such as, in 2003, the National Science Foundation of USA launched the "scientific and technological research on network information infrastructure reform." This research suggests that, with the development of computing technology, information technology and communications technology, scientific research will enter into a new era. The basis of scientific research is a series of activities based on "network information infrastructure" (Cyber infrastructure), which created various forms of scientific knowledge environment, and changed the traditional methods of scientific research. Knowledge environment is formed by the digital library, large scientific instruments, scientific data, and tools, etc. It is an integrated knowledge environment gathered by scientific and technological resources. Under this environment, the efficiency of technological innovation and technological innovation can be improved. Cyber infrastructure is not a linear expansion of the scale and the level of national infrastructure investment, but represents a new direction of national knowledge infrastructure which a technologically advanced country is building in the 21st century. The so-called "advanced computing network infrastructure" is an integrated infrastructure including six elements of hardware, software, information, services, personnel and institutions, relying on the basic information technology, such as computing, storage and communications, etc., and, by means of network, operating systems and middleware, provides the following services: high-performance computing; data, information, and knowledge management; observation, measurement and manufacturing; interfaces, visualization; collaboration. These services will achieve customization according to the application of different disciplines and the projects. The ultimate goal is to build a knowledge environment catering to the different applications of scientific research, engineering, education, industry and so on.

3.2 Source of Innovation

Public knowledge platform has a profound impact on ways to acquire knowledge and information by innovative subjects.

As for Scientific research institutions and researchers, scientific communication approaches have experienced a revolution. Traditional communication approach of science can be classified to two kinds: formal and informal. Formal communication refers to the communication among scientific research institutions or personnel through the exchange of scientific and technological journals. Informal communication refers to the dialogue, letters, seminars and other direct communication. In cyberspace conditions, the boundary of formal and informal scientific communication has become blurred. Innovation diffusion can take advantage of electronic journals and way of talking to the network and the cyberspace has accelerated the diffusion of innovative knowledge among scientific and technical institutions and personnel.

For enterprises, cyberspace can promote the formation of innovation networks. Technical innovation is not an independent act of business or individual, technological innovators are among the relationship of both competition and cooperation, and this mode of competition and cooperation has intensified by cyberspace. With increased competition, accelerated technological innovation speed and rapid market changes, innovation activities of enterprises have entered network innovation phase with multi-party cooperation and interactive wound, and by the use of cyberspace connection, innovation networks of enterprises play an active role in technological innovation in the practice. For example, the Silicon Valley, after years of construction and development, has formed innovation networks in which innovative subjects promote each other. Innovation network is the flow of information between different actors in innovative process, with the aim of technical innovation. Cyberspace has created an interactive environment for knowledge and information exchanges in innovation network and promoted the realization of technological innovation. Through innovation networks, enterprises can integrate and make flexible use of inventions of internal and external and innovative services, and, making products, services and business models achieve the best profit.

For the public, cyberspace provides a space for the accumulation of innovative ideas. Internet as a public and neutral platform changes ways of ideas exchange. Voluntary exchange of personal information in cyberspace is

very convenient, and cyberspace provides for the public an environment for dissemination and exchange of innovative ideas and learning.

3.3 Realization of Digital Equity

Dissemination of traditional knowledge is of a small minority, only in a small scale to achieve the spread and use of knowledge, while public knowledge platform can achieve a wide range of knowledge dissemination.

From a global perspective, the realization of digital equity has achieved the international community's attention. Auspices of the United Nations have aroused great emphasis in countries on bridging the digital divide and information poverty issues. The United Nations in the "Biwako Millennium Framework for Action" initiated to create an inclusive, barrier-free and rights-based society, with the idea that in the information society, retrieving and using information is a basic human right, to ensure that under different conditions information can be accessible for anyone.

Many foreign public knowledge platforms, in the building taking into account of the information needs of the users of different ages, using different languages and people with disabilities, make sure that different groups can enjoy a wealth of knowledge, to achieve a digital equity between different groups of users. For example, the U.S. Library of Congress supplies specialized services for all types of users, including children and families, librarians, publishers, researchers, teachers, etc., to make sure that each category of users can get personalized service.

3.4 Promoting the Comprehensive Development of Society

Public knowledge platform provides a solid foundation for social development. One reason is that public knowledge platform can provide users with structured knowledge, which is the crystallization of human civilization, and plays a fundamental role in knowledge production and innovation. On the other hand, the individual through the use of public knowledge platform can initiatively combine individual knowledge with public knowledge to enhance the individual value, thus contributing to overall social development and progress.

4 Policy Recommendation for Construction of a Sound Public Knowledge Platform in China

China has entered the ranks of middle-income countries, when the primary problem is to solve the problem of sustainable development, the key of which is the popularity of various levels of knowledge in society and the ability of improving technological innovation. Innovation capacity needs a lot of valuable knowledge of innovation to support, and the state should formulate a series of policies to ensure the smooth implementation of public knowledge platform.

4.1 Create a Harmonious Policy and Regulatory Environment in Favor of Diffusion of Public Knowledge

It is suggested that the state authorities issued a series of policies and regulations conducive to the spread of public knowledge to eliminate and shield the constraints which affect knowledge dissemination. These suggestions are: to recommend cooperation of relevant departments to discuss and check policies and regulations impeding the spread of public knowledge, to adopt measures to improve policy and regulatory environment for the dissemination of public knowledge; to establish a policies and regulations assessment mechanism as to the dissemination of public knowledge.

4.2 Promoting Formation of Public Knowledge Platform Though Information Sharing

In cyberspace conditions, production, dissemination and use of information and knowledge develop on an unprecedented scale, which becomes the powerful driving force for technological innovation, human progress and coordinated development of economy and society. Improving the using level of information resources can provide based security for innovation and diffusion of technology. China has rich scientific and technical information resources, but the construction of science and technology information resources by fragmentation and administrative division seriously affects the overall effect of scientific and technical information resources serving innovation

Improving sharing and utilization of information resources, can effectively promote the diffusion and circulation of innovation knowledge in society, and then, information as a link, promote the formation of innovation networks.

Open access to content: to create preconditions in the organization, law and technology, standardization and interoperability, and in accordance with relevant guidelines and standards to achieve universal accessibility of public websites; among consumers to create awareness about the risks of technological progress, and to take necessary measures for issues such as digital protection, privacy, consumer rights and protection of vulnerable groups in society.

4.3 Develop Strategies to Promote Digital Equity

Public knowledge platform aims to establish an open, fair and reasonable network knowledge system to ensure all people, including rich and poor, men and women, the elderly and young people, people of different regions can share the results of public knowledge platform. States should develop a basic strategy promoting digital equity, identify short-term and long-term goals, coordinate all

978-1-935068-76-1 © 2011 SciRes.

areas and systems' efforts to eliminate the digital divide, and then progressively realize digital equity and promote overall social development.

References

[1] Wang Fang. "Social information welfare and its Realizations". Journal of Library and Information Service,2009,53(18),pp.30-34.
[2] Zhang Yanning,Song Zhengfeng. "Present Condition of China's S&T Information Field and Analysis of Development Countermeasures", Journal of the China Society for scientific and technical Information,2007,26(5),pp.790-795.
[3] Liu Qiang, Zeng Minzu."The Construction of Sustainable S&T Information Service in China". Proceedings of the 45th Anniversary of S&T Information Conference,pp.25-32.
[4] Liu Luyin,Guan Jialin. "Study on the Service Management of S&T Resources". China Science & Technology Resources Review, 2008, 40(1), pp.16-19.

Scientific Research

eHealth Program for Connecting Communities of Care

Neil BERGMANN[1], Ying SU[2]

[1]*School of Info. Tech. and Elec. Eng., the University of Queensland, Brisbane 4072 Australia*

[2]*Information Quality Lab, RSPC, Institute of Scientific and Technical Information of China, Beijing, China*

Email: n.bergmann@itee.uq.edu.au, suy.rspc@istic.ac.cn

Abstract: The aim of this pilot study was to investigate, specify, design, prototype, evaluate and analyze new, state-of-the-art software methodologies and architectures specifically tailored to eHealth applications, which can be used by the Industry Partner to develop new, successful eHealth products that will substantially improve the delivery of care to people with Chronic Disease. These new methodologies and architectures will be based on Participatory Design Principles, and Model Driven Architectures, and 6-month eHealth program was to improve elders' autonomous access to and use of health-related information in the form of videoconference from a human-centered web site. The content of the program included participants' mastery of basic computing skills and accessing and enhancing participants' interest in seeking health related knowledge and information via the Internet. Data were collected in months 2 (pretest) and 4 (posttest) using questionnaires and open-ended questions. The overall learning experience was positive. Our eHealth program can be viewed as a prototype for designing technology platforms for the delivery of professional healthcare services to home-based older adults with chronic disease.

Keywords: eHealth; Chronic Disease; Participatory Design Principles; Model Driven Architectures

1 Introduction

Modern Information and Communications Technologies (ICT) have demonstrated enormous benefits in supporting information and workflows in a wide variety of industries. Significant investment is currently occurring worldwide in Health Informatics Systems (HIS) which seeks to capture these productivity benefits in the healthcare industry. Unfortunately, many HIS deployments have been much less successful than deployments in other industries. This is because of poor stakeholder engagement, caused in part by poor product design and implementation.

The aim of this paper is to investigate, specify, design, prototype, evaluate and analyze new, state-of-the-art software methodologies and architectures specifically tailored to eHealth applications, which can be used by the Industry Partner to develop new, successful eHealth products in the area of Chronic Disease management and Community Care.

These new methodologies and architectures will be based on Participatory Design Principles, and Model Driven Architectures, and 6-month eHealth program was to improve elders' autonomous access to and use of health-related information in the form of videoconference from a human-centered web site.

The remainder of this paper will summarize the current state of the art, describes the aims, goals and objectives of this project in more detail, and then provide illustrative case study to demonstrate health promotion among community-dwelling older adults.

2 State of the Art

The increasing international pressure to accelerate HIS adoption will deal with some of these issues, such as promulgation of standards, funding of national eHealth infrastructure, and national level HIS adoption. However, there is also still a need to address the significant issues associated with poor quality and inappropriate eHealth software. This project particularly deals with issues of software quality, usefulness, evaluation, stakeholder buy-in and interoperability, by investigating the use of some powerful new design techniques which have shown promise in overcoming software quality.

2.1 Participatory Design

Participatory Design (PD) is a technique which has been used successfully for more than 20 years in the design of technology[1]. The key principle of PD is to actively involve the real end users of the system in the design, implementation and evaluation of the systems that they will use. CSPR [2] describe the following principles of PD:

- Respect the users of technology, regardless of their status in the workplace, technical know-how, or access to their organization's purse strings. View every participant in a PD project as an expert in what they do, as a stakeholder whose voice needs to be heard.
- Recognize that workers are a prime source of innovation, that design ideas arise in collaboration with participants from diverse backgrounds, and that technology is but one option in addressing emergent problems.
- View a "system" as more than a collection of software encased in hardware boxes. In PD, we

978-1-935068-76-1

see systems as networks of people, practices, and technology embedded in particular organizational contexts.

- Understand the organization and the relevant work on its own terms, in its own settings. This is why PD practitioners prefer to spend time with users in their workplaces rather than "test" them in laboratories.
- Address problems that exist and arise in the workplace, articulated by or in collaboration with the affected parties, rather than attributed from the outside.
- Find concrete ways to improve the working lives of co-participants by, for example, reducing the tedium associated with work tasks; co-designing new opportunities for exercising creativity; increasing worker control over work content, measurement and reporting; and helping workers communicate and organize across hierarchical lines within the organization and with peers elsewhere.
- . Be conscious of one's own role in PD processes; try to be a "reflective practitioner."

PD has been widely used in many areas, although its use is not without criticism[3]. PD can be a time consuming and iterative process, and its benefits are sometimes hard to quantify. Its use has often been restricted to early stages of design, with its use in later, more technical aspects of the ICT design process harder to manage. There have been only limited reported uses of Participatory Design in the HIS field, but some of these have shown useful benefits[4,5]. Pilemalm and Timpker present what they call a third generation PD approach for HIS development, and a key aim of this research is to evaluate the practicality and usefulness of this current state-of-the-art PD approach, and to further enhance the approach to fit within the current Australian National eHealth Strategy[3].

2.2 Model Driven Design (MDD)

This project also intends to undertake research in the area of Model Driven Design. This is a technical ICT design approach which divides software design into two distinct phases. Firstly, the software functionality is modeled in a modelling language such as Unified Modelling Language (UML). UML provides a method for visualizing and modelling software architecture, including the actors in the system, the business processes, the software processes, database transactions and messaging. Importantly, a UML model of a system is independent of the underlying implementation technology. In the second stage of software design, the UML description is translated, either manually or automatically, into a particular implementation, using specific software components, database systems, messaging systems, and distributed systems middleware.

The Object management Group (OMG) has developed one particular MDD approach called the Model Driven Architecture (MDA), which is built upon other OMG standards such as UML and CORBA. One criticism of MDA is that UML is a very general modelling framework, and others have suggested that it may be more appropriate to work with domain specific modelling languages for HIS. However, as Walderhaug et al notes [6]: "Creating DSMLs for the healthcare domain is a daunting task, and requires extensive investment of resources and time".

One of the key research issues in this project is to evaluate the usefulness of model based design as a way of capturing the workflows and information flows involved in the example HIS applications chosen for this project, and then using those models to drive the development of the software implementations for this project. The Industry partner in this project, ltxysoft (http://www.ltxysoft.cn/), have their own HIS platform, called human-centered eHealth (**HeH)**, and part of this project will be to identify how best to use model-driven design to develop specific HIS implementations, from the HIS specifications developed within the Participatory Design process described above.

3 Aims, Goals and Objectives

This project undertakes research to enable the more efficient and effective development of HIS. For this project, one particular model of care will be used as the case study for new HIS software development. Importantly, it is outside the scope of this project to undertake a detailed clinical study of the effectiveness of a particular model of aged care, or chronic disease management in terms of improved health outcomes – although that is a future goal of the overall program of research by the research team. The evaluation in this project will be on the effectiveness of new software design techniques to support new models of care. It has been shown many times in the eHealth area that it is necessary to get the software support right first, before the health benefits of new ICT-enabled care models can be meaningfully evaluated.

The care model on which this software project is based is referred to as a Electronic communities of care. The target population is the elderly and those with chronic diseases who are living independently in the community, often with the help of friends and family. Those in the target audience are often dealing with several co-morbidities, and may have complex care needs. The Electronic communities of care model uses ICT to coordinate the activities of patients, their informal carers, primary health carers and allied health carers through access to appropriate patient centred health information.

This care model is based on the development and coordination of an individualized patient-care plan, which is developed amongst the whole community of carers, and the patient themselves. It is also based on active involvement of the patient in the monitoring and management of their conditions through access to suitable self-monitoring

systems, and high-quality, web-based information resources.

The research team includes experts in gerontology and aged-care nursing, and the care model has been developed in consultation with them. The Australian Government Medicare Chronic Disease Management initiative [7] supports integrated allied health and general medical practitioner care, however there have been limited reports of how well this initiative has been operationalised [8]. Additionally, Medicare funds referrals to geriatricians for a comprehensive assessment of an older person, covering areas such as current active medical problems; past medical history; medication review; immunisation status; advance care planning arrangements; current and previous physical function including personal, domestic and community activities of daily living; psychological function including cognition and mood; and social function including living arrangements, financial arrangements, community services, social support and carer issues.

3.1 Aim

The aim of this project is to specify, develop, implement and evaluate software support for this Chronic Disease Management initiative through a newly developed Health Information System. This HIS allows appropriate electronic access by health professionals, informal carers and patients to their individualized healthcare plans, and also to relevant web-based information sources to support patient-centred management of Chronic Disease. The system will be developed in cooperation with a focus group of users from all parts of the care community using the principles of Participatory Design explained earlier. The principles of Model Driven Design will be used to provide a means of describing the system's operation separate from the underlying implementation technology, which will be based on the Industry partner's exisiting software products.

3.2 Goal

The goal of this project is to develop new models for aged care. New models of healthcare are currently being actively investigated in many countries. Some key features of these new models of care based on high-quality eHealth technology [9] are:

- Engaging consumers: Patients should be fully engaged in their own health management, supported by information and tools that enable informed consumer action and decision making.
- Transforming Care Delivery: New patient-centred/clinician-guided workflows will be facilitated by making more complete, timely and relevant patient-focused data and clinical decision support tools available in a secure manner to patients, carers and clinicians at the point of care.
- Improve Population Health: eHealth systems enable much more detailed analysis on the health outcomes and cost effectiveness of new treatments and programs.
- *Transforming Funding Models*: New models of care delivery need to be supported by new models of funding which ensure that eHealth infrastructure is appropriately and equitably funded for patients, carers and clinicians.
- *Managing Privacy, Security and Confidentiality*: Consumers in a healthcare system supported by state-of-the art HIS must have confidence that their personal health information is private, secure and used with their consent in appropriate and beneficial ways.

3.3 Project Objectives

Given the above background, the project aims and objectives can now be more clearly stated:

1) The project aims to specify, prototype and evaluate new eHealth software to support Chronic Disease Management based on the concept of an Electronic communities of care, and building upon the existing software components available from the Industry partner.

2) The project aims to investigate the applicability of Participatory Design techniques for the specification, prototyping and evaluation of Electronic communities of care software, including the need for such software to meet the needs of a very diverse set of users – clinicians, allied health workers, informal carers and patients.

3) The projects aims to develop new modelling frameworks, based on modelling languages such as UML, which are suitable for specifying Electronic communities of care eHealth software. The project also aims to evaluate the applicability and usefulness of mapping models described in this framework into applications based on the Industry Partner's HeH software suite.

4) The project aims to investigate the integration of existing and newly-developed web-based information sources on patient-centred Chronic Disease management into the Electronic communities of care software suite, and to evaluate the perceived usefulness of such information sources for patient-centred care.

5) The projects aims to provide a framework for incorporating home-based and point-of-care based health monitoring devices (blood pressure, temperature, weight, blood glucose, activity monitors) into the Electronic communities of care health records.

6) Although a formal clinical trial of the medical effectiveness of an Electronic communities of care approach to patient-centred chronic disease management is outside the scope of this project, the project does aim to improve elders' autonomous access to and use of health-related information in the form of videoconference from a HeH Web site.

4 Methods

4.1 Sample

978-1-935068-76-1 © 2011 SciRes.

This was an illustrative case study. Forty older people attending community centers were invited to participate in an eHealth education program. The inclusion criteria were (a) being an older person who scored 24 or above on the Mini Mental Status Examination and (b) being able to speak, understand, and read Mandarin [10].

4.2 EHealth Program: Objectives, Implementation, and Evaluation

The 6-month eHealth program (2 week per month) was held in the activity room of an elderly center. The objective of the eHealth program was to improve elders' autonomous access to and use of health-related information in the form of videoconference from a HeH Web site. The content of the eHealth program included mastery of basic computing skills and accessing HeH Web sites that provide health information regarding exercise. Participants were invited to perform exercise as recommended on the HeH web sites. The contents and expected outcomes of the program are itemized in Table 1.

For the design of the Web sites, PD guidelines that specify design criteria for older adult users were used and yielded intuitive, user-friendly Web sites. Pre- and post-questionnaires about the eHealth program, a knowledge test, and focus group interviews were used. Participants were asked to complete a questionnaire in month 1 (pretest) and in month 4 (posttest). A 5-point Likert scale (1, not very much; 5, a great deal) was used.

Focus groups were conducted at the end of the eHealth program for participants and staff working in the elderly center. Data collection ceased when saturation was achieved. The interviews were audiotaped and transcribed. Related content was clustered by themes and further analyzed into categories for comparison and discussion.

Table 1. Contents of the eHealth Program

Month	Contents	Expected outcomes
1	To introduce basic knowledge of operating a computer and accessing the Internet To promote awareness of eHealth by using a computer	Participants will be able to: Switch on and off the computer Operate the computer using mouse, and keyboard Access Internet Gain awareness and interest in using a computer
2	To evaluate the challenges of using technology Participate in health education (via a Web site from department of health) regarding the importance of exercise for older persons To access the Internet and a health service To increase health knowledge	Participants will be able to: Access health information Web sites Answer a set of questions regarding the importance of exercise before and after browsing health information Web sites Get more correct answers after browsing the health information Web sites
3	To examine the usability of HeH Web site To use of the Web site layout, links, and content To reinforce skills in operating the computer and browsing Web sites regarding exercise for older people	Participants will be able to: Demonstrate the operation of a computer and accessing the Internet Open and browse a health-related Web site Apply the health knowledge acquired to perform an actual demonstration on suggested exercise for older people
4	To evaluate benefits of an Internet-based educational support intervention delivered to older chronically ill adults. To reinforce skills in operating the computer and browsing Web sites regarding exercise for older persons To collect feedback on the program	Participants will be able to: Operate a computer and access the Internet Evaluate benefits of an Internet-based educational support intervention delivered to older chronically ill adults.

5 Data Analysis and Results

5.1 Participant Data

Forty older people (14 males, 26 females) participated in the study (Table 2).

Table 2. Participant Characteritics of the Older Persons

Characteristics	Frequency	Percent	p value
Sex Male Female	 14 26	35 65	0.026
Age (mean ± SD) Male Female			65.32±2.331 66.49±4.211
Living circumstances Living alone Living with family	 24 16	 60 40	0.132
Educational background No formal education Primary level Secondary level	 12 18 10	 30 45 25	0.034
Availability of computer at home Available Unavailable	 10 30	 25 75	0.000

Mastery of computer skills, interest, and health-related knowledge gained via the eHealth program.

Table 3 shows a significant increase in mastery of computer skills and interest in accessing health information via the Internet after the 6-month eHealth program; knowledge of health information in the area of chronic disease also increased significantly ($p < 0.05$). Participants were able to operate the computer and access health information via the Internet, and they expressed confidence in their ability to do so. Participants were also found to have increased their knowledge after reading information on the Web sites, with a significant increase in giving correct answers to health-related questions.

Table 3. Mastery of Computer Skills, Interest, and Health-Related Knowledge Gained via eHealth Program (N=40)

	Pre (Month 2)	Post (Month 4)	p value[a]
Mastery of computer skills			
Basic computer operation skills	4	35	0.006
Log on to the Internet	11	37	0.000
Browsing and searching for information on the Internet	8	33	0.001
Confidence in accessing the Internet individually	3	36	0.002
Self-perceived ability to access health-related information via the Internet	4	31	0.000
Interest in accessing health information via the Internet			
Interest in using computers	4	34	0.000
Interest in the eHealth program	7	38	0.002
Interest in seeking health information via the Internet	3	28	0.000
Frequency of using the Internet to access health information	6	21	0.000
Knowledge of health information (chronic disease)			
Correct answer regarding chronic disease	4	23	0.002
Frequency of communicating with each other and health professionals	0	28	0.001
Feeling the necessity of preventing regular chronic disease	12	36	0.043
Usefulness of using the Internet to seek health information	2	32	0.002

Data are means (SD).

aWilcoxon Signed Rank Test was used.

**p < 0.05 was considered statistically significant.*

5.2 Focus Group Interview

Focus group interviews revealed that elderly center staff was impressed by the active involvement of the older persons in the eHealth program. Staff felt that providing such programs through elderly centers would be useful. The following comments, made by participants during the focus group interviews, highlight their reactions to the eHealth program and their concerns about using technology.

Explore the technology. "It provides an opportunity to explore technology that we had not imagined before." "I could obtain health information. There is a lot of information out there, and I was able to use it."

Language barrier and lower education level. "We don't know many words." "People think we are old fashioned, so there's no need to learn about computers." "We can't even type. How are we supposed to manage Chinese typing?"

Physical disabilities. "Poor eye–hand coordination makes it difficult for us to use the mouse." "When sitting for a period of time, I get severe low back pain. It annoys me."

Social support. "At least we have each other," and "These meetings are the highlight of my week."

Health issues. "I can't do as much but I can still get around the house, do some cooking and watch television" and "I always take my medications at the right time—things would get worse if I didn't."

6 Discussion and Conclusion

The proportion of the population aged 65 and over is predicted to increase from 12% to 25% by 2051 in Australia with comparable increases in the percentages of the elderly in the populations of most other countries. The numbers in this age group are predicted to grow very fast, from 2.2 million in 1997 to about 4.0 million in 2021 and about 6.3 million in 2051. China has a population of over 1.3 billion people, of whom 160 million are aged 60 and older; the largest aging population in the world. In addition, China's aging population is estimated to increase at a rate of 5.96 million per year from 2001 to 2020 and then 6.2 million per year from 2021 to 2050, and is expected to exceed 400 million by 2050, accounting for 30% of its total population. Ageing is related to health issues such as declining health status, and increased costs of care.

Evidence suggests that patients with effective self-management skills make better use of health care professionals' time and have enhanced self-care [11] and the Australian federal government has allocated an unprecedented $515 million for implementation of guided chronic disease self-management over 5 years. In 2009, China's State Council passed a long awaited medical reform plan which promised to spend 850 billion yuan (123 billion U.S. dollars) by 2011 to provide universal medical service to the country's 1.3 billion population [12].This proposed program will support that initiative by providing a standard framework for patient information across diseases.

The present study demonstrated an innovative way of using the Internet to search for health information and improve health promotion among community-dwelling older adults. With the collaboration of community services, the use of technology in the form of an eHealth program would be an effective tool in providing health education to older people. Computer skills, interest in accessing health information, and health knowledge gained were found among the 40 participants. It is suggested that attending the eHealth program helped to build their confidence in searching health-related Web sites and that they benefited from information obtained from these Web sites. Therefore, the eHealth program and the use of Internet technology will be a new trend in health promotion and health education among the older population.

As older populations and their vulnerability to health problems increase, eHealth education can help to enhance their health status. The use of eHealth education can help promote healthy habits and lifestyles and increase health knowledge among older people. Continual and regular provision of eHealth programs for older people will bring positive health outcomes to older adults and to society as a whole.

The analysis of the project results show that older

978-1-935068-76-1 © 2011 SciRes.

adults with a chronic disease can be trained to use computer hardware and software to access the Internet and negotiate links of a HeH Web site specifically designed for older adult users. The aim of the program was to deliver an in-home healthcare intervention to chronically ill older adults who manage their own healthcare, frequently without the support of family members or friends. The challenge was to demonstrate that older adult computer nonusers could be trained to use hardware and software to access a Web site that was designed to address the computer skills, interest, and health-related knowledge gained via the eHealth program. A second challenge was to show that the analysis of the online support group discussion themes would be clinically relevant in terms of the concerns shared by older adults living with a chronic disease. In terms of responses to the use of technology to access health services, the qualitative analysis of the follow-up interviews showed that older adults with limited prior experience using computers could be trained to manipulate hardware and software to access the Internet and negotiate the HeH Web site links. Some participants reported initial hesitation in learning to use the equipment. However, as familiarity with the HeH Web site increased, participants became more confident in using each of the links and were pleased with their ability to acquire a new and useful skill. At follow-up, what the participants valued the most about the technology platform was being able to communicate with others and being able to access a healthcare support program from the comfort of their homes.

Session analyses of the HeH online group model of intervention with older disabled adults showed that the intervention could be professionally delivered and that the extracted group discussion themes appear to address key issues of importance to persons with a chronic health condition. For example, the analysis showed that the disabled feel marginalized and in some ways lose a sense of positive self-identity. Although not measured, we can speculate that depression may be a comorbid condition with many chronic diseases. Despite the limitations of our online group videoconferencing, the HeH Web site videoconferencing format appears to support group cohesion similarly to what occurs in face-to-face groups. Analysis of the participants' perception of the usefulness of the online support group intervention follow-up interviews showed that participation in the online support groups reduced the participants' sense of isolation and loneliness, while encouraging maintenance of optimal healthcare strategies.

Acknowledgment

We would like to thank NNSFC (National Natural Science Foundation of China) for supporting Ying Su with a project (70772021, 70831003).

References

[1] C. T. Kello, "Participatory design: Principles and practices - Schuler,D, Namioka,A," *American Journal of Psychology,* vol. 109, pp. 630-635, Win 1996.

[2] Computer professionals for Social responsibility. (2005, *What Is Participatory Design?* Available: http://cpsr.org/issues/pd/introInfo/

[3] S. Pilemalm and T. Timpka, "Third generation participatory design in health informatics - Making user participation applicable to large-scale information system projects," *Journal of Biomedical Informatics,* vol. 41, pp. 327-339, Apr 2008.

[4] K. A. Thursky and M. Mahemoff, "User-centered design techniques for a computerised antibiotic decision support system in an intensive care unit," *International Journal of Medical Informatics,* vol. 76, pp. 760-768, Oct 2007.

[5] C. Weng, *et al.*, "Participatory design of a collaborative clinical trial protocol writing system," *International Journal of Medical Informatics,* vol. 76, pp. S245-S251, 2007.

[6] M. Chaudron, *et al.*, "Experiences from Model-Driven Development of Homecare Services: UML Profiles and Domain Models," in *Models in Software Engineering*. vol. 5421, ed: Springer Berlin / Heidelberg, 2009, pp. 199-212.

[7] A. G. D. o. H. a. Ageing. (2009, Chronic Disease Management (CDM) Medicare Items. Available: http://www.megpn.com.au/Docs/MBS%20items/Overview%2520Fact%2520Sheet.pdf

[8] K. Grimmer-Somers, *et al.*, "Enhanced Primary Care pilot program benefits Type II diabetes patients," *Australian Health Review,* vol. 34, pp. 18-24, 2010.

[9] e. Initiative. (2008, Improving Our Nation's Health and Healthcare Through Information Technology. Available: http://www.ehealthinitiative.org/sites/default/files/file/eHealthInitiativeConsensusPolicyExecutiveSummaryDec2008.pdf

[10] M. M. Y. Tse, *et al.*, "E-health for older people: The use of technology in health promotion," *Cyberpsychology & Behavior,* vol. 11, pp. 475-479, Aug 2008.

[11] J. E. Jordan and R. H. Osborne, "Chronic disease self-management education programs: challenges ahead" The Medical Journal of Australia, vol. 186, pp. 84-87, 2007.

[12] Y. Su, *et al.*, "Human-centered eHealth: Current opportunities, challenges and the way forward for China," in *The Global People-centred eHealth Innovation Forum*, BMJ Publishing Group, 2011, pp. 25-27.

Scientific Research

eQualityHealth Program for Managing Data & Information Quality in SOA-based eHealth Systems

Ying SU[1], Omar BOUCELMA[2]

[1]*Information Quality Lab, RSPC, Institute of Scientific and Technical Information of China, Beijing, China*

[2]*LSIS UMR CNRS 6168, Université Paul Cézanne Aix-Marseille 3 Domaine, Universitaire de Saint-Jérôme, Avenue Escadrille Normandie-Niemen, France*

Email: suy.rspc@istic.ac.cn, omar.boucelma@lsis.org

Abstract: Disasters are non-routine events in societies, regions, or communities that involve conjunctions of physical conditions with social definitions of human harm and social disruption. Disaster management is a multifaceted process aimed at minimizing the social and physical impact of these large-scale events. IT has become a critical tool for facilitating the communications and information-processing activities in managing disasters. The objective of this paper is to provide a service oriented architecture (SOA) for eHealth systems enhanced with data and information (DIQ) tools and quality-driven service composition techniques in to enable effective and accurate decision making in the context of disasters. We identify the most important issues regarding the quality evaluation and we proposed several solutions, following the SOA paradigm.

Keywords: eHealth; Disaster management; Service Oriented Architecture (SOA); Data Quality, Information Quality; Web Service

1 Introduction

Disasters are non-routine events in societies, regions, or communities that involve conjunctions of physical conditions with social definitions of human harm and social disruption. Natural, technological, and willful (terrorist initiated) sources of disasters all cause dramatic losses of life and property [1]. Disaster management is a multifaceted process aimed at minimizing the social and physical impact of these large-scale events. It is thus not surprising that IT has become a critical tool for facilitating the communications and information-processing activities in managing disasters [2]. Especially important is the role of Web Services [4] and the Service Oriented Architecture (SOA) [5] for computer to computer communication in helping to support this capability SOA is a conceptual architecture that specifies interoperable, modular, loosely-coupled, self-contained and self-describing applications, systems or services that interact only at well-defined interfaces [6]. Analogously, web services are loosely-coupled, interoperable and modular network addressable application modules which can be invoked across heterogeneous networks to access and process information [7].

Web services offer a framework for modular composition of evolvable information systems [8]. However, most of current web service implementations do not guarantee the levels of information quality delivered to their users [9].

The objective of this paper presents a managing platform for better disaster medicine by assessing data and information quality provided by an SOA-based eHealth System using the quality metrics of geospatial data and health information, which is a novel step towards the building eHealth of trust. Then DIQ assessing model for disaster management is given in detail. These will enable disaster medicine specialist and volunteer rescue workers to make more accurate, timely and all-around decisions before, during, and after a disaster situation.

The remainder of this paper will summarize the current state of the art, describes the aims, goals and objectives of this project in more details, and then provide architectural components of eHealth system.

2 State of the Art

2.1 Data Quality

The topic of information quality has increasingly generated many research works. Data quality is a crucial problem in companies and organizations, where investments for the IS must be justified and reevaluated. Comprehensive surveys on data quality can be found in [10,11]. Data quality is generally described through a large set of quality dimensions, attributes or factors [12]. Literature aims at defining quality factors and metrics, proposing quality models including these factors and metrics [14], enabling the quantitative evaluation of quality factors [15], proposing taxonomies of factors and metrics [13].

In [16], data stream quality is assessed based on data provenance, which is used to computes trust scores for the stream presents load shedding mechanisms which take into consideration data quality and presents quality assessment models for objects and fusion operators in the context of wireless sensor networks.

2.2 Quality Driven Web Service Chaining

A service chain is defined in ISO 19119 as “a sequence of services where, for each adjacent pair of services, oc-

978-1-935068-76-1

currence of the first action is necessary for the occurrence of the second action" [ISO/DIS 19119] From a user perspective, service chaining is the linking together of standardised data and processing services into a workflow to produce results that are not predefined by the service providers. For instance, a user could use some EHR (electronic health record) data delivered by a web service as inputs to a data quality assessment or a data mining service, and then redirect the outputs of such a processing service to a decision making service.

The Object management Group (OMG) has developed one particular MDD approach called the Model Driven Architecture (MDA), which is built upon other OMG standards such as UML and CORBA. One criticism of MDA is that UML is a very general modelling framework, and others have suggested that it may be more appropriate to work with domain specific modelling languages for HIS. However, as Walderhaug et al notes [1]: "Creating DSMLs for the healthcare domain is a daunting task, and requires extensive investment of resources and time".

WfMC, the Workflow Management Coalition has defined XPDL (XML Process Definition Language) [WFMC] for saving and storing process models reliably, while the OMG came up with BPMN (Business Process Model Notation) the most commonly used notation business process modelling. Today XPDL remains the only standard means for the serialization of models using BPMN. Service chaining may result in a workflow implemented in using BPEL (Business Process Execution Language) and executed through a BPEL engine, or directly submitted to an XPDL engine.

3 Objectives and Originality

In this project, we focus on quality issues in eHealth systems. We consider autonomous and dynamic environments, where data sources are distributed, heterogeneous and where the set of considered data sources may evolve significantly over time Our goal is firstly to propose data quality evaluation tools for eHealth systems, and secondly to take into account quality aspects for service chaining in eHealth systems.

The care model on which this software project is based is referred to as an Electronic communities of care. The target population is the elderly and those with chronic diseases who are living independently in the community, often with the help of friends and family. Those in the target audience are often dealing with several co-morbidities, and may have complex care needs. The Electronic communities of care model uses ICT to coordinate the activities of patients, their informal carers, primary health carers and allied health carers through access to appropriate patient centered health information.

This care model is based on the development and coordination of an individualized patient-care plan, which is developed amongst the whole community of carers, and the patient themselves. It is also based on active involvement of the patient in the monitoring and management of their conditions through access to suitable self-monitoring systems, and high-quality, web-based information resources.

The research team includes experts in gerontology and aged-care nursing, and the care model has been developed in consultation with them. The Australian Government Medicare Chronic Disease Management initiative [2] supports integrated allied health and general medical practitioner care, however there have been limited reports of how well this initiative has been operationalised [3]. Additionally, Medicare funds referrals to geriatricians for a comprehensive assessment of an older person, covering areas such as current active medical problems; past medical history; medication review; immunisation status; advance care planning arrangements; current and previous physical function including personal, domestic and community activities of daily living; psychological function including cognition and mood; and social function including living arrangements, financial arrangements, community services, social support and carer issues.

4 Project Conceptual Framework

In order to enable efficient data and information quality (DIQ) management, three key issues are raised: (i) evaluating DIQ, (ii) analyzing DIQ and (iii) improving DIQ. These three problems correspond to the main components of our quality framework represented in figure 1.

4.1 Quality Evaluation

Considering an existing information system, one problem we are interested in is to assess data quality for both stored data and data streams. We will identify the relevant quality dimensions and propose an evaluation framework. This framework will address the following issues:

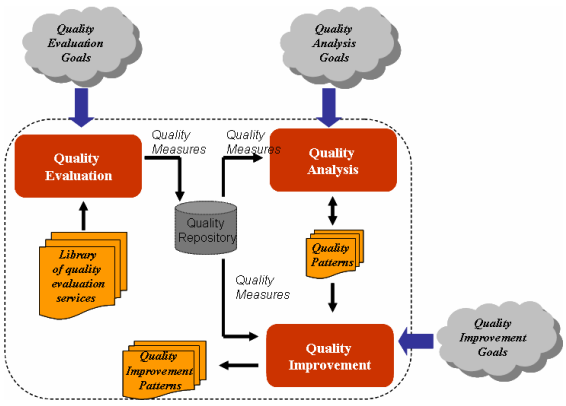


Figure 1. Quality Framework

1) Assessing the quality of stored data, possibly resulting from the integration of several data sources and therefore depending on the quality of these sources. The relevant quality dimensions and associated metrics will be identified.

2) Assessing the quality of streaming data. A first problem is to assess the quality of the data stream at the sensor level and to identify the relevant quality dimensions and metrics. A second problem is to propagate the quality properties through the processing of the data stream before it is presented to the user.

4.2 Quality Analysis

Beside the evaluation of the different quality dimensions, another important aspect is the analysis of the dependencies which may exist between the dimensions. Improving the value of a dimension may increase or decrease the value of another. One of our objectives is to provide the environment to enable the storage and analysis of the quality measures in order to exhibit the possible dependencies between the dimensions.

4.3 Quality Improvement

We are also interested in the improvement of data quality if the results of the evaluation show that the actual quality is far behind the requirements. We will adopt a pattern-based approach and we will propose a set of generic and reusable quality improvement patterns corresponding to different quality dimensions.

5 SOA eHealth

In Fig. 1, we proposed a general architecture consisting of four types of components: Legacy System component, Service Abstraction component, Semantic Service component and a Presentation component.

5.1 Administrative System Component

Registration, admissions, discharge, and transfer (RADT) data are key components of eHealth. These data include vital information for accurate patient identification and assessment, including, but not necessarily limited to, name, demographics, next of kin, employer information, chief complaint, patient disposition, etc. The registration portion of eHealth contains a unique patient identifier, usually consisting of a numeric or alphanumeric sequence that is unidentifiable outside the organization or institution in which it serves. RADT data allows an individual's health information to be aggregated for use in clinical analysis and research. SOA aggregates data and functionalities from three structurally independent and heterogeneous, real world sources:

Table 1. Web Services Protocol Stack

Metadata Exchage	Composition	BPEL, WS-Notification
	Quality of Service	SLA, WS-Security, WS-Reliable Messaging
UDDI	Description	WSDL, WS-Policy
	Messaging	XML, SOAP, WS-Addressing
	Transports	HTTP , HTTPS, SMTP
Network Level	Application	TCP/IP

1) GeoMan: An important feature will be the support offered by the GIS Web services, which will offer to authenticated patients the information about the known pharmacies, clinics, medical offices in their proximity.

2) DataPool: it comprises two types of databases. The one located at the Legacy system level is an operational database that records data sent by the sensors (they are named "source databases"). The other is a warehouse and is located at the central system level. All information is stored at this level.

3) InfoAgent is an Instant Messaging client facilitating lightweight communication, collaboration, and presence management built on top of the instant messaging protocol Jabber. The information can be accessed using an automated authentication and authorization system that allows a user that is authenticated and authorized on a Legacy System, to be automatically authenticated and authorized on the central system.

5.2 Service Abstraction Component

We distinguish between two classes of services: data and smart. The former refers to the three data sources introduced above, and are exposed by means of web service:

- Geo *service*: this service provides Geographic information over a specific rectangular spatial area.
- *Health services*: using the *DataPool* data each service in this set returns detailed information on a specific type of rest centre within a given circular area. For example, the 'getHospitals' Web service returns a list of relevant hospitals.
- Information services: these services allow presence information on online users to be accessed.

Smart services represent specific health planning reasoning and operations on the data provided by the data services. In particular, we created a number of DIQ assurance services that manipulate GIS data according to health specific requirements semantically described (e.g. health service centers with air conditioning system, hotels with at least 40 beds, easier accessible hospital, etc.). The criteria used were gained from our discussions with the health governors.

5.3 Semantic Service Component

The following ontologies reflecting the client and provider domains were developed to support Web Service Modeling Ontology (WSMO) descriptions:

- Domain Ontology: it represents the concepts used to describe the services attached to the data sources.
- HCI Ontology: it is composed of Human-Computer Interaction (HCI) and user oriented concepts, and allows lowering from the semantic level results for the particular interface which is used.

- Archetypes Ontology: it aims to provide a cognitively meaningful insight into the nature of a specialized object.
- Spatial Ontology: it describes GIS concepts of location, such as coordinates, points, polygonal areas, and fields. It also allows describing spatial objects as entities with a set of attributes, and a location.
- Context Ontology: it allows describing context *n–tuples* which represent a particular situation. In the emergency planning application, context *n-tuples* have up to four components, the use case, the user role, the location, and the type of object.

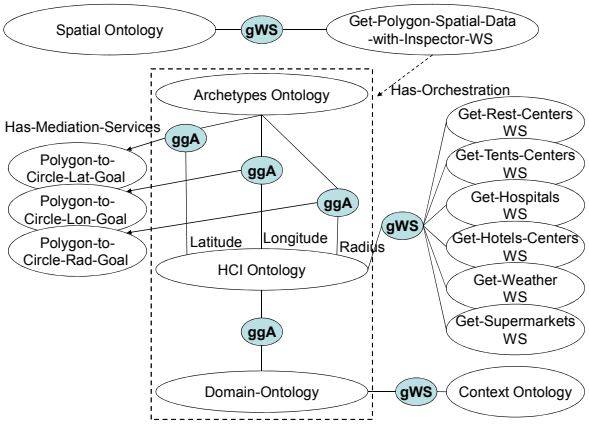


Figure 2. Ontology-driven Semantic Service Component

The central part of the framework is a taxonomy of DIQ dimensions. Once we identified the causes of DIQ variance, we were able to develop a taxonomy of DIQ dimensions that would allow us to evaluate the DIQ variance caused by these causes in a systematic and meaningful way. The original taxonomy of DIQ dimensions in the framework was based on semiotic theory. The four DIQ-categories or views are shown in Table 2:

Table 2. Taxonomy of DIQ Dimensions

DIQ Sorts	Data & Information Quality	
	Criterion	***Description***
Syntactic DIQ	Conformability	Is the data free of contradictions or convention breaks?
	Integrity	Is the scope of metadata adequate?
	Timeliness	Is the information processed and delivered rapidly without delays?
Semantic DIQ	Complete	Is the scope of information adequate?
	Concise	Is the information to the point, void of unnecessary elements?
	Accuracy	Is the information precise enough and close enough to reality?
	Currency	Is the information up-to-date and not obsolete?
Pragmatic DIQ	Applicability	Can the information be directly applied? Is it useful?
	Clarity	Is the information understandable or comprehensible to the target group?
	Value	Is the information add value to your work
	Interactivity	Can the information process be adapted by the information consumer?
Physical DIQ	Accessibility	Is there a continuous and unobstructed way to get to the information?
	Security	Is the information protected against loss or unauthorized access?
	Maintainability	Can all of the information be organized and updated on an on-going basis?
	Speed	Can the infrastructure match the user's working pace?

5.4 Presentation Component

The application user interface is based on Web standards. XHTML and CSS are used for presentation and JavaScript is used to handle user interaction together with AJAX techniques to communicate with IQA.

A small portion of eHealth management workflow represented in terms of WSMO descriptions is shown in Figure 2.

Get-Health-Spatial-Data-with-Inspector-Goal represents a request for available images within a disaster area. The user specifies the requirements as a target area, a sequence of at least three points (a polygon), and a shelter type (e.g. hospitals, tents). As mentioned above the set of Health services each return potential images of a specific type with a circular query area. The obtained results need to be inspected in order to return only images correlated to emergency-specific requirements (for example an earthquake). From a WSMO point of view the problems to be solved by this particular portion of the Ontology-driven Semantic Service layer included:

- Discovering the appropriate Health service; each definition is linked to the Get-Spatial-Data-Goal by means of a unique Get-Web-Service Agent (shown as gWS).
- Meditating the difference in area representations (polygon vs. circular) between the goal and Web services;
- Composing the retrieve and inspect data operations. Below we outline how the WSMO representations in Figure 2 address these problems.

Get-Health-Spatial-Data-with-Inspector-Goal is associated with a unique Web service that orchestrates, by simply invoking three sub-goals in sequence. The first gets the list of polygon points from the input; the second is Get-Spatial-Data-Quality-Goal; finally, the third invokes the smart service that inspects the list of spatial data. The first two sub-goals are linked by means of three get-goal Agent (depicted as ggA) that return the centre, as a latitude and longitude, and radius of the smallest circle which circumscribes the given polygon.

SOA offers a number of promises such as cost-

efficiency, reusability, flexibility, agility, and adaptability. However, a significant barrier to the widespread adoption of SOA-based systems for disaster management is how to cope with the problem of data and information quality.

Based on the ontology definitions, we define the following quality metrics for a context tuple T. $|T|$, $|T_A|$, $|T_I|$ $|T_M|$, and $|T_C|$ denote the cardinalities of the sets T, T_A, T_I, T_M, and T_C, respectively.

Accuracy of T, measured as $\alpha_C = |T_A|/|T|$, is the probability that a tuple in T accurately represents an entity in the real world.

Inaccuracy of T, measured as $\beta_C = |T_I|/|T|$, is the probability that a tuple in L is inaccurate.

Mismembership of T, measured as $\gamma_C = |T_M|/|T|$, is the probability that a tuple in T is a mismember.

Incompleteness of T, measured as $\chi_C = |T_C|/(|T|-|T_M|+|T_C|)$, is the probability that an information resource in the real world is not captured in T. Because T_A, T_I, and T_M constitute T, we have $0 \le \alpha_C, \beta_C, \gamma_C \le 1$ and $\alpha_C + \beta_C + \gamma_C = 1$.

Using these identities in the definitions of a_R, β_R, γ_R and χ_R, it is easily seen that $\alpha_R = \alpha_C$, $\beta_R = \beta_C$, $\gamma_R = \gamma_C$ and $\chi_R = \chi_C$. We show the algebra for χ_R here: $\chi_C = \dfrac{|T_C|}{(|T|-|T_M|+|T_C|)}$

The algebra of metric projection is defined as follows: $\Pi_S C_1 = C_O$ The probability that a tuple is accurate in R is therefore given by

$$\alpha_R = (\alpha_C + \beta_C)^{(k_{p\to K}+k_{p\to Q})/k_S} \gamma^{q_{p\to K}+q_{p\to Q}}$$

$$\beta_R = \left[(\alpha_C + \beta_C)^{k_{p\to K}/k_S} \gamma^{q_{p\to K}}\right]\left[1-(\alpha_C + \beta_C)^{k_{p\to Q}/k_S} \gamma^{q_{p\to Q}}\right]$$

$$\mu_R = 1-(\alpha_C + \beta_C)^{k_{p\to K}/k_S} \gamma^{q_{p\to K}}$$

$$\chi_R = 1-\frac{1-\chi_C}{1-\mu_C}\left[(\alpha_C + \beta_C)^{k_{p\to K}/k_S} \gamma^{q_{p\to K}}\right]$$

The algebra of the Cubic Product operator is defined as follows: $T_1 \otimes T_2 = T_O$

Metrics, we have

$$|\alpha_R| = \frac{|T_{1A}|}{|T_1|}\frac{|T_{2A}|}{|T_2|} = \alpha_1 \alpha_2$$

$$|\beta_R| = \alpha_1 \beta_2 + \alpha_1 \beta_1 + \beta_1 \beta_2$$

$$|\gamma_R| == \gamma_1 + \gamma_2 - \gamma_1 \gamma_2 .$$

The prototype was developed using ESRI ArcSDE software driven by a multiple users interface developed in Java.

Not all users evaluate quality based on the same criteria. Based on the ISO 19113 standard, quality indicators were defined and stored hierarchically within a relational database. The user can eventually adapt some items further or add more metadata about the indicators. One may select among different graphical representations to illustrate each indicator (See in Figure 3 and Table 3.)

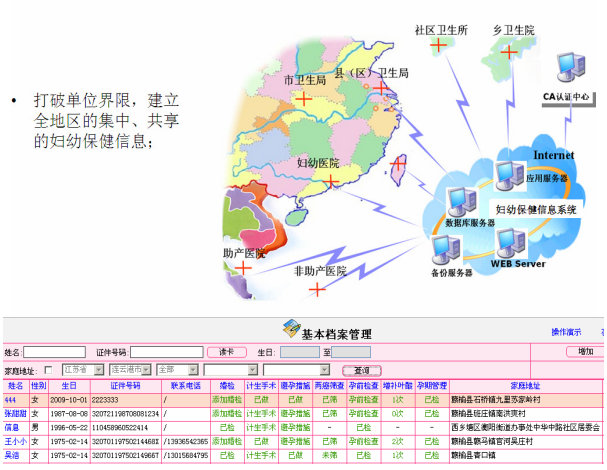


Figure 3. SOA-based eHealth and Interface

Table 3. DIQ metrics of Assessed Electronic Health Records

Dimensions	Nursing	Lab	Clinical	Radiology
accuracy	0.63	0.43	0.33	0.65
		0.49	0.61	0.34
Resolution	0.74	0.68	0.66	0.78
			0.78	0.73
		0.82	0.45	0.74
completeness	0.72	0.62	0.61	0.76
		0.84	0.63	0.53
consistency	0.81	0.86	0.68	0.72
		0.64	0.75	0.86

6 Discussion and Conclusion

The paper provided an analysis of use of the SOA paradigm in the context of an e-health Web-based system. We proposed a Conceptual Framework and several important components regarding certain aspects such as the Administrative System, Service Abstraction, the Semantic Service, and a Presentation component.

Our approach was focused on the autonomous and dynamic environments, where data sources are distributed, heterogeneous and where the set of considered data sources may evolve significantly over time Our goal is firstly to propose data quality evaluation tools for eHealth systems, and secondly to take into account quality aspects for service chaining in eHealth systems.

Unlike existing approaches for data quality assessment, we are interested in quality evaluation in a disaster context, and our aim is to propose algorithms and techniques

978-1-935068-76-1 © 2011 SciRes.

for quality evaluation in this context, which is one of the originalities of this project. This raises the following questions:

1) What are the specific constraints imposed by the disaster situation for the execution of quality evaluation tools?

2) How to capture these constraints and how to model the execution context?

3) What are the limits of existing quality evaluation techniques in such a context and how to overcome them to enable effective data quality in emergency situations?

Another objective of this project is to target any kind of data that can be considered in an eHealth system, be it structured/stored or stream/sensor data. While data quality issues have been widely studied in the literature for conventional databases, evaluating the quality of streaming data is still an open problem and very few approaches have been proposed to address this issue.

Our aim in this project is also to provide the stakeholders with a SOA eHealth platform where web services can be deployed and reconfigured independently. Unlike most of the current approaches, our goal is to guarantee the levels of information quality delivered, and to perform service chaining taking into account the quality of the delivered data.

Acknowledgment

We would like to thank NNSFC (National Natural Science Foundation of China) for supporting Ying Su with a project (70772021, 70831003).

References

[1] E. L. Quarantelli, "Where We Have Been and Where We Might Go," in What Is A Disaster?New Answers to Old Questions E. L. Quarantelli, Ed., ed London: Routledge, 2005, pp. 234-273.

[2] R. R. Rao, et al., Improving Disaster Management: The Role of IT in Mitigation, Preparedness, Response, and Recovery. Washington, DC: National Academies Press, 2007.

[3] Y. Sam and O. Boucelma, "Customizable Web services description, discovery and composition: An attribute based formalism," Silicon Valley, CA, United states, 2007, pp. 143-146.

[4] Rath, B., Donato, J., Duggan, A., Perrin, K., Bronfin, D.R., Ratard, R., et al. (2007). Adverse health outcomes after Hurricane Katrina among children and adolescents with chronic conditions. Journal of Health Care for the Poor and Underserved, 18(2), 405-417.

[5] Azuma, T., Seki, N., Tanabe, N., Saito, R., Honda, A., Ogawa, Y., et al. (2010). Prolonged effects of participation in disaster relief operations after the Mid-Niigata earthquake on increased cardiovascular risk among local.

[6] Y. Shen, et al., "Multi-index cooperative mixed strategy for service selection problem in service-oriented architecture," Journal of Computers, vol. 3, pp. 69-76, 2008.

[7] C. Zhou, et al., "QoS-Aware and Federated Enhancement for UDDI," International Journal of Web Services Research, vol. 1, pp. 58-85, Apr-Jun 2004.

[8] Y. Sam, et al., "Dynamic web services personalization," San Francisco, CA, United states, 2006, p. IEEE Computer Society Technical Committee on Electronic Commerce.

[9] Y. Su, et al., "Assuring Quality of Geo-information for Disaster Management," in 13th International Conference on Information Quality, MIT, USA, 2008, pp. 341-355.

[10] Batini C. and Scannapieco, M. 2006. Data Quality: Concepts, Methodologies, Techniques, Springer Verlag.

[11] Peralta, V. 2006. Data Quality Evaluation in Data Integration Systems. PhD thesis, Université de Versailles, France and Universidad de la República, Uruguay. http://tel.archivesouvertes.fr/docs/00/32/51/39/PDF/these.pdf.

[12] Rath, B., Donato, J., Duggan, A., Perrin, K., Bronfin, D.R., Ratard, R., et al. (2007). Adverse health outcomes after Hurricane Katrina among children and adolescents with chronic conditions. Journal of Health Care for the Poor and Underserved, 18(2), 405-417.

[13] Wang, R.Y., Storey, V.C. and Firth, C.P. 1995. A Framework for Analysis of Data Quality Research. IEEE Trans. Knowl. Data Eng., 7(4), 623-640.

[14] Lin, S., Gomez, M.I., Gensburg, L., Liu, W., & Hwang, S.A. (2010). Respiratory and Cardiovascular Hospitalizations After the World Trade Center Disaster. Archives of Environmental & Occupational Health, 65(1), 12-20.

[15] Pipino, L.L., Lee, Y.W. and Wang, R. 2002. Data Quality Assessment. Communications of the ACM, 45(4), 2002.

[16] Chronaki, C.E., Kontoyiannis, V., Charalambous, E., Vrouchos, G., Mamantopoulos, A., & Vourvahakis, D. (2008). Satellite-Enabled eHealth Applications in Disaster Management-Experience from a Readiness Exercise. Computers in Cardiology 2008, Vols 1 and 2, 1005-1008.

Session 2

Knowledge Organization and Knowledge Discovery

分论坛 2

知识组织与知识发现

Multilingual Concept Taxonomy Generation Based on Multilingual Terminology Clustering

Chengzhi ZHANG[1,2]

[1]*Dept. of information management, Nanjing University of Science & Technology, Nanjing, 210094*

[2]*Institute of Scientific & Technical Information of China, Beijing, 100038*

Email: zhangchz@istic.ac.cn

Abstract: The paper presents two methods of bilingual concept taxonomy generation, namely, cross language terminology clustering and mixed language-based terminology clustering. According to the experimental result of four special domains, it shows that: If the parallel corpus is used to cluster multilingual terminology, the method of mixed language-based terminology clustering is better than the method of cross language terminology clustering.

Keywords: multilingual concept taxonomy; multilingual terminology clustering; cross language terminology clustering; parallel corpus; mixed language

基于多语术语聚类的多语概念层次体系生成研究

章成志 [1,2]

[1] 南京理工大学信息管理系，南京，中国，210094

[2] 中国科学技术信息研究所，北京，100038

Email: zhangchz@istic.ac.cn

摘　要：本文从跨语言聚类和独立于语言（即混合语言）的聚类两个角度对多语言概念层次聚类问题进行对比研究。通过四个领域类别的数据进行测试，结果表明：仅依据平行语料进行多语言术语聚类，基于混合语言的多语言术语聚类策略优于跨语言的术语聚类策略。

关键词：多语言概念层次体系；多语言术语聚类；跨语言术语聚类；平行语料；混合语言

1 引言

层次体系（Taxonomy）构建是一项需要资源支持的、自顶向下、并且耗时的工作[1]。层次体系构建的方法包括手工构建、现有层次体系的调整、利用层次体系引擎（自动构建层次体系）以及混合方法等四种方法[1]。自动构建层次体系的方法包括：基于聚类的方法、基于词汇-句法模型匹配的方法（Lexicon-syntactic patterns）、利用机读词典（MRD）搜索词汇关系的方法[1]以及基于启发式规则的方法。基于聚类的层次体系构建方法，主要依据术语的分布特性，即同类术语在一定程度上具有类似的语境[1-6]。基于词汇-句法模型匹配的方法主要依据文本中出现的一些模式关系，如 Hearst 在 1992 年提出的几种上下位关系获取模式[1][7]。基于机读词典获取层次关系的方法，主要是利用 WordNet 等词典获取初始的层次关系，然后再通过用户反馈等手段减少概念间的关系[1][8]。基于启发式规则的方法，如 Sanderson & Croft 根据两个词语（分别为 A，B）在文本集中的出现频次信息，若 P(A|B)≥0.8，且 P(B|A)<1，则认为 A 是 B 的上位词[9]。

相对于单一语言的层次体系自动生成研究而言，关于多语言层次体系自动生成方面的研究较少。Christopher C. Yang 等人在中英文平行语料基础上，利用 Hopfield 网络和关联受限网络（associate constraint network）等模型进行了跨语言同义词表构建的研究[10,11]。Hans Hjelm & Paul Buitelaar 在可比语料基础上，利用层次聚类方法生成层次体系，他们的研究结果表明基于可比语料生成的层次体系，比基于单语语料生成的层次体系要准确[12]。Keita Tsuji & Kyo Kageura 在平行语料基础上，利用图论方法对翻译等价对进行聚类，从而生成日英同义词表[13]。与本文最接近的工作是，从学术论文数据库中获取日英关键词语料，通过图论方法生成双语关键词类簇[14]。

与以上工作不同的是，本文研究的是多语言概念

978-1-935068-76-1

层次体系的生成，聚类的对象是多种语言的核心术语集合，聚类的目标是生成多语言概念层次体系。此外，本文主要关注多语言聚类的策略，从跨语言聚类和独立于语言（即混合语言）的聚类两个角度对多语言概念层次聚类问题进行相对全面的研究。本文还研究了多种术语相似度计算结果与跨语言类簇合并阈值对跨语言术语聚类结果的影响。

2 基于多语术语聚类的概念层次体系生成

本研究利用 AP 聚类算法进行术语的聚类，其中的关键问题是术语相似度计算与聚类算法选择。图 1 为概念层次体系生成流程图。

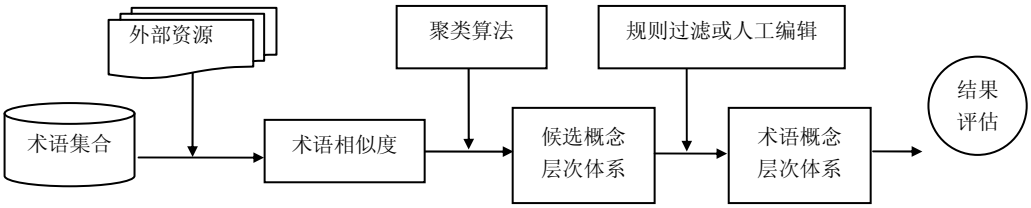


Figure 1. Flowchart of Concept Taxonomy Generation Based on Terminology Clustering
图 1. 基于术语聚类的概念层次体系生成流程图

术语相似度计算方面，本研究以词林为语义体系，进行术语的语义相似度计算，并以语料进行基于大规模语料统计的词语相似度计算，综合词语间的字面、语义及语料库上的相似度得到词语最终相似度。聚类算法方面，本研究采用 AP 聚类算法进行聚类，得到术语聚类结果。下面对术语聚类中涉及到的几个关键问题进行说明。

2.1 基于多语术语聚类的概念层次体系生成策略

根据聚类对象的不同，多语术语聚类可以采用两种策略，即：跨语言的概念层次体系生成策略（不同语种可以进行映射的概念层次体系）与混合语言的概念层次体系生成策略。

2.1.1 跨语言的概念层次体系策略

跨语言的概念层次体系是指不同语种可以进行映射的概念层次体系。如图 2 所示，本研究首先分别对中文和英文的术语进行聚类，分别得到中文与英文的概念层次体系，然后根据中英文双语概率词典或人工编辑的方法，对两种语种概念层次体系进行合并处理，从而得到跨语言的概念层次体系。其中，中文术语相似度计算采用综合相似度计算方法，综合利用字面相似度、语义相似度、基于语料库的相似度等；英文术语相似度为基于语料库的相似度。

双语概念层次体系中合并处理的基本策略为：将目标类簇中每个元素与源类簇的范例进行相似度计算，如果相似度高于源类簇中相似度（源类簇中每个成员与范例间的相似度）最低值，则认为该元素隶属于源类簇，如果目标类簇元素隶属于源类簇的比率超过给定阈值，则将目标类簇合并到源类簇，否则等待下一轮合并处理。

2.1.2 混合语言的概念层次体系生成策略

多语言的概念层次体系是包含多种语言术语的概念层次体系。如图 3 所示，本研究利用平行语料计算不同术语之间的相似度，然后调用聚类算法对包含不同语种的术语集合进行聚类，得到多语言的概念层次体系。

中英文术语相似度计算的依据为它们在平行语料上的分布情况，通过互信息、Dice、LogL 值等得到。基于混合语言的术语聚类，可以考虑不同语种在多语言本体或语义体系上的相似度、在平行语料上的相似度。本文在实验中，仅考虑中英文术语在平行语料上的相似度。混合语言的术语聚类结果为包含中英文术语在内的双语类簇。

2.2 术语的综合相似度计算

2.2.1 基于《同义词词林》的词语相似度计算

《同义词词林》[15]（亦称义类词典，以下简称《词林》）由梅家驹等学者编纂的一部对汉语词汇按语义全面分类的词典，收录词语近 7 万。《词林》将词义分为大类、中类、小类三级，共分 12 个大类，94 个中类，1428 个小类，小类下再以同义词原则划分词群，每个词群以一标题词立目，共 3925 个标题词。本文采用语义距离来计算词汇间的相似度。词语相似度计算过程的细节参见文[16]。

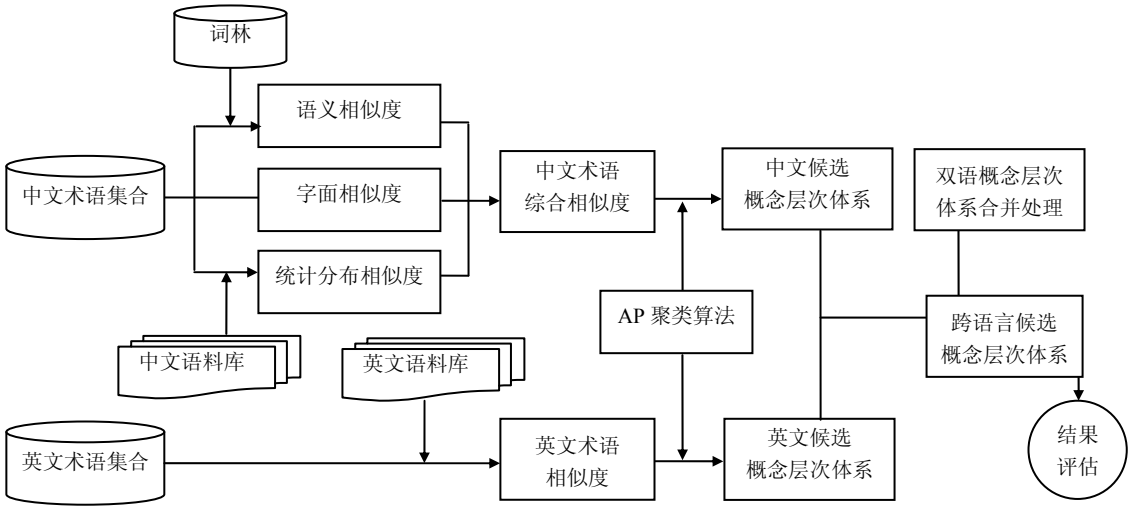


Figure 2. Flowchart of Concept Taxonomy Generation Based on Cross Language Terminology Clustering
图 2. 跨语言的概念层次体系生成流程图

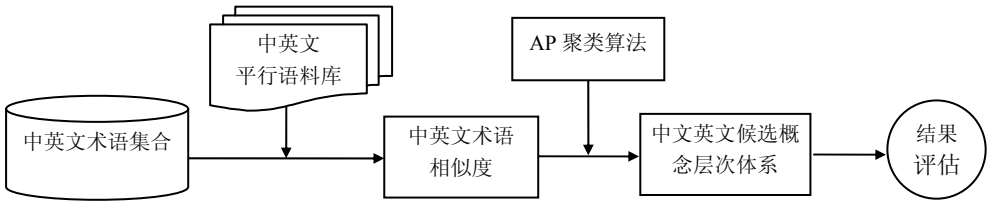


Figure 3. Flowchart of Concept Taxonomy Generation Based on Mixed Language Terminology Clustering
图 3. 混合语言的概念层次体系生成流程图

2.2.2 基于 Web 的词语相似度计算

基于 Web 的词语相似度计算方法，一般是将词语作为查询式，通过搜索引擎的返回结果的数目来度量不同词语之间的相似度。

2001 年，Turney 利用搜索引擎 AltaVista 搜索结果为依据，通过如下的公式计算两个词的相似度[17]。

$$\text{score}(A,B) = \text{hits}(A \text{ AND } B) / \text{hits}(B) \quad (1)$$

其中，hits(A AND B)为词语 A，B 同时出现的返回文档数，hits(B)为词语 B 查询返回文档数目。

2007 年 Rudi & Paul 将 Google 返回结果作为依据，将词语相似度计算问题转换为归一化的 Web 距离问题[18]。由于通过 Google、Baidu 等搜索引擎计算词语间的相似度比较耗时（主要是这些搜索引擎对一定时间段内的查询次数有限制），因此本文直接利用本地语料库代替基于 Web 的词语相似度计算，计算方式如公式 3 所示。

$$\text{Sim}_C(A,B)=\text{Log}_2\left[\frac{f(A,\ B)\ *\ N}{f(A)\ \cdot\ f(B)}\right] \quad (2)$$

其中，N 为语料库的规模。这里采用的是互信息公式计算词语在语料库上的相似度，还可以利用 Dice、LogL 等统计公式法计算词语间的相似度[16]。

2.2.3 术语的综合相似度

根据词语的字面相似度[19]、语义相似度以及基于语料库的统计分布上的相似度，进行线性组合[16]，得到 A 与 B 的综合相似度，计算公式如下：

$$\text{Sim}(A,B) = \alpha \cdot Sim_W(A,B) + \beta \cdot Sim_S(A,B) + \gamma \cdot Sim_C(A,B) \quad (3)$$

其中，$Sim_W(A,B)$ 为术语 A，B 的字面相似度，$Sim_S(A,B)$ 为术语 A，B 的语义相似度，$Sim_C(A,B)$ 为基于语料库的统计分布相似度（进行归一化后）。α、β、γ 分别为三种相似度的权重，且 α+β+γ=1。

本文在试验中，中文术语相似度计算采用字面、语义及语料库等三种可选方式进行计算，英文术语、中英文术语间相似度是通过平行语料上的相似度计算得到。

2.3 利用 AP 聚类算法进行术语聚类

Affinity Propagation (AP) 聚类是 2007 年在 Science 杂志上提出的一种新的聚类算法[20]。它根据 N

978-1-935068-76-1 © 2011 SciRes.

个数据点之间的相似度进行聚类，这些相似度可以是对称的，即两个数据点互相之间的相似度一样(如欧氏距离)，也可以是不对称的，即两个数据点互相之间的相似度不等。这些相似度组成 N×N 的相似度矩阵 S(其中 N 为有 N 个数据点)[21,22]。

AP 算法不需要事先指定聚类数目，相反它将所有的数据点都作为潜在的聚类中心，称之为范例（Exemplar）。以 S 矩阵的对角线上的数值 s (j, j)作为 j 点能否成为聚类中心的评判标准，这意味着该值越大，这个点成为聚类中心的可能性也就越大，这个值又称作参考度 P (preference)。聚类的数量受到参考度 P 的影响，如果认为每个数据点都有可能作为聚类中心，那么 P 就应取相同的值。如果取输入的相似度的均值作为 P 的值，得到聚类数量是中等的。如果取最小值，得到类数较少的聚类[21,22]。

AP 算法中传递两种类型的消息：假设有两个点 i 与 j，r(i,j)表示从点 i 发送到候选聚类中心 j 的数值消息，反映 j 点是否适合作为 i 点的聚类中心。a(i, j)则从候选聚类中心 j 发送到 i 的数值消息，反映 i 点是否选择 j 作为其聚类中心。r (i, j)与 a (i, j)越强，则 j 点作为聚类中心的可能性就越大，并且 i 点隶属于以 j 点为聚类中心的聚类的可能性也越大。AP 算法通过迭代过程不断更新每一个点的吸引度和归属度值，直到产生 m 个高质量的范例，同时将其余的数据点分配到相应的聚类中[21,22]。

在 AP 聚类算法中，各个聚类参数分别说明如下[21,22]：

similarity：数据点 i 和点 j 的相似度记为 S(i, j)，即：点 j 作为点 i 的聚类中心的相似度。一般用欧式距离来计算，本文使用聚类对象所对应向量的夹角余弦来计算相似度。

preference：数据点 i 的参考度或偏好值，称为 P(i)或 S(i, i)，是指点 i 作为聚类中心的参考度。一般取 S 相似度值的中值或者最小值。

Responsibility：R(i,j)用来描述点 j 适合作为数据点 i 的聚类中心的程度。

Availability：A(i,j)用来描述点 i 选择点 j 作为其聚类中心的适合程度。

Damping factor：阻尼系数，主要是起收敛作用的。

R(i,j)与 A(i,j)的计算公式分别如下：

$$R(i,j)=S(i,j) - \max\{A(i,k)+S(i,k)\} \quad (k\in\{1,2,...,N, \text{但 } k\neq j\}) \tag{4}$$

$$A(i,j)=\min\{0,R(j,j)+\sum_{k}\{\max(0,R(k,j))\}\} \quad (k\in\{1,2,...,N,\text{但 } k\neq i \text{ 且 } k\neq j\}) \tag{5}$$

AP 算法的具体工作过程为：先计算 N 个点之间的相似度值，将值放在 S 矩阵中，再选取偏好值（一般取偏好值的中值或者中值的 2 倍）。设置一个最大迭代次数，迭代过程开始后，计算每一次的 R 值和 A 值，根据 R(j,j)+A(j,j)值来判断是否为聚类中心（一般指定 R(j,j)+A(j,j)>0 时为一个聚类中心），当迭代次数超过最大值（即 maxits 值）或者当聚类中心连续多少次迭代不发生改变时终止计算[21]。

本文直接利用 Frey 等提供的 AP 聚类工具[1]对词语进行 AP 聚类。AP 聚类的初始结果作为候选的两层概念体系。通过调节 AP 聚类算法中的参考度 P 的大小获得不同聚类数目的聚类结果：偏好值小聚类数目则少，偏好值大聚类数目则多。因此，在较小偏好值的聚类后，进行偏好值大的聚类，可以得到下一个层次的聚类结果[23]。

3 实验结果与分析

3.1 测试数据与处理过程

本文采集了图书情报、法律、经济以及语言文字等四个领域类别的专业论文题录信息，包括学术论文的中英文篇名、中英文关键词等字段。本文以中英文关键词字段为测试数据，首先从中英文关键词字段中抽取中英文双语核心术语，并在此基础上通过多语言术语聚类得到双语概念层次体系。测试数据情况如表 1 所示，其中，图书情报类的规模最大，在中英文关键词字段中，不同的中文关键词总数为 25728 个，英文为 41729 个；语言文字类规模最小，不同的中文关键词总数为 12467 个，英文为 14047 个。

通过文[24]的双语核心术语抽取方法，分别得到以上四个领域进行中文和英文的核心术语。本文取每个领域中，术语得分位于前 10000（Top-10000）的术语集合作为待聚类的对象。在实验测试中，我们分别取 Top-100、Top-200、Top-300、Top-500、Top-1000 等五种术语集合大小，调用不同组合的相似度计算方法，根据 AP 聚类算法生成不同聚类策略下的多语言概念层次体系。AP 聚类算法中，为使得聚类结果类簇数目尽量少，本实验设置偏好值为聚类对象相似度值中的最小值。

[1] http://www.psi.toronto.edu/affinitypropagation/

Table 1. Test Dataset
表 1. 测试数据说明

标识	类别	中图分类号	平行语料规模	关键词总数		去重后关键词总数	
				中文	英文	中文	英文
1	图情	G25, G35	42954	124060	116471	25728	41729
9	法律	D9	7047	26046	25960	12743	16680
F	经济	F	31928	113024	113260	42872	58370
H	语言文字	H	5858	22055	22625	12467	14047

3.2 概念层次体系生成结果评价方法

概念层次体系的评价方法包括内部评价法与外部评价法。内部评价是对概念层次体系进行直接评价，而外部评价是指通过概念层次体系的应用效果，对其进行间接评价。本文采用内部评价方法，对生成的多语言概念层次体系进行评价，主要评价指标包括：聚类结果的净相似度（Net Similarity）[23]与聚类类目数。

①净相似度

Net Similarity=

$$\sum_{i=1}^{N}\sum_{j=1}^{M_i}\left(Sim\left(Term[i,j],\text{Exemplar}[\text{i}]\right)+\text{Preference}[\text{i}]\right) \quad (6)$$

其中：N 为聚类结果的类簇总数；M_i 为类簇 i (1≤i≤N) 中除范例以外的成员元素个数；$Sim(Term[i,j],\text{Exemplar}[\text{i}])$ 为类簇 i (1≤i≤N)中除范例以外的成员元素与范例之间的相似度；$\text{Preference}[\text{i}]$ 为类簇 i 中范例的偏好值；Net Similarity 表明 AP 聚类结束后目标函数所达到的最大值[23]。在本文中，我们将 Net Similarity 作为聚类结果评价指标之一，Net Similarity 越大，则聚类性能越优，即：概念层次体系越优。

②聚类类目数

聚类结果的类目层次反映了概念的分布情况。本文以聚类结果的一级类目数（即范例个数）为考察指标。一级类目数越大，则说明概念分布越分散。反之，一级类目数越小，说明概念分布越集中。

3.3 实验结果与分析

本文针对图书情报、法律、经济、语言文字等四个领域类别的数据进行多语言概念层次生成的实验。本节分别从不同多语言术语聚类策略下的聚类结果分析、跨语言聚类结果影响因素分析两个部分进行说明。

3.3.1 不同多语言术语聚类策略下的聚类结果分析

为了比较两种不同策略的多语言术语聚类结果，在实验中，设定相似度为基于平行语料的相似度，并采用互信息方式计算得到，合并阈值为 1.0。

针对四个领域类别，分别在单一语种概念数为 100、200、300、500、1000 等 5 种数量情况下，进行多语言术语的聚类。其中，跨语言的术语聚类，是在中文术语聚类和英文术语聚类完成后，进行跨语言的类簇合并来实现的；混合语言的术语聚类，则是各取给定数量的中文和英文术语，混在一起，依据平行语料进行去部分语种的相似度计算，然后再进行聚类。

四个类别的多语言术语聚类结果分别如表 5~8 所示。每一项聚类结果数据中包括：中文、英文、跨语言、混合语言等四种情况下的净相似度与一级类目数情况。列出中文和英文的术语集合聚类情况，主要是为了考察跨语言类簇的合并情况：中文类簇集合和英文类簇集合中有多少类簇被合并。通过表 2~5 可以看出，四个领域类别的跨语言聚类结果中，与没有进行类簇合并时相比，类簇数据减少，这说明相近的中英文术语在一定程度上被合并到同一类簇。基于混合语言的多语言术语聚类策略，聚类后的净相似度总体上高于跨语言的聚类结果的净相似度，而聚类的一级类目数低于跨语言聚类结果的一级类目数。

该结果表明，仅利用平行语料进行多语言术语聚类时，基于混合语言的多语言术语聚类策略要优于跨语言的聚类策略。

需要指出的是：在没有平行语料的情况下，基于混合语言的聚类策略由于无法计算不同语言术语间的相似度，造成该方法无法生成多语言概念层次体系；而跨语言的多语言术语聚类策略不受此限制。只要有每种语言的语料库或者语义体系，都可以进行该语种的术语间相似度计算，然后依据外部的翻译概率词典或双语词典，对不同语言聚类结果进行映射或合并。

综上，依据平行语料进行多语言术语聚类，基于混合语言的多语言术语聚类策略优于跨语言的术语聚类策略。在无平行语料，但有外部双语翻译概率词典或双语词典时，混合语言术语聚类策略失效，但跨语言的多语言术语聚类策略可以用来生成多语言的概念层次体系。

Table 2. Terminology Clustering Results of Library & Information Science
表 2. 图书情报类术语聚类结果

概念数	聚类类别	Net Similarity	一级类目数
	中文	976.4647	24

978-1-935068-76-1 © 2011 SciRes.

100	英文	1520.7656	23
	跨语言	1126.9561	45
	混合语言	2967.1406	49
200	中文	*	*
	英文	5702.5732*	46
	跨语言	*	*
	混合语言	11071.3525	84
300	中文	19942.2207	78
	英文	19941.9707	71
	跨语言	21631.1250	132
	混合语言	*	*
500	中文	37385.9023	109
	英文	40640.8086	100
	跨语言	42688.9805	189
	混合语言	72078.1797	204
1000	中文	218520.5781	225
	英文	*	*
	跨语言	*	*
	混合语言	353323.5000	316

注：“*”代表该参数配置下，AP 聚类算法未收敛，下同。

Table 3. Terminology Clustering Results of Law
表 3. 法律类术语聚类结果

概念数	聚类类别	Net Similarity	一级类目数
100	中文	1917.2438	26
	英文	1957.1853	27
	跨语言	2033.8512	49
	混合语言	3052.8347	40
200	中文	6941.5396	47
	英文	9055.7344	55
	跨语言	7483.4126	92
	混合语言	10431.9346	78
300	中文	16793.1504	77
	英文	18759.5059	77
	跨语言	17640.6367	145
	混合语言	22970.8203	111
500	中文	48848.9102	131
	英文	57087.8555	128
	跨语言	50547.1875	247
	混合语言	63655.1602	175
1000	中文	283823.5625	285
	英文	297318.6250	258
	跨语言	277761.8750	494
	混合语言	340132.1250	348

Table 4. Terminology Clustering Results of Economy
表 4. 经济类术语聚类结果

概念数	聚类类别	Net Similarity	一级类目数
	中文	1095.4644	22
	英文	1520.7656	23
100	跨语言	1123.9259	42
	混合语言	2247.9729	50
200	中文	3939.4097	41
	英文	*	*
	跨语言	*	*
	混合语言	8077.0718	97
300	中文	9245.7383	68
	英文	9973.5361	57
	跨语言	9859.2607	119
	混合语言	16509.6367	139
500	中文	26805.6094	101
	英文	28775.5371	111
	跨语言	28195.1035	203
	混合语言	44530. 8086	214
1000	中文	118912.7656	213
	英文	147548.7969	249
	跨语言	129804.8594	444
	混合语言	191058.8438.	396

Table 5. Terminology Clustering Results of Language
表 5. 语言文字类术语聚类结果

概念数	聚类类别	Net Similarity	一级类目数
100	中文	*	*
	英文	2224.1824	25
	跨语言	*	*
	混合语言	3456.6028	48
200	中文	7925.4922	49
	英文	7842.3325	53
	跨语言	8578.4023	99
	混合语言	12509.3564	87
300	中文	20205.7988	74
	英文	19742.4629	78
	跨语言	20400.0977	143
	混合语言	28978.9883	125
500	中文	60569.9063	130
	英文	58765.8242	128
	跨语言	62438.5977	242
	混合语言	*	*
1000	中文	262151.9375	286
	英文	264577.7500	272
	跨语言	258419.0781	510
	混合语言	336053.5313	374

3.3.2 跨语言术语聚类结果影响因素分析

①相似度计算方法对多语言聚类结果的影响

本文在实验中，利用图书情报领域作为实验对象。

相似度计算方法包括字面、语义以及语料库等三种基本方法，然后对这些方法按照不同权重比率进行线性组合，分别得到不同的跨语言聚类结果。其中，聚类的术语数目为 500，合并阈值为 1.0。

表 6 给出了不同相似度计算方法对跨语言聚类结果影响的情况比较。从表 6 可以看出，三种基本的相似度计算方法，聚类后得到类目数的排序为：

语义 < 字面 < 语料库。

Table 6. Influence of Similarity Computing Method on the Cross Language Terminology Clustering
表 6. 相似度计算方法对跨语言聚类结果的影响

相似度计算依据	中文概念层次体系一级类目数	跨语言层次体系一级类目数
字面	68	131
语义	29	123
语料库	109	189
字面+语义(权重分别为 0.3, 0.7)	25	50
字面+语料(权重分别为 0.7, 0.3)	35	63
语义+语料(权重分别为 0.7, 0.3)	29	124
字面+语义+语料库(全文分别为 0.1, 0.8, 0.1)	38	114
字面+语义+语料库(全文分别为 0.2, 0.7, 0.1)	46	129
字面+语义+语料库(全文分别为 0.2, 0.6, 0.2)	48	132

不同的相似度计算方法的组合，生成的类目数都不相同。仅考虑两种相似度的情况下，字面与语义结合的方法，生成的类目数较少。三种都考虑的情况下，加大语义相似的权重时，得到更少的聚类类目数。

需要指出的是，关于不同相似度计算方法生成的不同聚类类目数，能否用来判断聚类结果的优劣，我们将此作为下一步需要研究的工作。另外，针对目标语言（如中文）可以依据的相似度计算方法或资源较多，生成的聚类体系比源语言的概念体系更好的情况，如何通过跨语言映射将源语言聚类结果合并到目标语言聚类结果中，也是下一步需要研究的工作。

②跨语言类簇合并阈值聚类结果的影响

以图书情报领域作为实验对象，相似度为基于语料库的相似度，聚类的术语数目为 500，考察不同的合并阈值对跨语言聚类结果的影响。表 7 给出了阈值分别为 0.5、0.6、0.7、0.8、0.9、1.0 等六种情况下的聚类结果的一级类目数情况。通过表 7 可以看出，阈值越低，则不同语言间的类簇合并“门槛”越低，类簇数减少的越多。设置较高的合并阈值（最高值为 1.0），则被合并的类簇个数减少，多语言聚类结果中的一级类目数较多。因此，在跨语言的多语言术语聚类应用中，除了根据聚类算法自身的参数设置（如 AP 聚类算法中的偏好值），可以根据合并阈值控制最终聚类结果中的类簇个数。

Table 7. Influence of Threshold of Clustering Results Merging on the Cross Language Terminology Clustering
表 7. 跨语言类簇合并阈值对跨语言聚类结果的影响

合并阈值	Net Similarity	一级类目数
0.5	21172.1641	121
0.6	30034.4453	141
0.7	32891.2305	162
0.8	36234.5781	181
0.9	42688.9805	189
1.0	75034.6797	210

4 结论

本文首先对层次体系生成、词语相似度计算等相关研究进行概述；然后以 AP 聚类算法为基础，进行多语言术语聚类，给出聚类结果。本文从跨语言聚类和独立于语言的聚类两个角度对多语言概念层次聚类问题进行相对全面的研究。通过 4 个领域类别的数据进行测试，结果表明：仅依据平行语料进行多语言术语聚类，基于混合语言的多语言术语聚类策略优于跨语言的术语聚类策略。在没有平行语料，但有外部的双语翻译概率词典或双语词典时，混合语言术语聚类策略失效，但跨语言的多语言术语聚类策略可以用来生成多语言的概念层次体系。

今后进一步的工作包括如下几个方面：首先，研究如何有效整合不同语种的概念层次体系的方法，从而可以通过一种语言概念层体系优化另一种语言的概念层次体系。另外，进一步考察多个层次级别的概念层次体系生成；尝试对不同策略的多语言概念层次体系进行集成的方法，达到优势互补的效果，从而提高多语言概念层次体系的质量；其次，进一步研究多语言概念层次体系评估的方法，与现有的专业领域概念层次体系或其中的一部分与机器生成的概念层次体系进行比较，通过相符程度评价机器生成的概念体系的质量；另外，还可以尝试通过应用角度进行多语言概念层次体系的评估；最后，尝试结合多种语义资源（如 WordNet）与大规模语料（如 WWW 语料），提高术语的相似度计算的正确率。

致　谢

978-1-935068-76-1 © 2011 SciRes.

本研究受教育部人文社会科学研究一般项目“多语领域本体的自动构建研究”(08JC870007)，国家自然科学基金项目“基于可比语料的多语言文本聚类研究”(70903032)资助。

References (参考文献)

[1] Eric Tsui, W M Wang, C F Cheung, Adela S M Lau. A concept-relationship acquisition and inference approach for hierarchical taxonomy construction from tags [J]. Information Processing and Management. 2010, 46(1): 44-57.
[2] Harris Z. Mathematical stucutres of language [M]. Wiley, 1968.
[3] Stephan Bloehdorn, Philipp Cimiano, Andreas Hotho. Learning ontologies to improve text clustering and classification [A]. In: Proceedings of the 29th Annual Conference of the German Classification Society (GfKl2005) [C].Magdeburg, Germany, 2006: 334-341.
[4] Philipp Cimiano. Ontology Learning and Population from Text: Algorithms, Evaluation and Applications [M]. Springer-Verlag, New York, NY, USA, 2006.
[5] Alexander Maedche, Viktor Pekar, Steffen Staab. Ontology learning part one – on discovering taxonomic relations from the web [A]. In: Proceedings of the Web Intelligence conference [C]. New York, USA, 2002: 301-322.
[6] Inderjeet Mani, Ken Samuel, Kris Concepcion, David Vogel. Automatically inducing ontologies from corpora [A]. In: Proceedings of the 3rd International Workshop on Computational Terminology [C]. Geneva, Switzerland, 2004: 47-54.
[7] Hearst M A. Automatic acquisition of hyponyms from large text corpora [A]. In: Proceedings of the 14th international conference on computational linguistics (COLING1992) [C]. Nantes, France, 1992.
[8] Alves A, Pereira F, Cardoso A. Automatic reading and learning from text [A]. In: Proceedings of the international symposium on artificial intelligence [C]. Kolhapur, India 2001.
[9] Mark Sanderson, Bruce Croft. Deriving concept hierarchies from text [A]. In: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval [C]. Berkeley, California, USA, 1999: 206-213.
[10] K. W. Li, C. C. Yang. Automatic Cross-Lingual Thesaurus Generated from the Hong Kong SAR Police Department Web Corpus for Crime Analysis[J]. Journal of the American Society for Information Science and Technology 2005, 56(3): 272-282.
[11] C. C. Yang, C. Wei, and K. W. Li. Cross-lingual Thesaurus for Multilingual Knowledge Management[J]. Decision Support Systems, 2008, 45(3): 595-605.
[12] Hans Hjelm，Paul Buitelaar. Multilingual Evidence Improves Clustering-based Taxonomy Extraction [A]. In: Proceeding of the 18th European Conference on Artificial Intelligence (ECAI2008). Patras, Greece, 2008: 288-292.
[13] Keita Tsuji，Kyo Kageura. Automatic generation of Japanese–English bilingual thesauri based on bilingual corpora [J]. Journal of the American Society for Information Science and Technology. 2006, 57(7): 891-906.
[14] Aizawa, Akiko and Kageura, Kyo: A Graph-based Approach to the Automatic Generation of Multilingual Keyword Clusters [A]. In: Recent Advances in Computational Terminology [C]. Bouligault Didier, Jacquemin Christian, L'Homme Marie-Claude, eds. Amsterdam/Philadelphia: John Benjamins, 2001: 1-27.
[15] Jiaju MEI. TongYiCi CiLin. Shanghai: the Shanghai Lexicographical Publishing House, 1983.
梅家驹. 同义词词林[M]. 上海: 上海辞书出版社, 1983.
[16] Chengzhi ZHANF and Zhentian BAI. Research on automatic indexing and classification. Nanjin: Southeast University,2010.
章成志，白振田. 文本自动标引与自动分类研究[M]. 南京: 东南大学, 2010.
[17] Peter D. Turney. Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL [A]. In: Proceedings of the 12th European Conference on Machine Learning[C], Freiburg, Germany, 2001: 491-502.
[18] Rudi L Cilibrasi, Paul M B Vit´anyi. The Google Similarity Distance [J]. IEEE Transactions on Knowledge and Data Engineering. 2007, 19 (3): 370-383.
[19] Keqiang WU. Development of post-controlled vocabulary used for retrieving economic information. Nanjing: Nanjing Agricultural University, 1999.
吴志强. 经济信息检索后控制词表的研制[D]. 南京: 南京农业大学, 1999.
[20] Brendan J Frey, Delbert Dueck. Clustering by Passing Messages between Data Points [J]. Science, 2007, 315 (16): 972-976.
[21] Wufeng KE. Web-based Chinese-English Bilingual Term Extraction. Beijing: Tsinghua University, 2009.
柯武峰. 基于 Web 的中英文术语自动抽取技术[D]. 北京: 清华大学, 2009.
[22] Affinitiy Propagation Clustering Algorithm [EB/OL]. [2009-12-31].
http://hi.baidu.com/river_hiter/blog/item/a5ae8c08fa34c02d6a60fb77.html/cmtid/86b188a0f0a87b8746106413
[23] FAQ for Affinity Propagation [EB/OL]. [2009-12-31].
http://www.psi.toronto.edu/affinitypropagation/faq.html#def
[24] Zhang Chengzhi. Extracting Chinese-English Bilingual Core Terminology from Parallel Classified Corpora in Special Domain [A]. In: Proceedings of Workshop on Natural Language Processing and Ontology Engineering (NLPOE 2009) in conjunction with Conference on Web Intelligence (WI/IAT-09) [C]. Milan, Italy, 2009: 271-274.

Analysis of NSFC's Cooperation and Exchange Based on S&T Monitoring

——A Case of Major International (Region) Cooperation and Exchange Project

Qingfeng DUAN[1,2], Xuefeng WANG[1], Xiaofan HENG[1], Donghua ZHU[1], Jianling CHEN[1]

[1]*School of Management and Economic, Beijing Institute of Technology, Beijing, China*

[2]*School of Economic and Management, North University of China, Taiyuan, China*

Email: dqf01@sina.com

Abstract: The S&T monitoring was proposed to be applied in NSFC based on the analysis of decision issues in science foundation management. It is helpful to quickly find the valuable knowledge from numerous data of science and technology, which also support the decisions of managers and experts. After the introduction of analyzing process, we illustrated the necessity of employing the S&T monitoring in NSFC and the characteristics of their interactions. Lastly, a case of the Major International (Region) Cooperation and Exchange Project in NSFC was empirically employed in the views of keywords and affiliations, which demonstrated its effectiveness.

Keywords: S&T monitoring; cooperation and exchange; knowledge mapping; correlation analysis

基于科技监测的科学基金合作交流分析

——以重大国际（地区）合作研究项目为例

段庆锋[1,2]，汪雪锋[1]，衡晓帆[1]，朱东华[1]，陈建领[1]

[1]北京理工大学管理与经济学院，北京，中国，100081

[2]中北大学经济与管理学院，太原，中国，030051

Email: dqf01@sina.com

摘　要：通过分析科学基金管理面临的决策问题，将科技监测方法应用于科学基金管理，从大量的科技数据中快速发现有价值的知识，为管理者及专家提供决策支持。介绍了科技监测的分析流程，分析了科技监测在科学基金资助管理中的应用必要性及两者的关系特征。最后以国家自然科学基金重大国际（地区）合作研究项目为例，从关键词及依托单位的视角进行了知识图谱分析，说明了方法有效性。

关键词：科技监测；合作交流；知识图谱；关联分析

在当代科学界科学家之间的合作与交流已经成为常态，对一个国家的科学水平及能力发展有着重要影响。科学基金制在我国科技创新体系中发挥重要作用，以国家自然科学基金(NSFC)为代表的科学资助计划对我国的基础研究发挥了重要作用。合作与交流类项目是国家自然科学基金资助体系中的一类重要类型，它的战略定位在于资助我国科学家同国际同行专家开展各种形式的合作研究及学术交流活动，以促进我国科学研究的发展，提高我国科技在国际的地位。

面对科学资助的海量事务性数据，一般的数量层面的统计分析所刻画的特征信息已经无法满足科学基金的管理需求，尤其需要从语义的层面理解科学技术的发展动态、趋势及潜在方向。科技监测已经成为海量数据环境下，快速有效掌握科技发展动态的方法，也受到国内外相关学者的关注，如 Porter[1]、朱东华[2]等。因此，本文将知识图谱的分析方法引入科学基金的决策管理之中，通过自动化的计算机监测和分析程序，从海量数据库中快速萃取出有价值的模式（知识），并通过图形的方式展示，为科学基金的决策提供支持，为科学基金的管理提供新的分析工具及思路。

1 科技监测

科技监测是一个多学科交叉、借助多种工具实现知识发现的过程，一般可以归纳为四个部分，包括数据采集、数据处理、技术分析、知识表达，如图 1 所示。

本文受到国家自然科学基金委主任基金项目（J0910016）的资助。

978-1-935068-76-1 © 2011 SciRes.

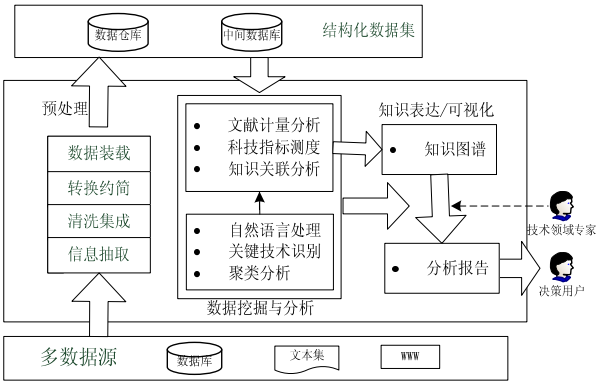


Figure 1. The framework of S&T technology
图 1. 科技监测的框架体系

数据采集的首要步骤为确定数据源，由于学术论文和发明专利是基础科学、应用科学研究成果的优质载体，因此世界上公开出版的检索数据库成为最佳的采集来源，如 EI、INSPECT、USPTO 等。随着智能信息处理技术的发展，非结构化数据也逐渐成为采集分析的对象，如一般的文本集、互联网中的页面信息等，多数据源的采集及融合处理将成为今后的发展趋势[2]。另外，通常数据集是与特定技术领域相关的，需要确定技术内容范畴并制定检索策略，其中头脑风暴法是确定可能技术关键词的一种有效方法[3]。

与一般的数据挖掘分析过程相似，数据的预处理是科技监测的不可缺少的过程。通过信息抽取、清洗集成、转化约简、数据装载阶段，将原始数据转换为去除噪声、便于分析的数据集。数据的挖掘与分析建立在文本挖据、文献计量、聚类分析、技术指标测度及计量回归分析等多种分析手段基础之上，以定量及半定量化的方法，分析特定领域技术发展的现状及趋势，为企业、组织达到其战略目标而提供决策支持。其中基于共现关系(Co-occurrences)的关联分析是常用的方法，可以对研究对象间的联系进行识别和度量。一般认为如果两个元素频繁地一同出现，则其具有某种内在关联性[4]，这种共现关系通常体现为两种形式，即共引(Co-citation)[5]和共词(Co-word)[6]。

将技术分析得到的模式（知识）以决策者易于理解的形式输出是科技监测获得有效应用的重要因素，其中知识图谱、技术拟合曲线等是重要的输出形式。多维标度法(MDS)[7]是有效地将多维数据映射到二维或三维空间的有效算法，通过计算机可视化程序生成图谱，形象地展示研究对象的关系及其结构[8]。在此基础上，自动及半自动化的科技监测分析报告生成技术，可以有效地提高技术分析、决策的效率和效果[9]。

2 科技监测与科学基金管理

科学基金通常由政府拨款或企业、团体、个人捐赠形成基金，通过自由申请、专家评审、择优支持等方法配置科技资源，在管理中引入竞争机制和激励机制，依靠科学家群体来管理科学活动[1]。在我国，国家自然科学基金的职能主要在于：制定和实施支持基础研究和培养科学技术人才的资助计划，促进科研资源的有效配置；协同国家科学技术行政主管部门制定国家发展基础研究的方针、政策和规划；同其他国家或地区的政府科学技术管理部门、资助机构和学术组织建立联系并开展国际合作等。

将科技监测的方法纳入科学基金决策的过程，对于其战略目标及上述职能的实施具有积极的促进作用，弥补了专家评价的主观性及对于海量数据的处理能力不足。不同于同行评议主要借助于专家的隐性知识进行评价，科技监测可以从数据库中挖掘、发现“隐藏”在海量数据中的知识；两者相互结合，可以为科学基金管理提供更加全面的决策支持信息，这些信息有助于决策者准确、快速地做出科技发展及资助形势的判断，并对未来技术发展趋势做出科学的预测。在科学基金战略目标的导向下，科技监测方法“挖掘”得到的知识，对于科学基金实现战略目标提供了支持，如图 2 所示。

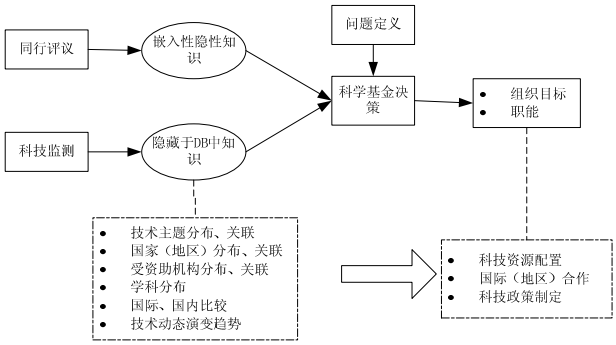


Figure 2. Relationship of S&T monitoring and NSFC management
图 2. 科技监测与科学基金管理的关系

科技监测的方法可以应用在不同的环境中，为达到组织的不同战略目标提供决策支持。尽管在不同情境下，科技监测的方法、关键技术基本相同，但在组织的不同战略目标导向下，科技监测的应用在目的、内容、功能方面存在一定差异。在企业、科研机构、科学基金三中应用情境下，科技监测的相应特征如表 1 所示。

Table 1. Characteristics comparison of S&T monitoring in different scenario
表 1. 不同情境下科技监测特征比较

	主要数据源	目的	内容	功能
企业情景	• 专利	• 获取企业的技术优势 • 获取企业的竞争力	• 识别技术现状、机会 • 识别竞争对手及特征 • 技术领域演变状况	• 制定 R&D 策略、技术竞争策略 • 制定 R&D 联盟策略、
科研机构情景	论文、专利 •	• 寻找科技创新点 • 提高科研创新能力	• 识别学科研究热点、趋势 • 前沿科学家、机构、地区、国家分布	• 选择合作研究、交流对象、途径 • 制定 R&D 策略
科学基金情景	资助数据、论文、专利 •	• 提高资助效率、效果 • 提高科学共同体满意度	• 资助的结构（科学家、机构、地区、国家） • 资助学科的热点主题 • 资助学科的主题演变 • 资助的国际、国内比较	• 优化科技资源配置 • 制定科技、资助政策 • 制定国际（地区）合作策略、途径

3 合作交流资助分析

3.1 关键方法

技术主题词的识别是项目内容分析的重要环节，建立在自动化分词程序的基础上，从项目标题及研究内容摘要中抽取出具有意义的技术关键词。对于分词程序抽取识别出的结果，进一步需要相关领域专家的检验及修正，最终得到精炼后的与项目关联的技术关键词集。本文根据分析的对象不同，采用了不同的关联矩阵生成方法。

技术关键词的关联分析建立在共生关系（Co-occurrences）的基础上，如果两个主题词出现在同一个项目中，则认为这两个主题词具有共生关系。

假定有 n 个项目，这 n 个项目名称中可能包含有 m 个技术主题词，则可建立{n 个项目*m 个技术主题词}的矩阵 X。在矩阵 X 中，项目 D_i 的主题词 $Term_j$ 的权值用布尔代数值表示，当 $Term_j$ 在项目 D_i 中出现时取 1，否则取 0，其中 n 是所有项目的总数，m 是所有项目所包含的主题词总数。基于矩阵 X，可以进一步得到{主题词*主题词}的共生关联矩阵 $T=X^T \cdot X$，其中矩阵 T 中任一元素 t_{ij} 表示第 i 个主题词和第 j 个主题词在同一个项目名称中出现的频数，即共现频数。主题词 i 与 j 的关联度为 t'_{ij}，其中 t'_{ij} 为矩阵 T' 中的元素，T' 为 T 的归一化矩阵。

对除了技术关键词之外的其它研究要素的关联分析采用了余弦相似度的度量方法。以项目的依托单位为例，设向量$(a_1,\ldots,a_i,\ldots,a_n)$和$(b_1,\ldots,b_i,\ldots,b_n)$分别表示依托单位 A 和 B 的特征向量，其中 a_i 表示主题词 $Term_i$ 在依托单位 A 所承担的所有项目中出现的频数，b_i 的含义同理，n 为主题词的总数。采用余弦相似度，定义 A 与 B 的关联度为

$$sim(A,B)=\sum_{i=1}^{n} a_i b_i \Big/ \sqrt{\sum_{i=1}^{n} a_i^2 \times \sum_{i=1}^{n} b_i^2}\ ,i=1,\cdots,n.$$

在关联矩阵的基础上，采用多维标度法（MDS）将研究对象映射到二维空间中，形成知识图谱，这样可以有效地发现要素之间的关系模式。将分析的结果，以自动或半自动的方式转化为人易于理解的模式，通过可视化的技术生成知识图谱，并输出生成监测分析报告。

3.2 项目资助概况

国家自然科学基金是我国最早成立的采用基金制的科学资助组织，目前已经形成了面上项目、人才项目和环境类项目的三大资助系列，是国家创新体系的重要组成部分。合作与交流类项目是隶属于环境类项目的重要资助类型，主要包括了重大国际（地区）合作研究、国际（地区）合作研究、在华召开国际学术会议、中德科学中心合作项目等。本文选取国家自然科学基金重大国际（地区）合作研究项目为科技监测对象，主要以关联分析和知识图谱的方法分析其项目资助的状况。

重大国际（地区）合作研究项目的资助目的在于为顺应科学研究国际化趋势，最大限度地利用国际科学资源，增强我国基础研究国际竞争能力，推动我国基础研究若干领域进入国际先进行列，它是国家自然科学基金委员会设立的专门面向国际科学交流的资助项目类型。

本文分析的样本数据来源于国家自然科学基金数据库，对源数据进行清洗，将错误数据和冗余数据删除，去除噪声和无关数据，并转换成有效形式，共抽取 239 条数据，时间跨度涵盖 2003 到 2008 年。

图 3 反映了按年代统计的重大国际（地区）合作

978-1-935068-76-1 © 2011 SciRes.

研究类项目资助情况的变化。从资助的金额和数量看，整体呈上升趋势，只是在 2006-2007 年稍有回落，在 2008 年继续上升，在 2008 年达到金额和项目数的最大值。项目数从 2003 年的 20 项，到 2008 年的 62 项，上升了 210%。这样的增长幅度在国家自然科学基金资助体系中是较为突出的，说明它是被重点发展、强化的资助项目类型。

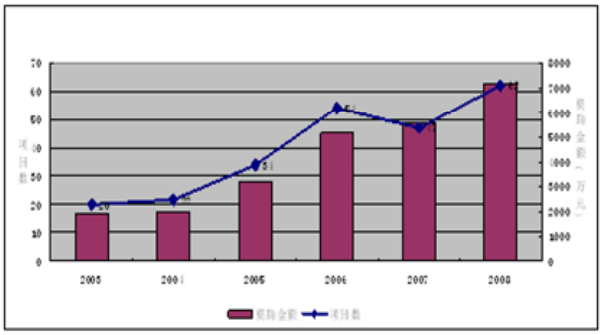


Figure 3. Sponsoring of Major International (Region) Cooperation and Exchanging Project
图 3. 重大国际（地区）合作研究项目资助状况

3.3 项目关键词分析

图 4 显示了资助项目技术关键词关联知识图谱，其中原点的大小代表关键词频率高低，关键词间的连线表示两者间存在关联。图中显示了关联度大于阈值 0.4 的连线，选取了出现频率前 20 名的关键词。技术关键词反映了该项目资助的技术热点，有助于有效地发现科学技术发展的趋势。鉴于图中的样本数不多，直接通过观察的方法可以大致将这些关键词划分为 6 个大类，如图中圆圈所示，分别以聚类中的最高词频技术关键词为代表，它们可以归纳为 1）纳米、材料；2）地震；3）气候、高原；4）生物、土壤；5）基因、细胞；6）生态、环境。这些技术关键词聚类基本反映了样本时间跨度内，国际（地区）合作交流项目的资助内容，六大聚类基本上代表了相应的六个资助热点及趋势，尽管它们之间存在不同程度的交叉，但是通过知识谱图可以很直观地理解项目资助的学科热点，为科学资助政策的制定提供有价值参考。

为了对比验证，对技术关键词进行因子分析，采用最大方差法进行正交旋转，得到八个主成分，累计方差贡献率达到 80.617%，如表 2 所示。通过进一步分析发现，前 4 个主成分与图中的相应聚类（按序号）对应，而后面 4 个主成分可以对应为聚类 5 和 6 中的部分主题。主成分 6 和 8 大致对应于聚类 5，主成分 5 和 7 大致对应聚类 6。通过研究发现，初始特征值大于 2 的主成分基本上能够与图中的聚类找到对应，采用这两种方法得到的结论基本一致。

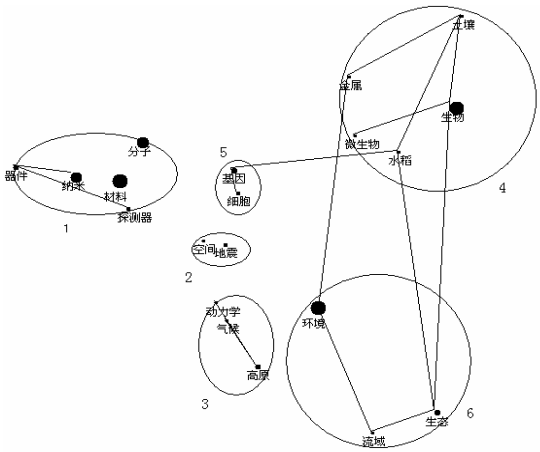


Figure 4. Keyword-based knowledge mapping of sponsored projects
图 4. 资助项目技术关键词知识图谱

Table 2. Keyword's Principal component of sponsored projects
表 2. 资助项目技术关键词主成分列表

主成分	初始特征值	方差贡献率%	累计方差贡献率%	旋转后的因子负载平方和	方差贡献率%	累计方差贡献率%
1	3.279	16.397	16.397	2.817	14.086	14.086
2	2.640	13.2	29.597	2.192	10.96	25.046
3	2.313	11.565	41.162	2.168	10.841	35.887
4	2.018	10.088	51.25	2.07	10.351	46.238
5	1.784	8.92	60.17	2.046	10.23	56.468
6	1.683	8.413	68.583	1.814	9.07	65.538
7	1.33	6.649	75.232	1.585	7.927	73.464
8	1.077	5.385	80.617	1.431	7.153	80.617

3.4 项目依托单位分析

图 5 显示了出现频率前 20 名的项目所属依托单位关联图谱，样本点之间的连线代表依托单位之间存在较强的研究项目方向上的一致性；也就是说如果两个依托单位之间的关联关系越强，则说明在研究方向上它们具有越高接近性。以项目的技术关键词向量为基础，关联性采用了余弦相似度刻画依托单位之间的接近程度。图 5 在图形分布上构成倾斜的倒 U 形结构，处于顶部的依托单位不仅获得的项目资助数量较多，而且与其它依托单位研究内容的关联性也较强。清华大学、北京大学、中国科学院化学研究所、中国科技大学和中国科学院南京土壤研究所处于关联簇的中心位置，它们基本构成了项目资助的热点单位，这些大

学及科研机构的综合实力及相关优势学科领域，在国内处于领先的地位。

经过进一步跟踪分析，汇总出这几个依托单位的主要一级学科分布，如表 3 所示。值得注意的是，尽管浙江大学和中国科学院高能物理研究所获得该资助项目数也很高，但其项目在内容上与图中其它依托单位的关联不大，说明它们具有相对独特的学科研究优势领域。

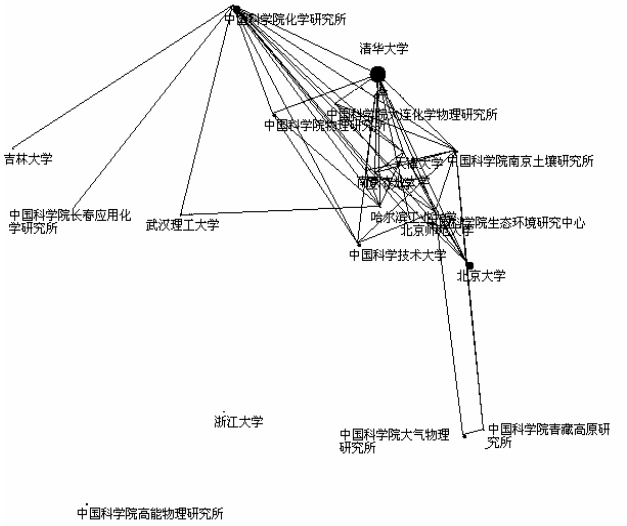


Figure 5. Knowledge mapping of projects' affiliations
图 5. 依托单位关联知识图谱

Table 3. Discipline distribution of projects' affiliations
表 3. 依托单位主要学科领域分布

机构	主要学科领域
清华大学	A05（基础物理学）；F04（半导体科学）；A02（力学）
北京大学	D01（地理学、土壤学和遥感）；D04（地球物理学和空间物理学）
中国科学院化学研究所	B03（物理化学）；B04（高分子化学）；B05（分析化学）
中国科学技术大学	A03（天文学）；A05（基础物理学）
中国科学院南京土壤研究所	D01 （地理学、土壤学和遥感）

3.5 合作国家分析

图 6 显示了重大国际（地区）合作研究项目的主要合作国家及地区的关联图谱。从分布的形态看，与图 5 所示的项目依托单位关联图谱很相似，整体上呈现倾斜的 U 型结构，其中合作最为频繁的主要集中在 U 形底部，在图 6 中集中在右下角区域。位于 U 形底部的国家（地区）可以被视为国际科学合作交流的中间枢纽，以美国、日本、英国、德国为代表的国家是世界科学研究的中心和学术前沿领导者。由这些核心国家向外呈扇形扩散，合作的数量及强度逐级减弱；在达 U 形顶端附近的国家、地区及组织则在合作项目的样本中处于边缘地位，我国与它们的项目合作相对不够紧密。从整体上合作的态势，可以看出我国对外科学合作交流的“马太效应”，一定程度上，也是在科学基金对外合作交流政策的导向下体现的有目的选择科学合作对象的结果。

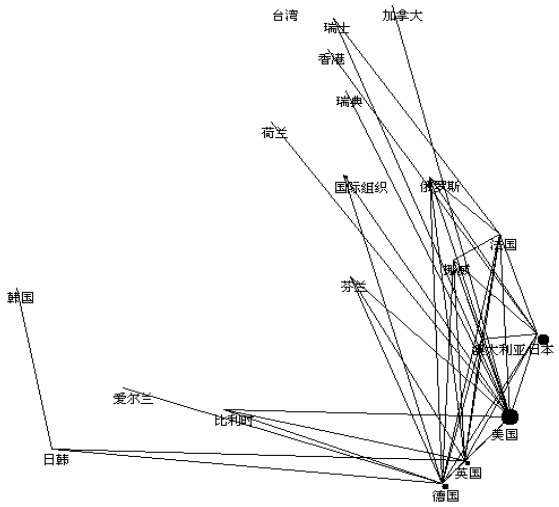


Figure 6. Knowledge mapping of cooperative countries (region) in projects
图 6. 项目合作国家（地区）关联图谱

4 结语

通过本文的技术关键词、依托单位及合作国家（地区）方面的实证分析，说明了该方法在科学基金管理中的有效性及应用前景。科技监测的方法优势在于以图形化的表达方法，将隐藏在海量数据中的深层次的知识呈现给决策者，并辅助决策者快速把握问题的关键特征。将科技监测方法应用于科学基金的管理，理论及实践上还处于探索阶段，相信随着科技监测理论及方法的不断完善将很大程度上促进科学基金管理的科学化，提高科技管理过程的效率。

References (参考文献)

[1] Porter A, Cunningham S. Tech mining: exploiting new technologies for competitive advantage [M]. Hoboken: John Wiley & Sons, 2005

[2] Zhu DH, Yuan JP. The S&T monitoring approach based on data mining [J]. Journal of industrial engineering Management, 2004, 118(14): 135-139.
朱东华，袁军鹏. 基于数据挖掘的科技监测方法研究[J]. 管理工程学报, 2004, 118(14): 135-139.

978-1-935068-76-1 © 2011 SciRes.

[3] Porter A L, Detampel M J. Technology opportunities analysis[J]. Technological Forecasting and Social Change, 1995, 49: 237-255.
[4] Kostoff R N. Research impact assessment: problems, progress, promise[C]. Miami FL: 1994.
[5] Small I I, Griffith B. The structure of scientific literatures [J]. Science Studies, 1974, 4: 17-40.
[6] Kostoff R N. Co-word analysis[M]. Boston: Kluwer, 1993: 63-78.
[7] Kang Yuhang, Su Jingqing, Diao Xiaochun. Technology opportunity analysis based on MDS[J], System Engineering, 2007, 25(12): 80-83.
康宇航，苏敬勤，刁晓纯. 基于多维标度的技术机会分析[J]. 系统工程, 2007, 25(12): 80-83.
[8] Chen C M. Visualising Semantic Spaces and Author Co-Citation Networks in Digital Libraries[J]. Information Processing & Management, 1999(35): 401-420.
[9] Porter A L. QTIP: Quick technology intelligence processes[J]. Technological Forecasting and Social Change, 2005, 72: 1070-1081.

RCPE Review on Agricultural Ontology Services Research

Xiaolu SU[1], Jing LI[2], Xianxue MENG[1], Yunpeng CUI[1], Ping QIAN[1]

[1]*Key Laboratory of Digital Agricultural Early-warning Technology Ministry of Agriculture, Agricultural Information Institute of Chinese Academy of Agriculture Science, Beijing, China*

[2]*China National Institute of Standardization, Beijing, China*

Email: suxiaolu@mail.caas.net.cn, lij@cnis.gov.cn

Abstract: This paper gives some brief introduction about the work of agricultural ontology services research group of Agricultural Information Institute of Chinese Academy of Agriculture Science, including the research on ontology construction theory and methodology, Large-scale Ontology Development Environment (LODE), and ontology application etc.

Keyword: ontology constructing theory; ontology constructing method; ontology development system; knowledge domain model; ontology search; ontology service

农业知识本体服务研究课题组工作回顾

苏晓路[1]，李景[2]，孟宪学[1]，崔运鹏[1]，钱平[1]

[1]农业部智能化农业预警技术重点开放实验室，中国农业科学院农业信息研究所，北京，中国，100081

[2]中国标准化研究院，北京，中国，100081

Email: suxiaolu@mail.caas.net.cn, lij@cnis.gov.cn

摘　要：本文回顾中国农业科学院农业信息研究所农业知识本体服务研究课题组开展的主要研究工作。对课题组提出的知识本体的构建理论与方法、大规模知识本体开发环境（LODE）的研发，以及知识本体在用户建模、文献检索与组合服务等方面的应用研究进行了简介。

关键词：本体构建理论；本体构建方法；本体开发系统；知识领域模型；本体检索；本体服务

1 引言

在国内外对知识本体这一研究主题有了广泛共识，该研究领域即将走向繁荣和活跃的背景下[1,2]，课题组结合本单位的工作特点和研究基础，确定了农业知识本体这个研究方向，全面系统地对本体的概念、理论、方法、工具和标准进行研究[3-12]，提出了基于元认知的知识抽象层次的理论和基于知识领域模型的大规模知识本体管理方法，并以该方法为基础研发 LODE 原型系统，以及相关的本体应用系统的开发与试验。

2 本体构建方法与系统

2.1 本体构建理论与方法研究

根据人脑的元认知机制（人脑中存在着的控制其自身思考活动的机制），提出了将知识划分为各个不同的抽象层次的需要。包含海量知识的大规模知识本体，必然是这样一种多抽象层次的知识体系。对于具有多个抽象层次的知识体系，单纯使用逻辑的方法不足以有效表达其内涵，需要建立一种推理与建模相结合的知识库系统框架，由此提出基于知识领域模型的大规模知识管理方法，以同时使用逻辑推理和模型仿真两种方法来处理具有多个抽象层次的知识本体。

知识领域模型的原理包括知识网络与知识距离、知识运用的局部性与领域的范围、知识抽象层次与知识领域结构、知识的“普遍－专门”联系与知识领域结构、知识领域的内容、知识领域边界与跨领域使用知识的方法，及知识领域模型的构建等方面问题[13]。

2.2 LOED 系统的开发

开展基于知识领域模型的大规模知识本体管理的研究，研发 LODE(Large-scale Ontology Development

978-1-935068-76-1

Environment）知识本体管理系统，作为多人协作大规模知识本体开发的工具。LODE 系统同时使用 C/S 结构和 B/S 结构。对于知识本体开发来说，以 C/S 为主，对于知识服务和信息服务来说，以 B/S 为主。目前已基本完成 LODE 客户端的开发和部分服务器端的开发。客户端的主要由知识领域管理、知识本体编辑、按领域管理词汇，以及文档管理和标引模块组成[14]。在服务器端已完成的部分有知识领域与知识本体的存储、知识配置管理、知识库查询等，计划开发的部分有知识融合、知识可视化、模型管理与模型推演等。LODE 系统完全完成后，将成为一个基于知识领域建模理论、支持多人并行开发大规模知识库、支持基于知识的建模与仿真、支持知识融合与知识发现的多功能综合知识开发与管理平台。LODE 系统知识领域管理界面如图 1 所示。

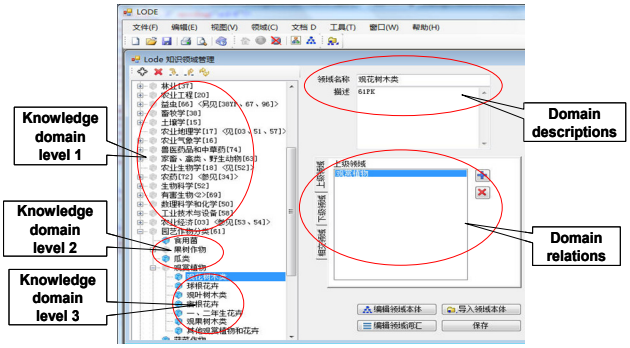


Figure 1. The knowledge domain management interface
图 1. LODE 系统知识领域管理界面

2.3 本体进化与可视化

开展本体进化研究主要是对用户检索行为和自动标引中未识别的词汇进行处理，通过计算词汇与本体中概念的关系，进行本体进化过程中概念的推荐。

在文献标引模块中将《农业科学叙词表》中的 6 万余条概念和词汇加入分词系统，增加了对农业科技信息分词的准确性；词汇与知识领域关系判断模块根据未识别词汇在文档中与已识别概念的位置关系和已识别概念所属领域的统计分析的结果，推荐潜在可能的领域，领域中与未识别词汇相关概念的数量如图 4 左侧领域名称后面的数字所示；词汇与概念间关系可视化模块将未识别词汇与概念间距离（直接联系与非直接联系）、相关概念间关系的类型（父子、相关）和关系的数量以图形方式显示出来，便于人工的选择，可视化界面如图 2 所示。

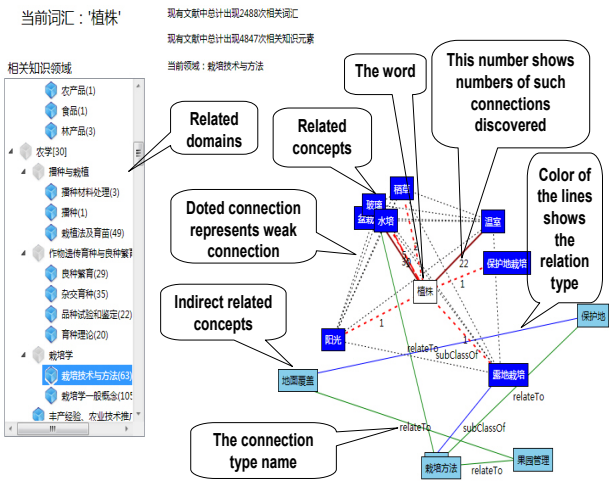


Figure 2. Visualization of unrecognized word's relation network
图 2. 未识别词汇与概念间的关系

3 本体构建与应用

3.1 农业科学初级本体构建

利用 LODE 系统客户端进行《农业科学叙词表》管理，包括 342 个领域、共有 5 万余条概念和 1 万余条同义词。领域分为三层，每个领域都可以拥有一个本体和一个词汇表，本体中包含领域中的概念，词汇表中包含指称知识对象的词汇，在一个确定的领域中一个词汇只能指称唯一一个知识元素，而一个知识元素可以有多个词汇来指称。本体使用 OWL 格式存储。

3.2 用户建模

采用知识本体与概念向量相结合的方法，开发了农业科技信息用户建模系统，构建基于知识本体的农业科技信息用户模型，试验结果表明，使用用户模型对检索结果排序可以改善的排序效果。用户建模及检索结果排序系统设计如图 3 所示。

系统从用户检索式、用户浏览题目和用户下载文章中的文摘抽取概念，为用户模型确定知识范围。由于所获取的用户知识结构的不完整性，以《农业科学叙词表》为参照依据本体，补充用户的知识结构，建立用户知识框架。采用 TF-IDF 算法的思想，用本体中的概念匹配文献中的词汇，用概念代替词汇统计用户的兴趣点，计算用户所关注文献中所有概念的出现频率，将其以向量的形式表示用户对概念的偏好程度[15-17]。

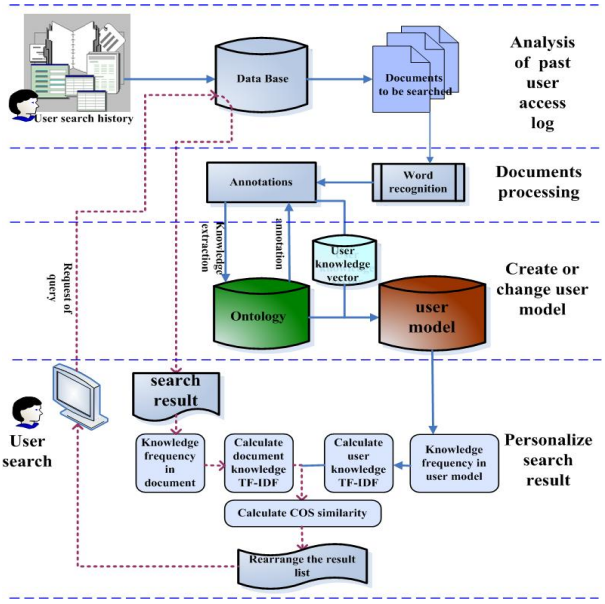


Figure 3. Architecture of user modeling
图 3. 用户建模系统设计

3.3 本体检索系统

知识本体检索系统是一个面向 Internet 可支持多本体同时进行复杂查询条件查询的系统，系统由知识本体查询中间件，知识本体查询 Web 服务，知识本体查询网站三部分构成，其中知识本体查询中间件可作为独立软件提供给其他开发者使用。

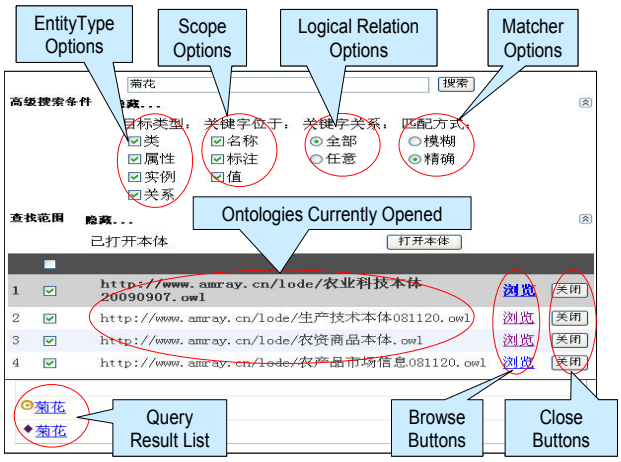


Figure 4. User interface of ontology query system
图 4. 本体检索系统界面

知识本体查询系统的基本功能包括：打开和关闭知识本体；同时从多个本体当中查找内容；对知识本体内容通过检索词进行查询；对本体元素的类型、检索词的位置、关系和匹配方式进行选择检索；检索结果的显示分为细览和概览，概览将符合查询条件的节点列表显示，细览准确展示检索词在本体中的节点位置以及与该节点相关的其他节点的信息。本体检索系统界面如图 4 所示。

3.4 LODE 检索系统

LODE 检索系统是一项基于知识网络构型相似匹配的检索研究，同时也是为了配合 LODE 系统的知识本体管理方法和文献处理方法而开展的一个初步的应用探索。检索系统利用知识本体，将用户的检索词所代表的知识元素及其之间的连接构成知识网络，与根据文档中的词汇所构成的网络相比较，找出其中相似度超过一定数值的文章提供给用户。

检索系统提供选择知识领域、知识元素和同义词进行检索（如图 5 所示），以及显示根据检索条件和知识元素生成的最大知识网络关系图和最小知识网络关系图。对检索结果可按照词语 TF/IDF、知识元素 TF/IDF 及最大图三种方式的排序结果比较[13]。

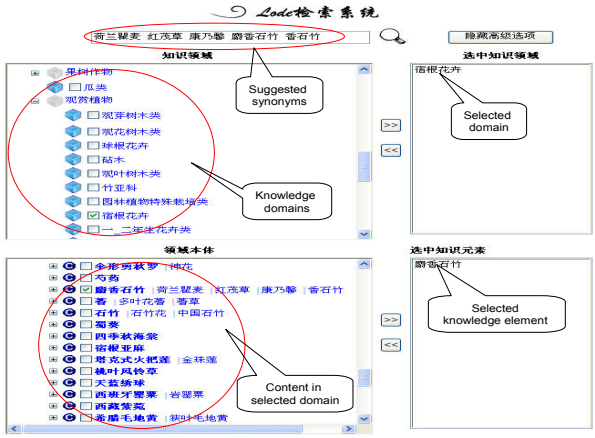


Figure 5. LODE search system (part of the page)
图 5. LODE 检索系统界面

3.5 文献服务组合

文献服务组合研究是针对互联网上众多的文献服务，通过使用智能代理技术和语义描述技术，让智能代理通过各个服务的语义描述获得对服务能力的知识，从而将服务有机地组合起来，用户通过统一的服务界面即可在后台自动调用一组相互配合的服务完成所需的任务。

文献服务组合包括主题浏览服务组合、文献信息提供服务组合和文献全文提供服务组合。服务组合所需的知识全部存在于信念库中，而信念库则根据不同

978-1-935068-76-1 © 2011 SciRes.

的需要的建立在不同的抽象层次上，如在主题浏览服务组合中，文献与主题间关系的推理，直接在信念库中将主题作为“类”，而将文献作为“个体”来处理，而主题合并的推理服务则将主题作为“个体”来处理，以便使用相应的推理规则。

4 其他

4.1 基于 AGROVOC 的检索

AGROVOC 是 FAO 研发的农业叙词表，本课题组使用其中文版进行基于叙词表的文献检索研究，开发了基于叙词表的文献检索系统和将叙词表与 Google API 相结合的农业搜索引擎。

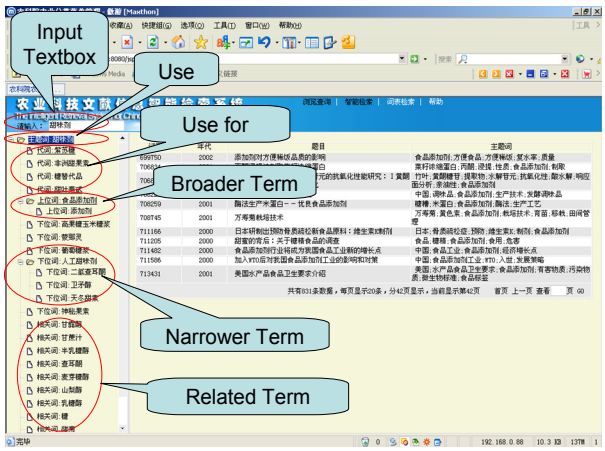


Figure 6. Search system based on AGROVOC
图 6. 基于 AGROVOC 的检索系统

Figure 7. AGROVOC based search engine
图 7. 基于 AGROVOC 的搜索引擎

基于 AGROVOC 的检索系统将叙词表中的概念及其相关概念的关系结构以 XML 格式组织在一起，在用户检索时将组织好的概念及其全部的相关概念和关系呈现给用户，用户可以在其中选择更适当的概念和词汇进行检索。检索系统界面如图 6 所示。

农业搜索引擎利用 FAO 叙词表和 Google API 建立一个农业垂直搜索引擎，引擎通过概念间的语义关系连接农业概念，利用概念树进行知识导航，将基于关键字的 Google 检索转化为概念检索，实现了用户输入词语的自动扩展和规范 [18]。基于 AGROVOC 的搜索引擎界面如图 7 所示。

4.2 顶级本体的 OWL 转换

我们尝试将开源的 OWL Full 表示的 SUMO 和 OpenCyc 顶级本体转换到 OWL DL，以便将顶级本体和使用 OWL DL 子语言构建的领域本体相结合进行推理。SUMO 顶级本体的知识元素的数量虽然远远少于 Cyc，但其中不符合 OWL DL 语法的情况比 OpenCyc v0.7.8b 版本中的情况复杂很多[19]。

从对 SUMO 和 OpenCyc 的转换过程可以看出，两个 OWL Full 表示的本体共同存在的问题是一个概念同时表示类和实例的情况。本课题组希望通过不同知识抽象层次和知识领域模型研究，从根本上解决此类问题。

4.3 元数据及著录系统

在参考 Dublin Core, 科技数据库核心元数据等标准的基础上，开展农业科技信息元数据核心元数据标准框架研究。农业科技信息元数据核心元数据标准框架包括农业科学技术信息核心元数据标准（草案）及其应用方案、扩展原则和使用指南。

元数据著录系统包括用户管理、元数据著录及浏览、元数据检索与审核以及相关编码的管理等功能。系统遵循标准的体系、内容和结构，尽量用程序自动控制与标准保持一致，如根据相关的知识库，主动检验用户著录内容的有效性[20]。系统界面如图 8 所示。

4.4 花卉学本体模型及推理系统

通过构建对 6,432 篇花卉文献中的 6,887 个不重复主题词进行归并、合并、归类和语义分析，最终共得到了花卉学物种概念、名词性概念、谓词性概念 3 种的概念集合（类），从主题词和现有概念中提取出了上下位和非等级两大类概念间关系；从现有概念间的关系中提取了子集定位和集合指向两大类函数；从现有概念和实例间关系提取出了公理 14 类；并利用 CycL 表示花卉学领域的概念、函数和公理[21]。

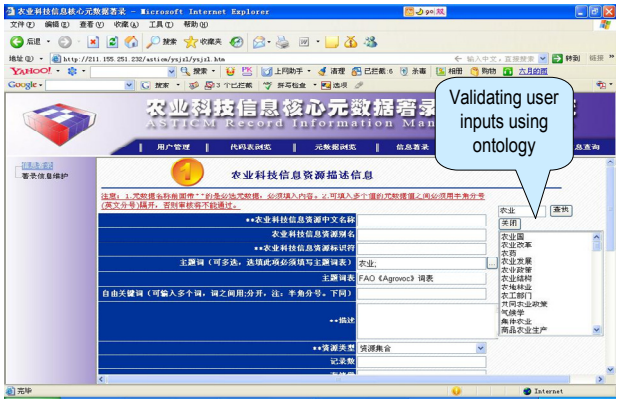


Figure 8. ASTICM Record Information Management System
图 8. 农业科技信息元数据核心元数据著录系统

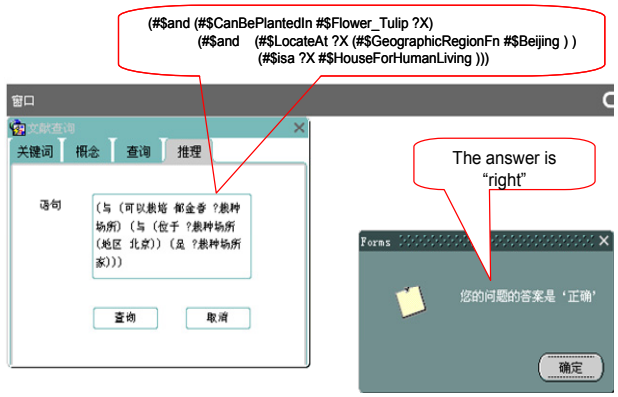


Figure 9. The asking & answering pages in FORS (Floricultural ontology retrieve system) answered the question "may tulip be cultured in Beijing domestically?"
图 9. FORS 系统回答“在北京居家可以栽培郁金香吗？”的问题输入页面和答案返回页面

花卉学文献试验性本体检索系统以构建的花卉学本体模型为基础，采用 OpenCyc 开源码项目的顶级本体结构和推理引擎，利用 Oracle 数据库应用程序实现了相应的检索功能和文献查询与知识录入界面。系统初步实现了推理判断、排除歧义、和概念检查与纠错等智能检索功能。系统推理界面如图 9 所示。

4.5 主动推荐式检索系统

主动推荐式检索系统将与用户输入的检索词相关的知识领域按照分类法的层次结构提示给用户，使用户能够对所查询的概念有一个完整的了解，并在其中选择概念的所在知识领域，系统根据用户的选择，在精确的知识领域范围内进行查询，以达到排除歧义和提高查询结果准确率的目的。概念与分类之间的关系从主题标引和分类标引中提取，抽取实验用概念 2270 条，相应的类目 164,377 条，平均每个概念对应 72 个类目。系统采用 Software AG 公司的 Tamino 纯 XML 数据库管理系统存储数据。使用 XML Schema 定义概念与分类结构之间的关系，使用标准的 XQuery 语言检索[22-25]。系统页面如图 10 所示。

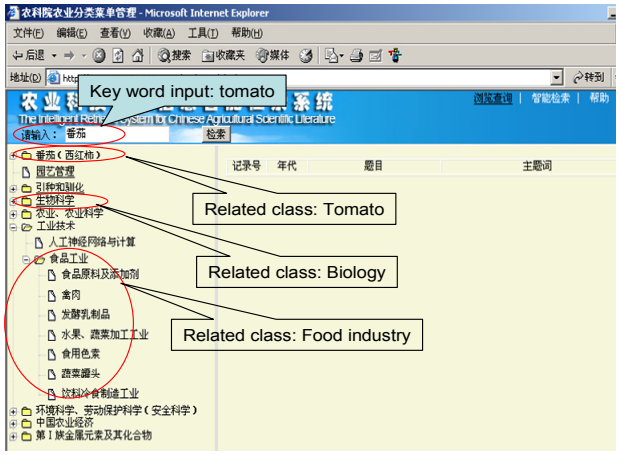


Figure 10. Searching page of initiative recommendation
图 10. 主动推荐式检索系统页面

5 结语

我们从知识本体的理论、方法、技术及应用各方面开展了广泛的探索，在应用研究中我们发现一个本体的质量，对应用实验的效果起着至关重要的作用，但是就本体的构建而言，其中仍存在理论和方法方面的问题尚未解决，如知识模型，常识性知识等，而这些问题严重制约着领域知识本体的开发与应用的发展，要开发具有实用价值的农业知识本体任重而道远。

致 谢

本文由国家科技基础条件平台建设项目“农业科学数据共享中心”；第四十批中国博士后科学基金“领域本体的构建方法与应用研究”；国家“十一五”科技支撑子专题“农村知识本体构建的共性技术研究”；国家 863 课题“农业科技信息移动智能服务技术研究”等项目资助。

References (参考文献)

[1] Tim Berners-lee. Semantic Web. http://www.w3.org/2000/Talks/1206-xml2k-tbl/ [EB/ OL] 2000.
[2] Lu Ruqian. Knowledge Engineering and Knowledge Science at the Turn of the New Century", Tsinghua University Press, 2001.P447-468 (Ch).
陆汝钤主编：世纪之交的知识工程与知识科学，清华大学出版社 2001，P447-468.
[3] Li jing, Su Xiaolu, and Qian Ping. The methodology of develop-

978-1-935068-76-1 © 2011 SciRes.

ing domain ontology. *Computer and Agriculture*, 2003. (7), 7-10 (Ch).
李景,苏晓鹭,钱平. 构建领域本体的方法[J]. 计算机与农业,2003,(7).P7-10.

[4] Li jing, Qian Ping. Distinctions and Relationships between Thesaurus and Ontology. *Journal of Library Science in China*, 2004. (1), 36-39 (Ch).
a) 李景,钱平. 叙词表与本体的区别与联系[J]. 中国图书馆学报,2004,(1). P36-39.

[5] [5] Gao Fan,Li Jing. Ontology and the Ralations between Ontology, Taxonomy and Thesaurus. *Library Theory and Practice*, 2005. (2),44-46 (Ch).
高凡,李景. Ontology 及其与分类法、主题法的关系[J]. 图书馆理论与实践,2005,(2). P44-46.

[6] Li Jing. Comparison of Important Ontology Specification Languages. *New Technology of Library and Information*, 2005. (1).1-4,8 (Ch).
李景. 主要本体表示语言的比较研究[J]. 现代图书情报技术,2005,(1). P1-4,8.

[7] Li Jing, Meng Liansheng. Comparison of Seven Approaches in Constructing Ontology. *New Technology of Library and Information*, 2004. (7),17-22 (Ch).
李景,孟连生. 构建知识本体方法体系的比较研究[J]. 现代图书情报技术,2004,(7). P17-22.

[8] Li Jing. Comparative Study of Major Ontology Editing Tools (Ⅰ). *Information Studies:Theory & Application*, 2006. (1), 109-111 (Ch).
李景. 主要本体构建工具比较研究(上)[J]. 情报理论与实践,2006,(1).P109-111.

[9] Li Jing. Comparative Study of Major Ontology Editing Tools (Ⅱ). *Information Studies: Theory & Application*, 2006. (2), 222-226 (Ch).
李景. 主要本体构建工具比较研究(下)[J]. 情报理论与实践,2006,(2). P222-226.

[10] Li Jing. Design of Bottom-level Structure of Database in Ontology-based Intelligent Retrieval System. *JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION*, 2004. (3),297-302 (Ch).
李景. 基于 ontology 智能检索系统的底层数据结构设计[J]. 情报学报,2004,(3). P297-302.

[11] Li jing, Qian Ping, and Su Xiaolu. Building Knowledge Portals Based on Ontology. *New Technology of Library and Information*, 2004. (2). 18-21 (Ch).
李景,钱平,苏晓鹭. 构建基于 Ontology 的知识门户[J]. 现代图书情报技术,2004,(2). P18-21.

[12] Li Jing. Current Trends on the Ontology Technology Standardization. *New Technology of Library and Information*, 2007, (8). 12-17 (Ch).
李景.本体技术标准化综述研究[J]. 现代图书情报技术,2007,(8). P12-17.

[13] Li Jing, Meng Xianxue, and Su Xiaolu. Development Method and Practice Research on Domain Ontology. China Agricultural Science and Technology Press. 2009 (Ch).
李景,孟宪学,苏晓路.领域本体的构建方法与应用研究. 中国农业科学技术出版社,北京, 2009.

[14] Su Xiaolu, Li Jing, and Meng Xianxue. Main Functions and Characters of LODE' s Client. *Agriculture Network Information*, 2009. (8), 63-65 (Ch).
苏晓路,李景,孟宪学. LODE 本体开发系统客户端的主要功能与特点[J]. 农业网络信息,2009,(8) , P63-65.

[15] Zhang Yu. The Study of Agricultural Scientific Information User Modeling System Based-on Ontology [D]. Graduate School of the Chinese Academy of Agricultural Sciences. 2009. (Ch)
张瑜. 基于本体的农业科技信息用户建模系统研究[D]. 中国农业科学院, 2009.

[16] Zhang Yu,Li Jing,Meng Xianxue. Summary of the Main Methods of Annotation for Network. *Library and Information Service*, 2008. (11).P20-22,68 (Ch).
张瑜，李景，孟宪学.网络标注的主要方法概述[J].图书情报工作，2008. (11).P20-22,68.

[17] Su Xiaolu, Zhang Yu, Meng Xianxue. Research on Providing Personalized Information Service By Modeling User's Knowledge Structure With Ontology. Computer & Computing Technologies in Agriculture-II (CCTA2009). 2010. 483-492.

[18] Cui Yunpeng, Tang Peng, Liu Shihong. Study of Agricultural Search Engine Based on FAO Agrovoc Ontology and Google API. Computer & Computing Technologies in Agriculture-II (CCTA2009), 2010.439-444.

[19] Su XiaoLu, Li Jing, and Meng Xianxue. Translating of Top Ontology Expression from OWL Full to OWL DL. *New Technology of Library and Information Service*, 2009. (2).39-45 (Ch)
苏晓路,李景,孟宪学. OWL Full 表示的顶级本体到 OWL DL 的转换研究[J]. 现代图书情报技术,2009,(2).P39-45.

[20] Cui Yunpeng. Study of Key Technologies of Agricultural Knowledge Management Based on Ontology. China Agricultural Science and Technology Press. 2009. (Ch).
崔运鹏. 基于本体论的农业知识管理关键技术研究.2009[D]

[21] Li Jing. Study on the Theory and Practice of Ontology and Ontology-based Agricultural Document Retrieval System--Floricultural Ontology Modeling . Graduate University of the Chinese Academy of Sciences. 2004. (Ch).
李景. 本体理论及在农业文献检索系统中的应用研究——以花卉学本体建模为例[D]. 中国科学院研究生院（文献情报中心）, 2004.

[22] Su Xiaolu, Qian Ping, and ZhaoQingling. Construction of Agricultural Scientech Navigation Knowledge Database and Intelligent Retrieval System. *JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION*, 2004. (6). 677-682(Ch).
苏晓路,钱平,赵庆龄. 农业科技信息导航知识库及其智能检索系统的构建[J]. 情报学报,2004,(6). 677-682.

[23] Su Xiaolu, Qian Ping, and Yan Yun. The Knowledge Organization in the Agricultural Intelligent Retrieval System. *New Technology of Library and Information*, 2005. (12). 34-38 (Ch).
苏晓路,钱平,颜蕴. 农业科技信息智能检索系统中的知识组织[J]. 现代图书情报技术,2005,(12). P34-38.

[24] ZHAO QingLing, QIAN Ping, and SU XiaoLu. 2006. Application of Ontology in Soil Knowledge Intelligent Retrieval System Based on Web，1st Semantic Web and Ontology(SWON2006), 3rd Web Information System and Application(WISA2006), COMPUTER SCIENCE 2006, 33(10A). 34-36.

[25] Cui XiaoYang, Qian Ping, and Su Xiaolu. Tamino Server and its application. *Agriculture Network Information*, 2004, (9). 28-30 (Ch).
崔晓阳,钱平,苏晓路. Tamino Server 及其应用[J]. 农业网络信息,2004,(9). P28-30.

A Study on Multi-echelon Centralized Spare Parts Inventory Control Model

Xiaojun YAN[1], Zhiya CHEN[2], Lijuan MAO[2], Xiaoping FANG[2]

[1]*Business School, Central South University, Changsha, China, 410083*

[2]*School of Traffic and Transportation Engineering, Central South University, Changsha, China, 410075*

Abstract: Building a multi-echelon centralized spare parts inventory control model by improving the traditional single-stage decentralized spare parts inventory control model, this paper makes a comparison between the two models. The findings of the study, with the same probability of stock shortage, the centralized control model can reduce inventory holdings of enterprises so as to lower down their inventory cost and reduce capital occupation. A empirical research of five production plant spare parts inventory control problem show that monthly inventory and monthly inventory costs decreased by 40.13% and 13.62% at the 0.15% probability of shortage by the multi-echelon centralized inventory control model, comparing to a single-stage decentralized inventory control model.

Keywords: multi-echelon; centralized management; spare inventory; shortage risk

The equipments of the branches of large smelt plant are generally similar, and interchangeable parts and standard parts are available for repairing. According to the location, the inventory control for the parts can be described as two kinds which are decentralization and centralization, and monopole inventory and multilevel inventory according to the location level. Because of the high price, large amount and sorts of the equipment parts, huge amounts of money and storehouses are necessary. Take Chalco (Aluminum Corporation of China Limited) for example, in 2005, the total equipment inventory take over 10 ha. with about 25.33 billion RMB. As a result, the research for the new inventory model is the best way to reduce the inventory cost and improve the inventory capacity.

1 Replacements Parts Inventory Research

Inventory control model can be divided into monopole control and multilevel control. In monopole control, the traditional inventory theory is available while the final result of the model is the optimal inventory of individual enterprise or department. The multilevel control is mainly based on the research for Supply Chain theory and started from the whole supply chain to get the overall optimum result. Qu Shoufeng[1] established the multilevel inventory control model which permits the item shortage which starts from the multilevel inventory control system lead by DC(Distribution Center) with several vendors and retailers to confirm the inventory policy of DC as a goal. Chow Zhiping[2] established the centralization-multilevel inventory control model, describing the content and the implementation methodology.

The research for the replacements parts inventory is divided into the following aspects: parts category, parts demand forecast, parts inventory control strategy and model, parts inventory control methods and item shortage and scraping. The frequently-used ABC taxonomy divides the spare parts into three parts which are Very important, important and common according to the value[3]. Another taxonomy is based on the velocity of circulation of the spare parts and the occupation of funds to divides into fast flow parts and low slow parts[4,5]. Cui Nanfang[6] has researched the slow flow inventory model with lifetime function forecast for the parts demands. Ghobbar[7] has researched the parts demands forecast model by survey the results of 13 forecast methods; and replacement parts control strategy is focused on the cost, system restriction and multilevel inventory system[8-11]. Cobbaert and Qudheusden[12] has established the EOQ improvement model analyzing the unexpected scrap risk of the parts inventory, discussing the influence of the parts scrap under the situation of full stock and shortage.

This article establishes the multilevel inventory model with stock shortage risk controllable, practicing the multilevel inventory control model in replacements parts inventory control strategy with inventory research, and compares the monopole safe inventory capacity and cost with multilevel ones under the same item shortage probability.

2 Two Replacement Parts Inventory Control Modes

The inventory management of most smelt plants are traditional monopole inventory control mode which is the individual plant purchase the replacement parts from retailers and kept their own stock. Because of the particularity and uncertainty of the equipment parts, the individual enterprise has a large scale of inventory with huge inventory cost in case of shortage. Basing on this, many large enterprises attempt to find a new inventory strategy. The following content is the description for monopole and multilevel inventory control modes.

978-1-935068-76-1

2.1 Distributed Monopole Inventory Control Mode

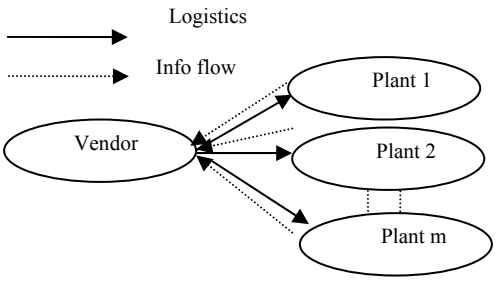


Figure 1. Distributed monopole inventory

In Figure 1, considering that a kind of part is provided by the only vendor to m plants, the manufacturer takes the inventory and fixed order cycle T strategy, and other variable quantity is fixed as follow:

D_{ij} is the demand in circle i for spare part of plant j and obeys the poisson distribution $p(\lambda)$, and the average is μ_{Dj}. R_{ij} is the demand in circle i for spare parts of plant j and s_j is the re-order point for spare parts of plant j which is the safety inventory. c is the unit cost of the spare parts and K_j is the fixed stock cost for plant j, and h is the unit inventory cost for spare parts, and the unit inventory cost in every plant is similar, and r is the unit transportation cost (yuan/unit item•km). d_{oj} is the distance from vendor to palnt j, and L_{ij} is the pre-order period for plant j who has purchased for i times and obeys the poisson distribution, the average is μ_{L_j}, and n is the annual order cycle with m is the plant quantity.

The inventory of every plant is as figure 2:

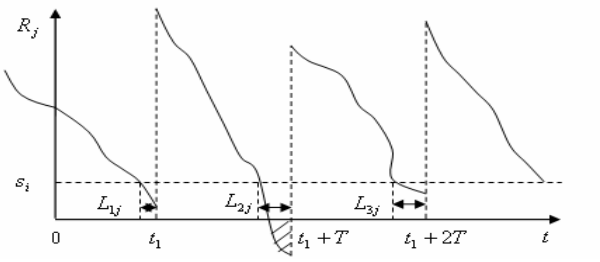


Figure 2. Inventory of plant J

The order cost of plant j in cycle i is $OC_j = K_j D_{ij}/R_{ij} + cD_{ij}$; the inventory turnover of plant j in cycle i is $R_{ij}/2$. The safety inventory of plant j is $s_j - \mu_{D_j}\mu_{L_j}$, and the inventory cost for plant j in cycle i is $SC_j = h\left(R_{ij}/2 + s_j - \mu_{D_j}\mu_{L_j}\right)$, and the transportation cost for plant j in cycle i is $HC = rD_{ij}d_{oj}$ and the annual cost for the plant is

$$Cf = \sum_{i=1}^{n}\sum_{j=1}^{m}\left[K_j D_{ij}/R_{ij} + cD_{ij} + h\left(R_{ij}/2 + s_j - \mu_{L_j}\mu_{D_j}\right) + rD_{ij}d_{oj}\right] \quad \text{(Formula 1)}$$

We can concluded from the total cost function that with the fixed demand and safety inventory, with the increase of order, the inventory cost increases as well while the order cost reduces. So it is necessary to find a exact order quantity to reduce the total cost which is cost function order demand to get the 0 first-order derivative, and the best order quantity can be calculated to be $R_{ij}^* = \sqrt{2K_j D_{ij}/h}$. In total cost function, the safety inventory can be controlled by stockouts function and get a reasonable safety inventory.

2.2 Multilevel Centralized Inventory Control for Spare Parts

The multilevel centralized inventory control mode is the headquarter inventory control center with several area hub stock which is increased at the base of monopoly inventory control and make it a multilevel centralized inventory control mode from the original monopoly one. The precondition of the mode is information sharing platform and all plants, hub stock and headquarter stock have shared the same information platform and every hub must grasp the demand of the spare parts for the plants and submit the purchase plan to headquarter inventory center. The hub makes the order with vendors according to the spare parts information and vendors will deliver the items to the hub stock directly with few parts are delivered to the headquarter inventory hub to be the safety inventory. The hub will deliver the items to the plants in this area. For the research convenience, it is supposed that the inventory of headquarter inventory hub in proportion with the safety inventory and the proportion is concerned with the stock size and quantity.

In figure 3, one part is provided to the hub and headquarter stock by the only vendor, and the hub will deliver the items to the plants. The area stock hub and headquarter stock held the inventory and take the fixed cycle order strategy T, and the variable quantity is supposed as follow: D_i is the demand for spare parts in cycle i for all centralized stocks and obeys the poisson distribution, the average is $\sum_{i=1}^{n}\mu_{D_i}$, D_i^q is the demand of stock p in cycle i and obeys the poisson distribution, and the average is $\sum_{i=1}^{n}\mu_{D_i^q}$, $D_i = \sum_{q=1}^{p} D_i^p$.

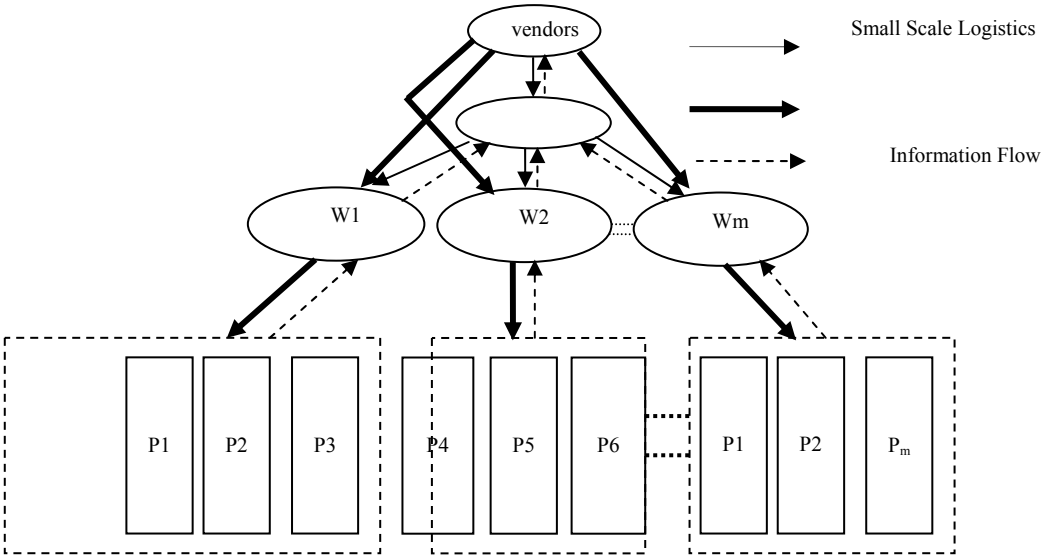


Figure 3. Multi-echelon Centralized Inventory Control Model

The demand for spare parts in cycle i for plant j which is in the charge of area center stock q is D_{ij}^{q} and obeys the poisson distribution, and the average is $\sum_{i=1}^{n}\mu_{D_{ij}^{q}}$, $D_{\mathrm{i}}=\sum_{q=1}^{p}D_{i}^{q}=\sum_{q=1}^{p}\sum_{j}^{m}D_{ij}^{q}$ (p is the quantity for all hubs), R_i is the spare parts order quantity for headquarter stock hub in cycle i ,and s is the sum of the safety inventory for all hubs according to the demand which is made by the headquarter. C is the unit cost, K is the fixed cost for each restock, and h is the unit inventory cost, L is the order lead times and obeys the poisson distribution, the average is μ_L, and r is the unit transportation cost (Yuan/ unit item • distance), d_q is the distance from hub q to plant j and the n is the total order cycles while m is the plant quantity.

The total cost for a enterprise is formed with inventory cost, order cost and transportation cost. The order cost in cycle i is $OC=K\,D_i/R_i+\mathrm{c}D_i=K\sum_{q=1}^{p}\sum_{j=1}^{m}D_{ij}^{q}\Big/R_i+c\sum_{q=1}^{p}\sum_{j}^{m}D_{ij}^{q}$ and the safety inventory is $s-\sum_{j=1}^{m}\mu_{D_j}\mu_L$. The inventory circular flow is $R_i/2$, and the total inventory cost in cycle i is $HC=h\left(R_i/2+s-\sum_{j=1}^{m}\mu_{D_j}\mu_L\right)$; the transportation cost in cycle i is $SC=\sum_{q=1}^{p}d_q D_i^q r+\sum_{q=1}^{p}\sum_{j=1}^{m}D_{ij}^{p}d_{qj}r$. There are some spare parts in headquarter inventory center, and the inventory cost is $HC_{总}=ahs$ .Because the out stock probability is small, headquarter seldom delivers the spare parts to plants, so the transportation cost in headquarter can be omitted. So the total annual inventory cost for the enterprise is :

$$C=\sum_{i=1}^{n}\left[K\sum_{q=1}^{p}\sum_{j=1}^{m}D_{ij}^{p}/R_i+c\sum_{q=1}^{p}\sum_{j=1}^{m}D_{ij}^{p}+h\left(R_i\Big/2+s-\sum_{j=1}^{m}\mu_{D_j}\mu_L\right)+\sum_{q=1}^{p}d_q D_i^q r+\sum_{q=1}^{p}\sum_{j=1}^{m}D_{ij}^{p}d_{qj}r+ahs\right] \quad \text{(Formula 2)}$$

In total cost function, when the total demand is fixed , inventory cost and order cost and transportation cost is changed with the change of order quantity. So it is necessary to find a exact order quantity to make the lowest cost. With the same method, the best order quantity in cycle i for area stock hub is as follow: $R_{\mathrm{i}}^{*}=\sqrt{2K\sum_{q=1}^{p}\sum_{j=1}^{m}D_{ij}^{p}\Big/h}$. By out stock function control, the safety inventory is reasonable.

3 Comparison of Two Inventory Control Modes

In multilevel inventory control, because of the centralized order, the safety inventory of every plant reduces. It can be calculated that with the same out stock ratio, all area center stock will held the safety inventory sum and the safety inventory s_i for every plant according to the poisson distribution.

978-1-935068-76-1 © 2011 SciRes.

It is supposed that the error probability for each spare part is p, the probability for K parts is $P\{X=M\} = C_n^M p^M q^{n-M}$, and $P\{X>M\} = 1-\sum_{k=0}^{M} C_n^M p^M q^{n-M} = 1-\sum_{k=0}^{M} \lambda^M e^{-\lambda}/M!$ ($\lambda = np$) is the out stock ratio with the inventory M. when P is similar, n is getting bigger while M smaller. So, by centralized controlling the spare part inventory, the safety inventory sum is smaller than the original safety inventory sum which is under the monopoly inventory control which is $s < \sum_{j=1}^{m} s_j$. The following formulas are the difference between the two inventory control strategies and the inventory cost, which is $\Delta C = Cf - C$:

$$\Delta C = \sum_{i=1}^{n}\left[\sum_{j=1}^{m} K_j D_{ij}/R_{ij} - K\sum_{q=1}^{p}\sum_{j=1}^{m} D_{ij}^q/R_i + r\sum_{j=1}^{m}\left(D_{ij}d_{oj} - \sum_{q=1}^{p} D_{ij}^p d_{qj}\right) - \sum_{q=1}^{p} d_q D_i^q r\right]$$

$$+\sum_{i=1}^{n} h\left(\sum_{j=1}^{m} R_{ij}/2 - R_i/2 + \sum_{j=1}^{m} s_j - s + \sum_{j=1}^{m}\mu_{D_j}\mu_L - \sum_{j=1}^{m}\mu_{D_j}\mu_{L_j} - as\right) \quad \text{(Formula 3)}$$

It can be concluded from above: ΔC is determined by order quantity and safety inventory with the fixed demand, order lead time and path, etc, and the order quantity is calculated by the best order batch. The safety inventory is determined by the out stock ratio. By centralized control to area centre stock with the same out stock ratio, when the size of centralized control gets bigger, so does $\sum_{j=1}^{m} s_j - s$ to reflect the advantage of centralized inventory control.

4 Case

The monthly demand for the replacement parts of all branch plants of a corporation obeys Poisson distribution, and the average is 7000 pcs with 0.5% error probability, then λ=35. The total demand of the five branch plants obeys the Poisson distribution as well and with the same error probability and 0.15% stockouts probability, the λ and the safety inventory s_i is as follow (for convenience, the only central stock is supposed):

Table 1. The λ and the safety inventory s_i

	Plant1	Plant2	Plant3	Plant4	Plant5
Monthly demand/piece	1200	1300	1500	1000	2000
λ	6	6.5	7.5	5	10
Safety inventory/piece	14	15	16	12	20
Regular order cost/Yuan	800	800	1000	700	1500
Best order quantity/piece	219	228	274	187	387
Distance to vendor/km	500	300	200	200	100
Distance to center/km	300	100	50	50	100
Averery order date/day	3	3	3	3	3

(r=0.02yuan/pcs·km, h=40yuan/pcs, k=1600yuan, R=748pcs, μ_L=3days, d=200kilos, a=10%)

If the parts inventory is managed centrally, it can be calculated that the area central safety inventory quantity is s=54, and safety inventory in home office is 6 while the total safety inventory of 5 branches is 77. According to the formula 3, put the above numbers in and monthly cost difference will be ΔC = 12079.

5 Results

The research for replacement parts in this article is combined with centralized inventory control theory of supply chain, considering the particularity of the parts inventory, establishing the centralized inventory control model and comparing with the monopole replacement parts inventory control. When the stockouts probability is 0.15%, the difference of the two modes is 291 pieces, and the inventory reduces 40.18%, and inventory cost reduces ￥12079.00. The total cost reduces 13.62% comparing with the monopole inventory control mode.

In addition, the model which the article has researched has a certain popularity, and some expensive and slow flow raw materials can take this inventory control model. Otherwise, the hypothetical lead time in the article is a random variable, which makes the model more practical to the actual situation. The model has a certain guide to the large production enterprises.

References

[1] Qu shoufeng, Research for multilevel stockouts control model under cyclicity demand[J]. Dongnan University references, 2007(28): 879-882.
[2] Zhou zhiping, Gao peng, Distributed inventory control discussion basing on supply chain management[J].Chinese management science, 2004.10(12):437-441.
[3] Cui rongchun, “4-ABC” Orthogonal Complete Control Act for spare parts[J]. Management engineering references, 2008. 4(22): 129-131.
[4] Xiong junxing, Xia fangchen, Tu haining , ABC disaggregated model for spare parts basing on BP neural network[J].Machine design and manufacture, 2008(2):215-217.
[5] Guo zhimin, Yan hongsen, Chen shihua, Wang xuewei, Research for spare parts inventory control methods[J].PC integrated manufacturing- CIMS.2003.11(9):100-104.
[6] Li jin, Cui nanfang, Ding liuming, automobile service parts

inventory construction basing on category management. Logistics technology 2006. 8: 73-75.
[7] Cui nanfang, Ding liuming, slow flow inventory model basing on spare parts life function. System project. 2006. 9(24): 24-31.
[8] Adel A. Ghobbar, Chris H. Friend.Sources of intermittent demand for aircraft spare parts within airline operations[J]. Journal of Air Transport Management, 2002, 8: 221-231.
[9] Si shubin, Sun shudong, Cai zhiqiang, the repairing parts inventory control mode and research with the supply cost restriction. Xibei University references, 2006. 5(24): 662-665
[10] Si shubin, Jia dapeng, Sun shudong, Wang ning, the multi-single repairing parts inventory control mode and research with the same service restriction. China machinery project express. 2007. 12(18):2844-2847.
[11] Dai gengxin, Zhang hui, multi-variety service inventory control strategy for spare parts with commercial restriction. Qingdao University references, 2007.4(22):85-87.
[12] Wang naichao, Kang rui, spare parts inventory optimized model and analyze basing on multi-restriction. Military references, 2009. 2(30):247-251.
[13] Sherbrooke, C.METRIC. a muliti-echelon technique for recoverable item control [J]. Operations Research, 1998.16:122-141.

978-1-935068-76-1

Research on Information Quality Evaluation Index System Based on User Experience

Bing LIU[1,2], Shuang LU[1], Jinru LI[1]

[1]*Management School of Tianjin Normal University, Tianjin, China, 300387*

[2]*School of Information Management, Wuhan University, Wuhan, China, 430072*

Abstract: Information quality determines the information value in social and economic activities, and customer satisfaction is the basis of the assessment in information quality. In the network, the experience in information exchange process directly affects the evaluation of information quality. The article first analyzes the basic content of user information experience and the internal relationship between user information experience and information quality evaluation. On that basis, it formed information quality evaluation system based on user experience, and analysis the indexes of the evaluation system further.

Keywords: user experience; information quality; information quality evaluation

1 Background and Issues

Information Quality (IQ) is the applicability of information for information consumer[1], is the degree of satisfaction for information users[2]. In the era of knowledge economy, and with the development of information technology and the accelerated process of social information, information quality decides the value of information in social and economic activities, and becomes a critical factor to the social development. Thus, it is an important issue for information management to establish a scientific and reasonable evaluation system so as to promote and enhance the information quality to meet the needs of users.

Domestic and international researches on the evaluation of information quality mainly focus on two aspects. One is based on data, information products and information flow. Lesca H. & Lesca E. (1995)[3], Wang R.Y. & Strong D.M. (1996)[1], Bovee M. et al. (2003)[4], Mouzhi G. & Markus Helfer (2007) [5], respectively researched the evaluation of information quality from the aspects of breadth, depth, and quantity etc, and constructed a system of information quality evaluation index from the two levels of product and process. Wang

Kan-chang & Gao Jianmin (2006) established the Information Quality Maturity Model (IQMM), which was referenced the basic concepts of software quality maturity evaluation, and form the quality evaluation index system on the same basis[6]. But Cha Xianjin & Chen Hong (2010) constructed an information resources quality assessment index system from multiple layers[7].

Another is based on user evaluation system. Su Qiang and Liang Bing (2000) constructed an information quality assessment system which is on the purpose of 'make users feel satisfied'[8]. Based on the impact of IQ, Naumann F. & Rolker C. (2000) constructed IQ evaluation dimensions from the aspects of subjective, objective and process: the user perception, the information itself and information access process[9]. Helfert M. (2001) combined semiotics and quality management together, and constructed information quality evaluation system. At the same time, it highlights the information process and information users' pragmatic level indicators[10]; In order to meet and exceed customer expectations, Kahn B. et al. (2002) constructed two-dimensional conceptual model of IQ which reflected the point of product and service quality[11].

In the ubiquitous information society, the Internet has formed information network structure, which has offered an opened and multi-dimensional space. Internet naturally and deeply integrated into people's daily life and work. 'Interaction is the essence of the Internet'. In this environment, a new generation of digital life and network services has deeply changed the users' information concepts and way of information consumption. With the increasing awareness of information personalization and the grading of participation, to receive, product and transfer of information has integrated together. The user is no longer a passive recipient of information services, but positively and proactively participates in the information process, and feels the changes in scenes and exchanges information, which gradually highlighted the dominance of users.

In this process, the user's active participation and information exchange behavior is perceived through the information implementation and experience. Through interaction and experience, information users will have a sense of satisfaction in the information process, this kind of experience, perception and satisfaction of the users will affect the quality of objective evaluation. In the environment of network, users are constantly require the im-

This paper is one of phasic fruits of General Project of China National Philosophy and Social Science Fund 'Optimize the user experience and perception and establish comprehensive evaluation index of information quality' (No. 10BTQ007) and China Postdoctoral Science Fund 'Research on the Public Interest Information Quality Evaluation from the Perspective of Users' (No. 20100480878)

provement of information quality, whether it can get an objective, comprehensive and systematic evaluation from the users' experience and interaction will affect the mutual trust between information users and service providers. Therefore, as long as we study the information quality on the perspective of users and build up a system, can we improve the information quality and truly meet customer needs and expectations.

2 Users' Information Experience in the Network Environment

The essential of information exchange in network environment is making people and system, and interpersonal communication come true through internet. This is a process that users and system, users and service interact and exchange information and perception of feedback with each other. In this process, users can get the information they need at any time and anywhere, however, this information behavior still can not get rid of users' way of thinking and psychological motivation[12]. Users' cognitive, emotional, intentional and other psychological factors determine their information behavior and information consumption process, and affect users' attitude towards products, systems, service attitude, and the users' evaluation to the quality of information, and satisfaction and other aspects[13].

User experience (User Experience, UE) is a purely subjective mental experience formed in the process of getting and using information[14]. User experience highlighted and reflected the emotion and idea that formed in the user interface (including products, systems and services) during the interaction process. In the ubiquitous information society, the network interaction is not only a process of forming and developing relationships, but also a process of user emotional experience. Through interaction and experience, users of information obtained a sense of satisfaction in the dynamic process.

User experience focuses on the content of user and information products, information systems and information services in the behavior of interaction, and focuses on the cognition, emotion and attitude in this process and behavior. User experience is the comprehensive reflection of instinct, behavior and rethinking in the process of information acquisition and use. Firstly, user experience fully reflects the complex and subjective mental experience behind the user information behavior during the process of the users' interaction with the system. Secondly, as purely subjective psychological feelings of information users, user information experience has showed out the objective psychological feeling in the behavior of information interaction, and this has a direct impact on the objective evaluation of the interaction. Thirdly, user experience can objectively reflect the real information needs of users, and it can reflect the information needs that in potential and the ones that haven't expressed clearly. Finally, user experience is not only the basis for understanding user behavior, but also for analysis and grasps the further behavior of the users. To some extent, user experience is not connecting with the product and service value itself.

Based on the analysis of people's psychological needs from psychologist Murray (Murmy), user experience can be divided into five categories: sensory experience, interactive experience, emotional experience, browsing experience and trust experience[15]. In the ubiquitous information society, with the improvement of the complexity of the network environment, customer's behavior of information gathering, browsing, retrieval and use intertwined, the interactive process of users and systems, and servants is converted to "real-time process. " In this process, user experience is not only a cross-section of information, but also a continuum. Therefore, based on the holistic perspective, according to the hierarchy of psychological perception in the information behavior and information exchange, the network environment can be divided into three levels, functional experience, technical experience and aesthetic experience. Among them, the functional experience describes the perception and feeling of users to the information that "can help users complete tasks", and is the basic experience that users to the utility of information product; technology experience describes the experience and psychological reflecting of users to the information that "can help users complete task efficiently"; aesthetic experience describes the feelings and attitudes of information users to the information that "can make users relaxed and complete the task."[16] The last two-level reflect the perception that the users for information products, and the "usefulness" of information systems.

In network environment, with the development of users' information needs and expectations, technology experience, aesthetic experience becomes the leading user information experience. Although the functional experience is still the primarily psychological feelings of evaluating the information products and systems, and is the basis of user behavior. But with the multiple growth of information quantity in the network, and diversification means of access to information, the initial impression and feeling directly determine and affect its evaluation toward information quality and leading the course of evaluation in the process of information interaction. Technical experience is the feelings and psychology formed in using the system in the process of interaction. Because of products, systems interface and design, the aesthetic experience is the pleasant emotional experience and response that formed in the process of interaction, and is the psychological feelings that it gets recognition respect on this basis. This two aspects, although is a purely subjective experience, it is the key source of feelings that users get initial impression in information interaction.

3 Dynamic Relationship between User Information Experience and Information Quality Assessment

978-1-935068-76-1 © 2011 SciRes.

Service marketing research shows that the evaluation which customers toward the service quality is the result of the comparison between expected service and perceived service. Improving customers' perceived service quality is the key factor to improve service quality, so as to match customers expected service and perceived service, and to minimize the gap between the two[17].

The analysis of user experience in network environment shows that information interaction and experience is a process of information consumption. In this process, the psychological feeling that users obtained in information experience is the result of the comparison between user information perception service and information expectation service, and this directly affect the evaluation process and the final product. Therefore, the users' evaluation to the information quality is the result of the comparison that can be interpreted as expectations of access to information and the information perceived in interaction.

- In network environment, the changing role of information users and frequent information exchange makes it an experience of a good information flow every time. In this process of flow experience, through interaction with the perception, the users will completely attracted by the information behavior and information process, they can get and use information in high-efficiency, and physical and mental pleasure, and may achieve a the feeling of pleasure and the peak of working condition[16]. In this state, the user's technical experience and aesthetic experience will directly affect users' evaluation to information service and products, and also affect users' information satisfaction and information quality evaluation.
- In True Moment theory, Swedish scholar Norman (Nomann R.) indicates that at a certain time and place (True Moment) the service experience formed had significant impact. In network environment, the user experience is the psychological perception formed through the 'true moment' that with system, products, service contact and behavior of interaction, and this kind of feeling will affect users' evaluation to the products, systems and services. The numerous true moments will form the reputation and image of information products and information services.
- In network environment, the process of user information interaction is a learning process of innovative and self-construction in a hidden way Therefore; the user information experience is a way of learning. Psychologists Bruner pointed out that learning including acquisition, transformation and evaluation the three links. In the process of information exchange, the user can take advantage of intelligent, interactive knowledge to access and process system, and progress effective feedback of the customized and interactive service that received according to the experience and feelings, and adjust its information behavior so as to achieve the target of knowledge searching, discovery and mining, and complete the evaluation to information quality in the process of experience.
- From the angle of the system and based on user experience and perception, and the service received by interactive, the quality of information products, services and experience is although subjective and emotional, it is the objective reflection of the user information needs and expectation. To manage information quality on this basis, it can not only verify the effectiveness of information exchange, enhance product and service availability and meet the user's information needs and expectations, but also establish an organic link between users and products. By improving the experience of optimizing products, systems and services, and channel expansion, it can lead to the new user experience and perception, and stimulate the user's creativity. It will help forming a virtuous spiral trend of user information quality.

It can be seen from the above analysis, in network environment, there exist a positive relationship between user information experience and information quality, they affect each other dynamically, effects, and carry spiraling trend. Only improve the degree of matches between user information expectation and perception, and try to narrow the gap between the two that we can improve the user evaluation to the information quality.

4 Information Quality Evaluation Index System Based on User Experience

4.1 Information Quality Evaluation System Based on User Experience

In network environment, this system including the objective evaluation to the content features of information, and including the continuous feelings and subjective evaluation, that is the comparison on the basis of user perception, and experience to the giving and receiving service, perception of system interaction results of experience. Thus, it can be seen on the point of views of users that the comprehensive evaluation is both objective and emotional.

On the basis of various types of information available both domestic and international, and considering the content of information quality and the forming of user information experience, it can built the quality assessment system in dimension layer and index layer in the network (Figure 1). Among them, the dimension layer is based on the perspective of information quality evaluation, and the index layer is from a different perspective of user experience gained with detailed elements of operational indicators.

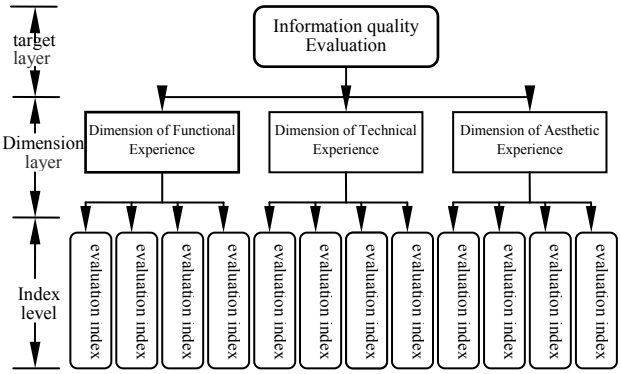


Figure 1. Information Quality Evaluation system based on user experience

According to the user experience in the network, information quality evaluation that based on user experience including three dimensions: functional experience, technical experience and aesthetic dimensions. Among them, functional experience is the perception of the users selected to the information that 'can help users complete tasks', and it is the interactive process of users behavior with the system. It is the subjective elements of perception and evaluation to the basic functions and content features of objective indicators; technical experience is the perception selected to 'help users efficiently complete the task', and it is users' interaction with information systems in which it can indicate the technical capabilities, technical indicators and evaluation of information services; aesthetic experience is the perception selected to 'make users relaxed and complete the task', and it is a purely emotional evaluation that the users felt about products, systems, friendly services and visual perception.

The three dimensions of quality evaluation system above, although have different contents and focuses, they have a strong internal correlation. Functional experience of which is the basic dimension, while technical experience and aesthetic experience is the dominant one.

Although the value of information has relatively objective evaluation criteria, the evaluation of basic functions and content features are gained from user experience with the development of interaction in network environment. To some extent, this is an evaluation of results, although it is subjective, it is a relatively objective evaluation with true features obtained from the user point of view. This dimension is the basic view and basis that user to information quality evaluation, without the functional experience, information quality evaluation system is meaningless.

In contrast, technical experience and aesthetic experience have the characteristics of the process of evaluating. Along with changes of information environment, and the development of technology, user needs and expectations are dynamically complex, users of information behavior and feelings in the process of the experience will affect the final evaluation of information quality. The good experience in process will compensate the disadvantage of the experience in result. In turn, the good experience in result will eliminate the dissatisfaction in process. Only makes the experience and feelings in the process as reference and criteria that it may embody the true evaluation of users. When the perception in process is less than expectations, even the functional experience can reach user requirements, that user evaluation to information quality may not very high; and when the perception in process is greater than the user expects, the user will be satisfied or produce the feelings of euphoria even functional experience is far from user requirements, that user evaluation of information quality will not be very low, even higher than expected.

4.2 The Indicators of Information Quality Based on User Experience

It can be seen from the above analysis, in network environment, the three constituent dimensions of information quality evaluation system which is based on user experience information are of subjective and emotional. According to the above dimensions of perspective, and the analysis of experience and feelings in information interaction. It may get the specific indicators of information quality access (See Table 1).

Table 1 The indicators of information quality based on user experience

Dimension of Functional Experience		Dimension of Technical Experience		Dimension of Aesthetic Experience	
Indicators	Description	Indicators	Description	Indicators	Description
(1) objectivity	content without personal bias	(1) security	privacy protection in process of access to information, virus restriction	(1) simplicity	interface, content and character setting clear and concise
(2) accuracy	correct information express, without mistakes	(2) intelligibility	information symbols must be able to understand and easy to understand	(2) interaction	friendly and convenient system and user interaction
(3) integrity	express a complete thought, describing one thing	(3) navigational	clear path, simple search	(3) proximity	column way, some content match user habits

978-1-935068-76-1 © 2011 SciRes.

(4) related	information units offered to customers must be correlated with each other	(4) timeliness	feedback and response speed, rapid response capability	(4) personalization	unique methods of operation, retrieval approach and information statements
(5) applicability	true value of information ,useful and helpful	(5) easiness	operation difficulty and learning difficulty	(5) extension	provide more information to guide users
(6) timeliness	speed of information updating is high, can be get in a short time	(6) fluency	smooth interaction, system response capability	(6) aesthetic	beautiful and generous interface settings
(7) right amount	moderate amount of information non-redundant	(7) convenience	provide convenient access to information channels and methods		

5 Conclusions

Along with the enhancement of personal awareness and the increase of participation in the process of information, customer satisfaction has become the basis for information quality evaluation. User experience in the process of information exchange, experience and the emotional reflection has positive relationship with information quality evaluation. The functional experience, technical experience and aesthetic experience in information interaction directly affect the information quality evaluation; information quality evaluation indication system formed on this angle can objectively reflect the users' experience to information quality.

Introducing the users' cognitive psychology law to the study of information quality evaluation is constructive in exploring information quality in the network. This article is only in the theoretical level study the user experience, and its internal relations with information quality, and is based on information quality evaluation system. In future research, it will study and verify the system on this theory research and then its findings will be more practical and help information institutions concern about the user experience in their products or services, and optimize the user experience and feelings, and even improve the evaluation results, and ultimately enhance customer satisfaction.

References

[1] Wang, R.Y. and Strong, D.M. Beyond accuracy: what data quality means to data consumers, Journal of Management Information Systems 12, (4) 1996, pp. 5-34.

[2] Huang K. T., Lee Y. W. & Wang R.Y. Quality Information and Knowledge Management. Publisher: Prentice Hall, 1999, pp. 271-350.

[3] Lesca H. & Lesca E. Gestion de l'information: qualité de l'information et performances de l'entrerpise. LITEC, Les essentiels de la gestion., 1995.

[4] Bovee M., Srivastava R.P., Mak B. A conceptual framework and belief-function approach to assessing overall information quality, International journal of intelligent systems, vol. 18, no 1, 2003,pp. 51-74.

[5] Mouzhi G. & Markus Helfert. A Review of Information Quality[J]. Proceedings of the IET China-Ireland International Conference on Information and Communications Technologies. 2007,(2): 951-958.

[6] Wang Kanchang, Gao Jianmin. Analysis on Enterprise Information Quality Status and Trends. Enterprise Management and Information Technology, 2006, (3):1-5.

[7] Cha Xianjin & Chen Minghong. Information resource quality assessment study. Journal of Library Science, 2010,136 (3): 46-55

[8] Su Qiang & Liang Bing. Information quality and its evaluation index. Computer Systems & Applications, 2000 (7):63-65.

[9] Naumann F. & Rolker C. Assessment Methods for Information Quality Criteria, Proceedings of the fifth International Conference on Information Quality. 2000.

[10] Helfert M. Managing and Measuring Data Quality in Data Warehousing, Proceedings of the World Multiconference on Systemics, Cybernetics and Informatics. 2001.

[11] Knight S. & Burn J. Developing a Framework for Assessing Information Quality on the World Wide Web[J].Informing Science Journal.2005(8):160-172.

[12] Qiao Huan. Information Behavior [M]. Beijing: Beijing Normal University Press, 2010:242.

[13] Taylor S.& Todd P. L Assessing IT Usage the Role of Prior Experience[J]. Management Information Systems Quarterly.1995, (1).

[14] Tullis, T. & Albert, B. user experience metrics. Zhourong Gang, etc. Translation [M]. Beijing: China Machine Press,2009:38-51.

[15] Yang Yixiang. Next stop: the user experience [M]. Beijing: China Development Press, 2010:1-30.

[16] Li Guihua. Information Service Design and Management [M]. Beijing: Tsinghua University Press, Beijing Jiaotong University Press,2009:124-125.

[17] Christian Gronroos. A service quality model and its marketing implications. European journal of marketing.1993:36-43.

Scientific Research

A Research Framework of Web Search Engine Usage Mining

Jimin WANG[1], Mingzi LILEI[1], Tao MENG[2]

[1]*Department of Information Management, Peking University, Beijing, China, 100871*

[2]*School of Electronics Engineering and Computer Science, Peking University, Beijing, China, 100871*

Email: wjm@pku.edu.cn

Abstract: The log files of search engines record the interactive procedure between users and the system completely. Mining the logs can help us to discover the characteristics of user behaviors and to improve the performance of search systems. This paper gives a framework on Web search engine usage mining, which includes the choice of data collections, the methods of data preprocessing, and an analysis and comparison of search behaviors from different countries. We also explore its applications on improving the effectiveness and efficiency of search engines.

Keywords: search engine; user log; web usage mining; user behaviors

1 Introduction

Search engine is an application software system used in the Web. It crawls Web information with certain tactics, processes and organizes them to provide users with information service [1]. Up to the year of 2003, there were about 3200 Web search engines of various types in 211 countries [2]. With the rapid growth of Web information, a variety of integrated large-scale or subject-oriented small search engines are continuing to increase.

The interaction process between users and search engines can be summarized as follows: the user enter a query string in the query box; after the analysis of the system, a few hundred or even thousands of records would be gained. A certain number of records (such as 10 ranked URL) form a query results page, in which each record represents a web page (a document) entry, which contains the title of the document, the location on the Web (i.e., Uniform Resource Locator, URL), the abstract of the web content and other information. Users can thus judge whether the record contains the content they are interested in, and decide whether to click on the URL and browse in detail.

Search engine logs record all the information of the interaction between the users and the systems. The forms of log of different search engines are different slightly, but generally speaking, they all include the visiting time, the IP address of the user, the queries they input, the URL they click, the clicking time and the serial number of the clicked URL. Those information are usually stored in some files with a certain format.

Search engine's log mining is a kind of the Web usage mining, which extracts meaningful modes from the user's query records, such as: how the searchers uses Web search engine; what information the user search on the Web; the searching behavior characteristics, rules and evolution trend of certain groups and individual user; the similarities and differences in the behavior of users from different areas or with differentiated themes; and how to use log analysis to improve searching system performance and so on.

Based on the existing researches on the theory, method and empirical study of search engine log mining, we present a framework on Web search engine usage mining, which includes: the data selection method, data preprocessing methods (Section 2); the main content and results of log files mining, and the comparative analysis of searching behavior characteristics of users from different countries (Section 3); the main methods of improving search engine system by the results of log mining (Section 4), and the prospect of future research (section 5).

This work was supported by the National Social Science Foundation of China(Grant No. 10BTQ050) and Humanities and Social Science Foundation of the Ministry of Education (Grant No. 09YJA870002)

2 Data Collection and Data Preprocessing

2.1 Data Set

2.1.1 Data Format

Most large-scale search engine systems log files contain two parts, user's query log and the click-through log. The query log is related to the query requests submit by searchers, which includes the query string, submitting time, the user's IP address, the page number (search results page shows 10 results per page, and the number is the page which the searcher clicks on) and other information. Take a simple query log record format from PKU TianWang search engine as an example:

```
Fri Mar 11 10:36:02 2005    // submitting time
162.105.146.*               // user's IP
Database                    // Cache Hit or not
Peking University           // query string
1                           // results page number
```

The click log is about when the user browses result and clicks on the record. It includes the time at which the user clicks on the page, the URL of the page, the user's IP address, the click page number (the page's position in

978-1-935068-76-1

the all query results), and the other click information. A click log from the PKU TianWang search engine is recorded like this:

Fri Mar 11 10:36:02 2005 // click time
162.105.146.* // user's IP
Peking University // query string
http://www.pku.edu.cn // the URL
2 // rank of click page

2.1.2 Datasets Choice

Usually, different data sets are selected and analyzed according to different research purposes. However, mainstream commercial search engine companies, considering commercial competition, are generally reluctant to furnish complete or do not want to provide their own log data, which to some extent restricts the study of search engine logs.

Related studies that have published are very different in the data sets selected, mainly in the time span: the most choose one day to study and analyze, such as in [3,4]; if conditions permit, some choose 1 week, several months or even years of data, such as in [5,6].

Generally speaking, the access rules, the content of query and the URL click behavior of group searchers are similar in a short time. For example, the query volume, click volume and page view of different users can be described by the hidden periodic model in time series method; the user's query and the process of clicking show the characteristics self-similarity, thus, the time span have a weak impact on the general characteristics study of users' searching behaviors.

2.2 Specific Terms

Some specific terms need to be defined preceding the research on search engine log mining. At present, the following concepts are being used extensively.

Term: it is a sequence of characters without any delimiter, including commas, end, colons, space bars and other specified descriptor symbols. Examples of terms are "search" or "log" etc. The choice of delimiter directly affects the results counting terms. In the log analysis, space bars are used as delimiters in the majority.

Query: what the users input in the searching box consists of one or more terms. Examples are "search engine" or "log analysis" etc. Query may include some logical operations, such as, and, or, not, etc. For a certain user, the first query he or she inputs are called initial query; If the subsequent queries are the same as the prior one, they are called repeat query; If not, they are named modified query.

Session: the whole query string sequence submitted by a certain user within a time interval, and the number of query strings is defined as length of the session. The intervals that distinguish different user's session may be several minutes, several hours or 1 day. The typical ones are 5 minutes, 15 minutes and 30 minutes or 1 day, etc. Statistical results will be different when different time intervals are used to segment the session. But from the perspective of the user's log, the session can be produced by a single user, common users or program grabbing.

2.3 The Methods of Data Preprocessing

Because original log files may exist incomplete or not consistent noisy data, data preprocessing is necessary before data analysis or pattern mining. It mainly includes data cleaning, user identification, sessions identification, English stems extracting and Chinese words segmentation, etc. Effective data preprocessing can improve the quality of mining model o reduce the time needed for mining.

(1) Data cleaning: it is to delete the data that has nothing to do with the mining tasks, such as deleting empty query caused by mal-operation of the users, removing punctuation and extra spaces according to the requirement.

(2) User identification: IP address are commonly used to distinguish different users, but due to the existence of local caches, proxy servers and firewalls, we cannot determine whether a query from a particular IP is a real single-user or not only from the user log. Researchers usually rely on the length of user session to identify certain users, such as deleting all the records that come from the same IP and whose number queries is over a certain threshold (e.g. 200) in a day.

(3) Session identification: it is to divide a user's access records into several single sessions. The time interval between the users' access to the system is commonly used to do the session identification. For example, when the time interval between two queries exceeds a set threshold, such as 30 minutes, then the user starts a new session.

(4) English stemming and Chinese words segmentation: English stemming can be used to reduce the size of the word space. The "stem" refers to the remaining part of a word after its prefix (suffix or affix) is deleted. For example, "compute" is the stem of "computer" and "computing". Since there is no delimiter between Chinese words, segmentation is needed; and different segmentation system have different algorithms, which lead to the diversity of final results.

(5) The transformation between capital and small letters: since most search engines are insensitive to capital and small letters, capital letters in queries need to be transformed into small letters to help the accumulation and consolidation of the information inquired.

3 The Content and Main Results of Log Mining

The main methods of Search engine log mining include: statistical analysis, modeling analysis and prediction, sequence pattern discovery, association rule mining, clustering analysis, etc., and the main content of mining

are: data analysis in the term level, query level, and sessions level; the characteristics of users' page viewing and clicks; the evolution trends of searching behaviors; the comparison of users' behavior from different countries; and how to improve search engine system via log mining.

3.1 The Main Statistical Indicators

There are several levels of the log mining pattern, according to current researches, the major statistical indicators and results are as follows:

(1) Term level: to analyze the single terms including term itself, the different usage of multiple languages and the wrong spelling condition and so on. For example, some researchers studied the usage of Chinese, English, Chinese and English mixing, and pure digital, etc.[4]. Deeper research results contains the English frequency distribution follow the power-law distribution (or Zipf distribution) [10,11].

(2) Query level: to study the number of terms in one query string input by the searcher, namely the query length, and to analyze the usage of Boolean operation (AND, OR, NOT) or phrase query. The major conclusions are: on average, English searching query contains 2.2 to 2.4 English terms, and the number of terms in one query is a Poisson distribution; Most Chinese searchers tend to input one term with Chinese characters, in which 2 to 4 Chinese characters is a majority; for the Web search engine, the usage of complex queries is in small scale.

(3) Sessions level: to study the length of sessions, the query revision and repeat condition, and the query submitted time interval and so on. The main outcomes including: most sessions only have one query, few users do revision; and most sessions' less than 15 minutes[5,11,13-15].

(4) Results page viewed: to study the viewing of a resulting page by a searcher while trying to locate relevant information such as Web Cache and users' viewing interval. The major statistical results are: most users view a fewer pages, usually 1-2; the time interval between pages is 2-3 minutes; only a small scale of searchers use Web Cache, take TianWang as an example, the click of Web Cache is 3.5% of the total hits[2,13,16].

(5) Hit URL: the study of the sum, the sequence number and the relation of the URLs clicked in one session or one query. Main results [2,6,9,11]are: the quantity of users' URL click comply with Heaps law, and URL click frequency is a Zipf distribution; the click behavior related to the page size; URL hit is of temporal locality, and the click process share self-similarity features, etc.

3.2 Characteristics and Comparative Analysis of Searchers from Different Countries

For the regional, cultural background and language difference user groups, may lead to a query behaviors and inquires content is different. According to the search engine's main user group location division, now already been analysis of search engine log has about 10.

The difference of users' geographical, cultural and language may lead to different searching behavior and searching content. According to the main users' location, more than 10 popular search engines have be analyzed:

United States: Excite [17]; AltaVista [5]

South America: TodoCL [18] (Chile)

Europe: AlltheWeb [19] (Norway); BWIE [20] (Spain); Fireball [21] (Germany)

Asia: NAVER [6] (South Korea); GAIS [13] (Taiwan, China); TianWang [4] (Chinese Mainland)

Although the time of log files of the researches above differ, and the sizes of the data sets are also greatly different, they each reflect some basic features of search engine users from a certain area, or under a certain cultural background. Several common points can be found by comparing and analyzing the research results.

(1) Queries input are generally short, usually containing 1 to 3 terms, two terms in English and other European languages and one in Chinese, Korean and Spanish.

(2) More than half of the users make only one query each time, and browse fewer results page, usually only 1 or 2 pages.

(3) Structures of the users' queries are relatively simple, and generally advanced search function are not used.

Furthermore, we can also find some statistical indicators of the behaviors of different search engine users are very different. The indicators are the average number of terms contained in every query, the proportion of only browsing one result page and so on. Thus we cannot extend an analysis of one search engine log to another fully. In particular, when it comes to large-scale search engines like Google, yahoo!, different strategies of system designing and services should be choose considering the differences of areas. The behaviors of user query tend to become simple, which demand us to design an interface that could be used by searchers as easily as possible.

3.3 Deep Mining

The deep mining of the search engine logs includes multi-tasking and the evolution trend of searching behavior, and the distribution of user access time. These studies require some knowledge of other fields, such as the query topic classification, time series analysis, and mathematical model building and so on. Other studies on the deep mining also include the co-occurrence of words in the query [22], hot topic & event discovery [23], named entity recognition [24], and specifically studies on the user searching behavior of a certain subject, such as multimedia searching, medicine and health searching, pornography searching and economic information searching, etc.[2].

3.3.1 Multitasking of Searching Behavior

User's information searching behaviors may include

978-1-935068-76-1 © 2011 SciRes.

more than one task (multitasking). For example, a user may search "computer" and "entertainment" kind of information successively. The research [25] showed that multitasking searching exist in different information environment (questionnaire survey, Web query, online databases, academic library). The papers [3,25-27] analyzed the multitasking searching characteristics of Excite and AlltheWeb which are respectively used mainly by American and Europeans .The results shows that on average every multitasking session includes three different themes, on every theme 4 to 5 queries are submitted. We have studied and analyzed the multitasking Web searching of TianWang with the data in 2002, and the results have shown that more than 1/3 users do multitasking Web searching; over 1/2 multitasking sessions include two different themes and 2 to 7 queries are made; the average time of a multitask is as twice as that of a general session; there are mainly three themes of TianWang users' multitasking searching: computers, entertainment and education; nearly 1/4 multitasking sessions include unclear information.

Multitasking Web searching, a common mode of information searching behavior, reveal that some users would enter Web search engines to inquire information only when they have several information needs. Multitasking Web queries are complicated in query modes, so more effective retrieval technology is needed to serve for such complicated searching structures.

3.3.2 The Evolution Trends of Searching Behavior

The information needs of a single user undergo constant changes, so do the query themes of user groups on the search engine. In order to analyze this migration, Spink selected samples of queries input by users in 1997, 1999 and 2001 from Excite search engine, each year with over 2000 queries. After manual classification (defining 11classes), Spink found that people's information needs are now changing "from e-sex to e-commerce "; that is to say, information needs on commerce have been increasing gradually while information needs on entertainment haven been decreasing comparatively among the Web user of Excite.

We have made sampling analysis on TianWang's usage logs of 2001-2005 and the results show: the number of the terms contained in one queries have tended to increase; the length of sessions has decreased year by year; the result pages users browse have become fewer and fewer; the time interval of browsing has become shorter; the number of Chinese characters in the queries have almost remained the same and queries with 2 to 4 Chinese characters are in majority; the proportion of the occurrence of clicking in the query results has tended to decrease; the relevance between the number of queries and times of clicking has become weaker; the themes of the Web user queries have changed more quickly.

3.3.3 The Distribution of Visiting Time

To know the distribution of the time of uses visiting systems is beneficial to the configuration of system sources. We have analyzed that of the TianWang and the results show that users visit it by rules. Three wave crests appear in a day: 10:30 a.m., 4:30 p.m. and 8:30 p.m. The least visiting happens during 3:00-7:00 a.m., which is similar to the data of CNNIC. Further researches show that during a short period TianWang users' visits are the same on weekdays (from Monday to Friday), and so are the distribution of visiting time. But they are slightly different from those in weekends (Saturday and Sunday) and the whole fluctuates. The visits can be described by hidden periodic model of time series.

4 Application: The Improvement of Search Engine System

Mining search engine logs can effectively improve the system's effectiveness, efficiency, service and other aspects of performance. Specifically, in the aspects of effectiveness, we can make use of user' queries, the user clicks the URL and other feedback to improve the quality of the results of sorting; in the aspect of system efficiency, the use of Cache replacement policy can improve the system response time; in the aspect of system services, We can find some Web queries similar to the given query, and provide recommendation services and so on.

4.1 Improvement of Ranking Quality

After the user submits a search request, a click on a certain URL on the results page generally indicates the user s' agreement to the URL, and in most cases, the URLs containing unrelated information will not be clicked. In 2002, Zhang Dell proposed a method that makes use of user click record to improve the quality of the ranking results [28]. The method was first used in the Chinese image information retrieval, and then Baeza-Yates applied it to text information retrieval [29], achieving good experimental results. Zhang named such methods as MASEL (Matrix Analysis on Search Engine) algorithm.

MASEL method tries to find the relationship among the users, the queries and the clicked URLs; The basic hypothesis is: good users submitted good queries, good queries returns good URLs, good URLs are clicked by good users. According to the user click records, the three basic variables are defined recursively. It is very similar to the recursive relationship established between the page's Authority and Hub in the Hits algorithm [30].

Baeza-Yates found out that MAEL is effective in a small-scale experiment using the logs of Chile's Todo search engine and pointed out the selection of log cycle as a certain impact on accuracy of the experimental results. In particular, when a log with a longer period of time was selected, the accuracy of multi-meaning words (query Lexical items) decreased.

Aiming at the sequencing of the results of meta search

engine, Joachims presented a method that uses click-through data as the training set to learn the retrieval function[31]; the experiment ranking results are superior to those of Google's.

4.2 Cache Replacement Policy

Using Cache in the client-side or the index-side of the search engine can greatly improve the average response time of user queries and improve the efficiency of the system. Statistical analysis of the distribution of users' searching is very concentrated (i.e., the localized feature of searching), which reveals the feasibility of using Cache: put the query results of the items with a high number of times of queries in the Cache, and the Cache of small capacity will be able to hit most of the user's queries, so larger Cache hit rate can be achieved with smaller space.

In the paper [10] Xie first discussed that using the Cache in the search engine system can reduce server load and reduce system response time. In the research[8] strategies of the TianWang search engine's several client-side Cache replacement are compared, whose results show that the Cache hit rates of LRU (Least Recently Used) and LFU (Least Frequently Used) are significantly better than that of FIFO (First In First Out), the effect would be almost the same as that of LRU. If a attenuation factor is given to LFU, the effect would be slightly better than the LRU results. Taking the complexity of implementation into account, LRU and FIFO are relatively simple, which LFU would attenuate in the process of replacement must traverse the entire Cache, and thus it spends much more replacement time LRU and FIFO, yet achieves effects almost the same as LRU. Considering all the elements above, the literature believes that LRU is the best Cache Replacement Policy.

4.3 Finding Relevant Queries

The query a search engine user enters is usually short, short words can express a comparably broad range of themes easily leading to ambiguity, and the user is often unable to accurately express his or her information needs. Due to the above reasons, sometimes users need to make ongoing amendments on their queries in order to find satisfied information. For the convenience of the user to amend queries, some search engines such as Google, Baidu, etc. have a list of related queries in the results page the systems return for users to make references when make amendments after they submit searching requests the first time, which make the users express their information need more accurately.

Traditional information retrieval systems use query expansion to find related queries, and the main methods are based on user feedback, partial or full information for query expansion etc. [32,33]. These methods typically rely on the specific contents of each document in a document set. It is complex to achieve and thus it is not much used in the practical retrieval system.

Based on the usage logs of search engines we find relevant query is a feasible method. The current studies consider two basic factors [16,34]: (1). If two queries contain the same lexical items, they may be relevant, and the more lexical items they share, the higher their similarity, and (2). As to two different queries, if the user clicks on the same URL in the query result then they may be relevant. In addition, other factors include: the number of queries, the query distribution of different users, and the number of clicking the same URL, the type similarities between corresponding documents of the clicked URL.

Then, we can utilize some method to combine these factors then find queries related to a certain given query; one way is to manually tag part of the training data, establish a regression model and determine the related Web queries by the extent of correlation [16], and multiple linear regression and support vector regression are the main regression methods. In general, the parameters of linear regression calculation is simple and can quickly predict the predicting results of each query, but the prediction accuracy is slightly worse; the parameters of support vector regression involve longer training time is longer but the accuracy is better. The research has addressed that prediction accuracy of the results of different types of queries (information type, navigation type, type of things [35]), differ greatly.

Only through the user click log to cluster the query is another way to discover related queries: firstly, a bipartite graph would be constructed, that is, based on the user click records, connecting some corresponding elements of query set and the clicked URL set; then cluster by agglomerative iterative algorithm, merging two queries and two URLs in turn until the end of the iterative algorithm [36]. A deficiency of this method is that "noisy data" cannot be effectively dealt with, that is, if the user mistakenly click on a URL, then two unrelated queries may be gathered together forever [37].

Association rules can also be used to identify relevant Web queries [38]. Specifically, the query is seen as an item in the association rules, and the session in the query log is seen as a transaction—a transaction is an aggregation of the queries submitted by a single user within a certain time interval. Then association rules are utilize to mine algorithm to find strong association rules, and then related Web searching can be found.

5 Conclusion and Prospect

With the help of the Web searching characteristic and laws abstracted from log files, we can not only improve the performance of search engine system, but also make some contribution to the information behavioral research. Large commercial search engine companies have been paying great attention to log mining researches, such as the cooperation between Baidu company and Peking university, with the establishment of "Chinese searching behaviors laboratory", SOHU company, cooperating with

978-1-935068-76-1 © 2011 SciRes.

Tsinghua university, has established a "Sohu searching technology laboratory" etc. The study of this field calls for the knowledge of information science, computer science, data mining, artificial intelligence, human-machine interaction, education psychology, cognitive science and other various aspects of knowledge. According to currently published papers, the achievements are primarily about English search engines, while the studies in Europe, Asia are relatively rare.

Aiming at the Web search engine logs, this paper proposes a framework on Web search engine usage mining , which through the entire process of log mining, involving used theory, technology and methods, as well as the summary of existing achievement . This framework can be a guide to the search engines study and similar Web log mining research.

There are still many issues worthy of further study in this field, among which one focuses on how to make use of the results of log mining to further improve the performance of search engines. Other studies include: how to divide the sessions more reasonably; what is the relationship between the searching behaviors of users from different regions and cultures; how to use the evolution rule of users' search behaviors to assess the performance of search engines; how to predict information needs according to visiting logs; what relationship exists between users' searching behaviors under complicated circumstances and in laboratories (or according to the findings from surveys); how to find out and identify the characteristics of users' searching behaviors of different types, such as genders, ages and knowledge backgrounds; what the similarities and differences of users' search content and behavioral characteristics on weekdays and in the weekend are; how to use some models of cognitive psychology to explain various kinds of query behavior. In order to deeply explore search engine logs. There are still many challenges to encounter by utilizing comprehensively knowledge from various fields.

References

[1] LI Xiaoming, YAN Hongfei, WANG Jimin. Search Engine: Theory, Technology and Systems [M]. Beijing: Academic Press. 2005.

[2] Spink A，Jansen B J. Web search: public searching on the Web [M]. Netherlands: Kluwer Academic Publishers. 2004.

[3] Ozmutlu S, Spink A, Ozmutlu H. A day in the life of Web searching: an exploratory study [J]. *Information Processing and Management*, 2004,40, P319-345.

[4] Wang Jimin, Chen Chong, Peng Bo. Analysis of the User Log for a Large-scale Chinese Search Engine [J]. *Journal of South China University of Technology*. 2004, 32(S), P1-5(Ch).

[5] Silverstein C. Analysis of a very large altavista query log [J]. *ACM SIGIR Forum*, 1999, 33(1), P6-12.

[6] Park S, Lee J H, Bae H J. End user searching: A Web log analysis of NAVER, a Korean Web search engine [J]. *Library & Information Science Research*, 2005. 27, P203-221.

[7] Wang Jimin, Peng Bo. Modeling Quantity of Users' Access for Search Engine [J]. *Computer Engineering and Applications*. 2004, 40(25), P9-11 (Ch).

[8] Wang Jianyong, Li Xiaoming, et al. Web Search Engine: Characteristics of user behaviors and their implication [J]. *Science in China (Series E)*, 2001, 31(4), P372-384(Ch).

[9] Wang Jimin, Peng Bo. User Behaviors Analysis for a Large-scale Search Engine [J]. *Journal of the China Society for Scientific and Technical Information*. 2006, 25(2), P154-162 (Ch).

[10] Xie Yinglian, O'Hallaron D. Locality in Search Engine Queries and Its Implications for Caching. in Proc. IEEE Infocom 2002，New Jersey: IEEE Press, 2002, P1238-1247.

[11] Baldi P, Frasconi P, Smyth P. Modeling the Internet and the Web, probabilistic methods and algorithms [M]. John Wiley. 2003.

[12] Jansen B J, Spink A, Saracevic T. Real life, real users, and real needs: a study and analysis of user queries on the web [J]. *Information Processing and Management*, 2000, 36, P207-227.

[13] Jansen B J, Spink A. How are we searching the World Wide Web? A comparison of nine search engine transaction logs [J]. *Information Processing and Management*, 2006, 42, P248-263.

[14] Lau T, Horvitz E. Patterns of Search: Analyzing and Modeling Web Query Refinement. in 7th Int. Conf. on User Modeling. Springer, 1999.

[15] He D, Goker A. Detecting session boundaries from Web user logs. in 22nd Annual Colloquium of IR Research 2000, UK: Cambridge. 2000.

[16] Wang Jimin. A Study on Web Using Mining in China Search Engine. A Post-doctorial Report in Peking University. 2005.

[17] Spink A, Jansen B J , et al., From e-sex to e-commerce: Web search changes [J]. *IEEE Computer,* 2002, 35(3), P133-135.

[18] Baeza-Yates R. Applications of Web Query Mining. in European Conference on Information Retrieval (ECIR'05). 2005. Spain: Springer.

[19] Spink A, Ozmutlu S, et al., US versus European Web searching trends [J]. *SIGIR Forum*, 2002, 32(1), P30-37.

[20] Cacheda F, Vina A. Understanding how people use search engines: a statistical analysis for e-business. in Proceedings of the e-Business and e-Work Conference and Exhibition 2001. Italy.

[21] Hoelscher C, Strube G. Web search behavior of internet experts and newbies [J]. *International Journal of Computer and Telecommunications Networking*, 2000, 33, P337-346.

[22] Wang P, Berry M W, Yang Y, Mining longitudinal Web queries: Trends and patterns [J]. *Journal of the American Society for Information Science and Technology*, 2003. 54, P743-758.

[23] Gu YQ，Cui JW，et al. Detecting Hot Events from Web Search Logs. In Proceedings of 11th International Conference on Web-Age Information Management. 2010. China.

[24] Xu G, Yang SH, et al. Named Entity Mining from Click-Through Data Using Weakly Supervised Latent Dirichlet Allocation. In Proceedings of 15th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2009.

[25] Spink A, Ozmutlu H C, Ozmutlu S. Multitasking information seeking and searching processes [J]. *Journal of the American Society for Information Sciences and Technology*, 2002, 53(8), P639-652.

[26] Ozmutlu S, Ozmutlu H C, Spink A. Multitasking Web Searching and Implications for Design. in ASIST'03: Annual Meeting of the American Society for Information Science and Technology. 2003. Long Beach, CA.

[27] Ozmutlu S, Ozmutlu H C, Spink A. A Study of Multitasking Web Searching. in IEEE ITCC'03: International Conference on Information Technology: Coding and Computing. 2003. Las Vegas, NV, USA.

[28] Zhang D, Dong Y, A novel Web usage mining approach for search engines [J]. *Computer Networks*, 2002. 39, P303-310.

[29] Baeza-Yates R. Query Usage Mining in Search Engines, in Web Mining: Applications and Techniques, A. Scime, Editor. Idea Group. 2004, P307-321.

[30] Chakrabarti S. Mining the Web: Discovering Knowledge from Hypertext Data [M]. Morgan-Kaufmann. 2003.

[31] Joachims T. Optimizing search engines using clickthrough data. in Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD), 2002.

[32] Baeza-Yates R, Ribeiro-Neto B. Modern Information Retrieval

[M]. Addison-Wesley-Longman. 1999.
[33] Cui Hang, Wen Ji-Rong, et al. Query Expansion by Mining User Logs. IEEE transactions on knowledge and data engineering, 2003. 15(4): 829-839.
[34] Wen Ji-Rong, Nie Jian-Yun, Zhang Hong-Jiang. Query clustering using userlogs. in Proceedings of the 10th World Wide Web conference. 2001. New York: ACM Press.
[35] Kang I H, Kim G. Query Type Classification for Web Document Retrieval. in Proceedings of the Twenty-Sixth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2003, Jul 28-Aug 1 2003. 2003. Toronto, Ont., Canada: Association for Computing Machinery.
[36] Beeferman D, Berger A. Agglomerative clustering of a search engine query log. in Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining. 2000.
[37] Chan W S, Leung W T, Lee D L. Clustering Search Engine Query Log Containing Noisy Clickthroughs. in Proceedings of the 2004 International Symposium on Applications and the Internet (SAINT'04). 2004.
[38] Fonseca B M, Golgheret P B, et al. Using association rules to discovery search engines related queries. in First Latin American Web Congress (LA-WEB'03). 2003. Santiago, Chile.

978-1-935068-76-1 © 2011 SciRes.

On Design of the Frame of the Metallurgical Information Database

Denan GU, Zhimin YU
China Metallurgical Information & Standardization Institute, Beijing, China
Email: gudenan@cmisi.cn, yuzhimin@cmisi.cn

Abstract: In this paper, subjects and functions of the Metallurgical Information Database are introduced. Design of the frame of the main subject, scientific and technical documents, is discussed minutely. Conclusions are gotten as follows: It is necessary for us to make the subjects more scientific, systematic and logical, but also take clustering of documents into account, and pay attention to the professional's custom of using databases when designing the frame of the Metallurgical Information Database; According to meanings of concepts, glossary should be corrected necessarily. If there is a small amount of documents under certain category, we can ignore it. When necessary, we can do some treatment about the upper and lower concepts, concept refinement and establishment of new categories. We need to be concerned about the development of domestic and international metallurgical standards. Use reference books with developing viewpoint.

Keywords: the Metallurgical Information Database; subjects; functions of the database; scientific and technical documents; design of the frame

试论冶金专业化信息检索系统架构设计

顾德南，于治民
冶金工业信息标准研究院，北京，中国，100730
Email: gudenan@cmisi.cn, yuzhimin@cmisi.cn

摘　要：本文介绍了冶金专业化信息服务平台的栏目设置和系统功能，详细论述了主打栏目——“科技文献”的架构设计，得出如下结论：冶金专业化信息服务平台架构的设计原则，既要考虑学科的科学性、系统性和逻辑性，又要兼顾文献聚类情况，同时还要尊重专业人员的使用习惯；根据概念的内涵，进行必要的术语修正；如果某一类目下文献量很少，则可忽略该类目；必要时，进行上下位概念的处理、概念细化和新类目设置；关注国内外冶金标准的发展动向；带着发展的眼光使用工具书。

关键词：冶金专业化信息服务平台；栏目设置；系统功能；科技文献；架构设计

1 引言

有研究表明，科技文献增长速度与时间成指数函数关系。面对海量信息，承担企业科技创新重任的科研人员常常因文献检索费时费力而感到困惑，因此有必要建立针对不同细分研究领域的专业化信息服务平台。在这一平台中，不仅将信息按不同细分领域进行有序的组织和排列，而且提供多元化、多层次、个性化的信息服务，使科研人员在最短时间内获得所需信息资源。

国家科技图书文献中心（以下简称中心）作为国家科技基础条件平台和国家科技创新体系的重要组成部分，近几年将“积极拓展个性化、专业化服务，提升中心整体的服务能力”纳入年度工作计划，我院作为中心的成员单位之一，也把开创新的服务模式列为工作重点。因此，建立冶金专业化信息服务平台，既有必要，也有可能。2008～2009 年，在中心的大力支持下，我院先后建立了炼铁和炼钢两个冶金专业化信息服务平台。2010 年，我院还将建立冶金主流程中最后一个专业化服务平台——轧钢平台。

Figure 1. First page of the Steelmaking Information Database
图 1. 炼钢专业服务平台首页

炼铁、炼钢和轧钢三个冶金专业化信息服务平台设在我院主办的中国冶金信息网首页上，采用专用服务器。其中，炼钢专业服务平台首页如图 1 所示。

2 栏目设置

冶金专业化信息服务平台各栏目的内容和信息来源如表 1 所示。

Table 1. Subjects of the Metallurgical Information Database

表 1. 冶金专业化信息服务平台栏目设置栏目名称	内容	信息来源
科技动态	中、英文产业政策、新技术、新工艺、市场行情、统计信息、会展消息、新建和在建项目等。	根据馆藏资源自行加工；搜索网上资源；搜集相关学协会、网站和信息机构的信息资源。
科技文献	中、英文炼铁科技论文，包括期刊、会议论文、学位论文、科技报告等。下设四级类目，分类导航。	以 NSTL 网上资源为基础，按细分专业进行文献组织。
文献检索	根据用户填写的文献检索委托书，代为检索，并提供全文。	馆内外资源
定题服务	根据用户需求，定期提供文献推送服务。	馆内外资源
特种文献	与炼铁专业相关专题咨询报告、高端图书等。	自建
全文提供	根据用户填写的全文提供申请单，提供全文。	馆内外资源
行业热点	有关炼铁热点问题的综述文章	根据馆藏资源自行加工；搜索网上资源。
资源推荐	炼铁方面的中英文期刊、会议文献、图书科技报告和网站等。	NSTL 资源
冶金企业	中外炼铁企业及其相关信息	自建
参考咨询	实时与非实时咨询	链接到 NSTL 参考咨询页面
其他服务	工程咨询、科技查新、专利信息咨询。	链接到中国冶金信息网相关页面

3 系统功能

冶金专业化信息服务平台具备以下功能：

3.1 浏览功能

所有栏目都要具备浏览功能，用户无需检索，可以直接浏览所关心子类的全部信息。

3.2 检索功能

科技动态、科技文献、特种文献、行业热点、资源推荐、冶金企业等栏目要具备检索功能。检索分为普通检索和高级检索。

3.3 请求全文功能

任何会员都可以通过“科技文献”和“全文提供”两个栏目的多个入口提出全文请求。工作人员可以在后台进行一系列操作，如接收请求、标注处理状态、调整页数等。

3.4 帐务管理功能

会员可以随时通过前台进行个人帐务查询。系统实现自动帐务管理，同时留有人工干预入口，工作人员可以通过后台进行帐务人工干预，如充值、修改页码等。

3.5 互动功能

为使平台更加贴近用户，在平台的显要位置提供接收用户意见和建议以及回复用户的功能。

冶金专业化信息服务平台采用 JAVA 语言开发底层的封装类，采用 ORACLE 数据库进行用户管理，采用 TRS 数据库作为网站全文检索平台。

4 “科技文献”栏目架构设计

“科技文献”为冶金专业服务平台的核心栏目，内容为炼铁、炼钢及轧钢中、英文科技论文的文摘信息，文献类型包括期刊论文、会议论文、学位论文、科技报告等。

“科技文献”是冶金专业化信息服务平台的主打栏目，其架构设计得是否科学、合理，是否符合科技人员的使用习惯，将直接影响平台的使用效果。因此，“科技文献”栏目的架构设计是决定整个平台成败与否的关键环节。

在此，以炼钢平台为例，阐述“科技文献”栏目架构设计的要点。

978-1-935068-76-1 © 2011 SciRes.

4.1 架构设计思路

“科技文献”栏目以 NSTL 网上资源为基础，采用树状分类导航，所设五级类目如图 2 所示。

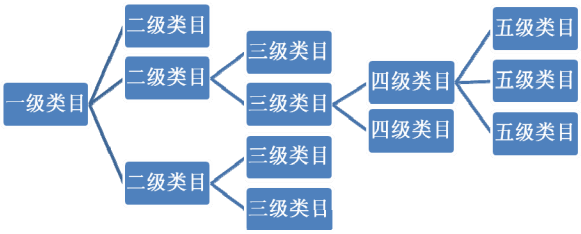


Figure 2. Pictrue of the five-grade items of the “Science and technology documents”
图 2. “科技文献”栏目五级类目示意图

“科技文献”栏目架构设计的基本原则是，既考虑学科的科学性、系统性和逻辑性，又兼顾检索速度和各类目下文献量的多少。所要实现的目标是，读者只需要点击类目名称，系统便会自动将该类目下的所有文献检索出来，并将文献题名分页显示于页面上。点击某篇文献的题名，便可进入下一级页面，进而全面了解该篇文献的题名、刊名（会议录名等）、年卷期、页码、关键词、语种、分类号、作者、作者单位、馆藏号码、摘要等信息。在任何一个页面，注册用户均可进行全文订购。

由于炼钢方法多种多样，钢种更是数不胜数，所以炼钢平台的架构很复杂。就“各种钢的冶炼”一大类而言，钢的分类方法颇费周折。相关资料很多，如《中国图书资料分类法》（第四版）（以下简称《中图法》）[1]、《中国冶金百科全书（钢铁冶金）》[2]、《钢铁工业主题词表》（第二版）[3]、《GB/T 13304.1-2008 钢分类.第 1 部分:按化学成分分类》[4]、《GB/T 13304.2-2008 钢分类.第 2 部分:按主要质量等级和主要性能或使用特性的分类》[5]、《冶金学名词》[6]、《炼钢学》[7]以及某知名网站使用的《钢的分类》。这些资料各有千秋，采取任何一种，都无法完全满足我们构建专业化服务平台的要求。

《中图法》是全国各类图书馆和信息研究机构广泛采用的大型检索语言工具书。经过多次修订，其体系结构、类目设置和标识符号等日臻完善。《中图法》在文献标引、存贮和检索方面发挥着无可替代的作用。因此，我们以《中图法》为基础，综合以上各种分类方法的优势，设计“科技文献”栏目的架构。

《GB/T 13304.1-2008 钢分类.第 1 部分:按化学成分分类》和《GB/T 13304.2-2008 钢分类.第 2 部分:按主要质量等级和主要性能或使用特性的分类》虽然很科学、系统，代表当今国内钢的最新分类趋势。但是由于在这两个标准中，根据化学成分和用途的不同，即使同一种钢，也可能被分到不同类别中，例如：根据 GB/T 13304.2-2008，同为“低合金结构钢”，有的属于“普通质量低合金钢”，有的则属于“优质低合金钢”。因此如果按这两个国标建立“科技文献”栏目架构，则不易于实现检索。这是我们为什么没有基于这两项国家标准设计冶金专业化信息服务平台“科技文献”栏目架构的原因所在。

4.2 术语修正

我们对《中图法》中的个别词汇进行了修正，例如：“特殊用途钢”包括“不锈钢”、耐热钢和电磁钢，根据其内涵，我们将其修正为“特殊性能钢”。

4.3 忽略次要类目

在设计架构时，我们适当忽略了个别条目。例如，《中图法》中的“碳素钢”中包含“极软钢”和“极高碳钢”，考虑到“极软钢”文献量很少，故加以忽略。

4.4 上下位概念的处理

《中图法》有时出于文献数量的考虑，把本属于某一类概念的下位概念单列出来，与上位概念设为同一级别。例如，“工具钢”本来与“结构钢”一样，属于“优质钢”的下位词，但《中图法》却把它与“优质钢”并列起来。参照某知名网站使用的《钢的分类》，炼钢平台把“工具钢”归属到“优质钢”中。同理，把“极高碳钢”划入“高碳钢”中，并更名为“超高碳钢”。

4.5 概念细化

在炼钢平台中，我们将某此概念进行了细化。例如，在《中图法》中，“不锈钢”一词没有进一步细分，但我们在检索过程中发现，仅在 NSTL 网站中文库中就能检索出关于“不锈钢”的文献 42000 余篇，显然过多。于是，我们将该类目细分为：奥氏体不锈钢；超低碳不锈钢；含钛不锈钢；马氏体不锈钢；双相不锈钢；铁素体不锈钢；无铬不锈钢；无镍不锈钢；不锈耐酸钢。

在《中图法》中，“低碳钢”也没有再进行分类，

根据实际文献量，我们由该类目衍生出“超低碳钢”。同理，“结构钢”中增加了“易切削钢”一项；“碳素结构钢”中细化为“普通碳素结构钢”和“优质碳素结构钢”；“合金结构钢”中增加了“表面合金化钢”（指渗氮钢、渗钒钢、渗铬钢、渗铝钢、渗碳氮钢、渗碳钢等）和“低合金结构钢”；在“特殊性能钢”中增加了“耐磨钢”和“耐蚀钢”。

概念细化需适可而止。在《钢铁工业主题词表》（第二版）中，“普通碳素结构钢”分为“甲类碳素钢”、“特类碳素钢”和“乙类碳素钢”三种，根据在 NSTL 网站实际检索结果，每一类文献数量都很少，甚至为 0。因此，在炼钢平台中，“普通碳素结构钢”不再细分。

4.6 设立新类目

根据读者的检索习惯，我们又增加了一类专门用途钢，其节选内容如表 2 所示。

Table 2. Part of the items of special usage steels
表 2. 专门用途钢各级类目节选

一级类目	二级类目	三级类目	四级类目	五级类目
六、各种钢的冶炼	2、优质钢	(4)专门用途钢	a.齿轮钢	
			b.船舶用钢	
			c.电工钢	(a)电机钢
			d.阀门钢	
			e.钢管钢	
			f.钢轨钢	
			g.管线用钢	
			h.压力容器用钢	(a)锅炉钢
			i.海洋用钢	(a)采油平台用钢
			j.焊条钢	
			k.航空及航天用钢	
			l.核工业用钢	
			m.化工用钢	
			n.建筑用钢	(a)桥梁钢
				(b)高层建筑用钢
				(c)钢筋混凝土用钢
			o.军工用钢	(a)兵器用钢(炮弹钢等)

4.7 国内外冶金标准发展动向

了解国内外冶金标准发展动向对于构建专业化信息服务平台架构很重要，在我国现行的国家标准中，已将“低合金钢”列为与“合金结构钢”等同的位置，而且是发展方向。两者之间有严格的界限。了解这一点后，我们便设计出如下关于“结构钢”的架构，如表 3 所示。

Table 3. Items of structural steels
表 3. 结构钢各级类目

一级类目	二级类目	三级类目	四级类目	五级类目
六、各种钢的冶炼	2、优质钢	(1)结构钢	a.碳素结构钢	(a)普通碳素结构钢
				(b)优质碳素结构钢
			b.合金结构钢	(a)稀土钢
				(b)表面合金化钢(渗氮钢、渗钒钢、渗铬钢、渗铝钢、渗碳氮钢、渗碳钢等)
			c.低合金钢	
			d.轴承钢	(a)滚珠钢
			e.弹簧钢	
			f.变压器钢	
			g.易切削钢	

978-1-935068-76-1 © 2011 SciRes.

4.8 工具书的作用

在建立炼钢平台的过程中，不可避免地要参考一些工具书。科学技术是不断发展的，因此，在使用这些工具书时，我们要考虑有些词条是否已经过时。另一方面，任何书籍的编写都有其历史背景，因此可能存在历史局限性，甚至出现一些错误的说法。鉴于这种情况，我们不应迷信任何一本工具书。在某工具书中，转炉炼钢中有“氧气蒸汽底吹”一项。经过向有关专家请教，得知炼钢不可能吹蒸汽，因为氧气底吹转炉炼钢时，钢水温度高达 1000℃以上，在这种温度下，水的体积将增大 5000 倍以上。这时假如蒸汽在炉内无自由发散余地，便会发生恶性爆炸事故。因此炉内严禁入水。但是每个炉子都需要使用大量水冷，因此要严密注意防止冷却水漏入炉内。于是我们在架构中删除了“氧气蒸汽底吹”这一类目。关于“氧气二氧化碳底吹”，专家认为准确的说是氧气底吹。

5 结论

⑴ 冶金专业化信息服务平台架构的设计原则，既要考虑学科的科学性、系统性和逻辑性，又要兼顾文献聚类情况，同时还要尊重专业人员的使用习惯。
⑵ 根据概念的内涵，进行必要的术语修正。
⑶ 如果某一类目下文献量很少，则可忽略该类目。
⑷ 必要时，进行上下位概念的处理、概念细化、设立新类目。
⑸ 关注国内外冶金标准的发展动向。
⑹ 带着发展的眼光使用工具书，考虑工具书编撰的历史局限性。

References（参考文献）

[1] Editorial Committee of China Document Classification. China Document Classification (4th edition) [M]. Beijing: Scientific and Technical Documents Publishing Press, 2000.
中国图书资料分类法编辑委员会.中国图书资料分类法（第四版）[M].北京：科学技术文献出版社，2000.

[2] Sub-editorial Committee of Iron and Steel Metallurgy, Editorial Committee of China Metallurgical Encyclopedia. China Metallurgical Encyclopedia (Iron and Steel Metallurgy) [M].Beijing: Metallurgical Industry Press, 2001.
中国冶金百科全书总编辑委员会（钢铁冶金）卷编辑委员会.中国冶金百科全书（钢铁冶金）[M].北京：冶金工业出版社，2001.

[3] Revising Group of Thesaurus for Iron and Steel Industry (2nd edition). Thesaurus for Iron and Steel Industry (2nd edition) [M]. Beijing: China Metallurgical Information & Standardization Institute, 1991.
钢铁工业主题词表（第二版）修订组.钢铁工业主题词表（第二版）[M].北京：冶金部情报标准研究总所，1991.

[4] China Iron and Steel Association. GB/T 13304.1-2008 Steel Classification. Part 1: Classification of according to chemical composition [S]. Beijing: Standards Press of China, 2008.
中国钢铁工业协会. GB/T 13304.1-2008 钢分类.第 1 部分:按化学成分分类[S].北京:中国标准出版社,2008.

[5] China Iron and Steel Association. GB/T 13304.2-2008 Steel Classification. Part 2: Classification of according to main quality classes and main property or application characteristics [S]. Beijing: Standards Press of China, 2008.
中国钢铁工业协会. GB/T 13304.2-2008 钢分类 第 2 部分按主要质量级别和主要性能或使用特性的分类[S].北京:中国标准出版社,2008.

[6] Approval Committee of Metallurgical Terms. Metallurgical Terms [M]. Beijing: Science Press, 2001.
冶金学名词审定委员会.冶金学名词[M].北京：科学出版社，2001.

[7] Peiran Zheng. Steelmaking [M]. Beijing: Metallurgical Industry Press, 1994.
郑沛然.炼钢学[M].北京：冶金工业出版社，1994.

Scientific Research

An Intercity Heterogeneous e-Government Data Integration Model Based on Topic Map Technology and Database Analysis

Lixin XIA, Fei YE, Li HUANG, Xiufeng CHENG

Department of Information Management, Central China Normal University, Wuhan, China

Email: xialx@mail.ccnu.edu.cn, jwcyefei@mail.ccnu.edu.cn, tomoyomavis@hotmail.com, xiufengcheng@gmail.com

Abstract: Data integration for the unified web-based information platform of heterogeneous e-Government system becomes the most important issue in e-Government system conformation. This paper proposes a new model of data integration based on in-depth analysis of Topic Map technology and Relational Database Management System (RDBMS)**.** This model uses OAI protocol for Metadata harvesting to automatically generate a topic index layer between the application layer and data layer of e-Government systems. To integrate the intercity heterogeneous e-Government systems, we derive topic maps in the topic index layer by identifying topic, association and occurrence of each topic map in the RDBMS, and then amalgamate topic maps by measuring their similarities.

Keywords: Topic map; e-Government; RDBMS; Database Reverse Engineering; OAI

1 Introduction

Rapid changes in information technology have brought the rapid development of e-Government. The focus of governmental information platform developing has shifted gradually from business informationization to the integration and sharing of information resources. As more and more independent e-Government system been built, problems of information sharing between these heterogeneous systems become more difficult. How to make sure the existing network infrastructure, business systems and information resources are properly used and effectively exchangeable is the key issue to the success of e-Government system integration.[1,2]. Therefore, how to implement the vertical structure of the e-Government to the horizontal data sharing is an important task that needs to be resolved urgently.

At present, most e-Government systems are using relational database systems (RDBMS, Relational Database Management System) for data management. The data structure of heterogeneous e-Government has the following characteristics: (1) Data heterogeneity. The difference in the structure and the storage model of each e-Government system, as well as the data management model and application process constitutes the data heterogeneity between systems; (2) Data element complexity. Data of government system comes from various industries and sectors and generated constantly, so the composition of data is relatively complex; (3) Data thematic. On the usage perspective view, despite the complex composition and large number of government data, the data can also be divided into a number of different organized topics which focused on the intersection and association of inter-related data of different departments; (4) Data environmental inconsistency. The various computer operating system (OS) and database systems (DBMS), as well as hardware and different system structures, formed the inconsistency of the data environment.

Based on the in-depth study of the advantages of topic map in the knowledge organization and information resources and the above data characteristics, this paper constructed data integration model of e-Government on the basis of topic map by using the reverse engineering in the heterogeneous database, the model make use of ER model and OAI (the Open Archives Initiative)-related topic map technology generation and use of the integration of topic map in order to realize data integration in heterogeneous e-Government systems integration, and support for inter-governmental sector business collaboration, enhance integration capability and coordination of e-Government systems. This method has the advantages of a relatively small amount of calculation, the economy and low cost, while ensuring the semantic data integration, scalability and flexibility.

2 The Feasibility of Topic Map Applications

A complete structure of e-Government system should at least include three levels: user layer, application layer and data layer. The user layer manages the identity authentication of user and provides interactive interface. Application layer is combination of application program and service program which are have close relationship to the theme. Data layer is all data collection of resources and in e-Government systems the Data layer are generally refers to an RDBMS. One purpose of the application of topic maps technology in e-Government system is to integrate the date layer of the affairs of the systems and Application layer. Another purpose is to increase a third-

978-1-935068-76-1

party topic maps index layer between the application layer and data layer. As shown in Figure 1, at first the index layer extracts data from the RDBMS, and use topic map to constitute a structured semantic indexing. The second the Application layer sends retrieval request to the user layer according to the operation of user layers. Index layer on the first response and semantic retrieving the preliminary and do the semantic retrieval roughly. Then according to the semantic retrieval the Index layer put the result to the actual data or provides the keywords which are after processing and semantic retrieval to the RDBMS. And at last put the eventual result to return user layer.

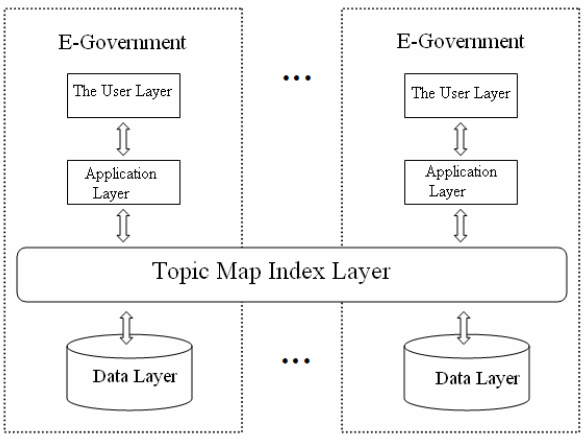


Figure 1. Topic Map Index Layer.

2.1 Relational Database

Relational database transfer complex data structure into a series of "two-dimensional form" use this binary relation to certify the links between entity types and entities. The designs of relational database include the two parts: relational schema and integrity constraints[3]. The various properties of relational database interrelated, interdependent and mutually constraining, thereby constituting a rigid whole. To avoid data redundancy, logical errors and operational anomalies, as well as improve access efficiency and accuracy, it must follow the appropriate rules (relational schema) while making the database design. Therefore, to achieve conversion from a relational database to the topic map, we should first follow a certain strategy which extracts these patterns of information from the semantic data level.

2.2 Feasibility Analysis

Topic Maps is essentially a dynamic, structured index which is independent of specific information resources. To guide the user to a specific address to obtain practical information resources through retrieving to an instance related to subject[4]. Therefore, using topic maps to integrate data in heterogeneous e-Government systems is a process of a mapping and navigation of database. It organizes abstract, isolated data in the database and builds an structured semantic web on the top of database which independent of specific technologies platform[5]. This method can effectively avoid a large number of tedious calculations in heterogeneous database standardization, data consolidation, data association. Although the merger method of traditional database can solve the heterogeneous data consolidation issues, but it can not properly resolve problems of the semantic heterogeneity, while which is the topic map application can solve.

For the feasibility of the conversion between the database and topic map, a simple example listed below. Suppose it is a government system database, three tables are as follows:

Table 1. Unit information

Unit Code	Unit	Corporation Code	Corporation
Unit_001	Overseas Investment Bureau	Staff_001	Mike
Unit_002	National Territory Bureau	Staff_002	John

Table 2. Position Information

Position Code	Position
Position_01	Director General
Position_02	Deputy Director General

Table 3. Staff Information

Staff Code	Staff Name	Unit Code	Certificate	Position Code
Staff_001	Mike	Unit_001	Undergraduate	Position_01
Staff_002	John	Unit_002	Graduate	Position_02

Shown in Figure 2, according to the three elements of Topic Maps TAO principle, analyzing information of three charts can identify Topic, Association and Occurrence in data information. Themes in charts include: unit, staff, position, certificate, national territory bureau and so on, among them the unit, staff and others are Topic type; Association types include: position, leadership, working relationships, etc.; Occurrence types include: corporation, Mick, etc.. The identified information will be describe separately under the element node stipulated by the XTM, generate XTM document, then to form three sub-topic maps respectively.

As the topic map has good scalability and can be integrated, it could be under the direction of the global model to establish a similarity analysis of three sub-themes graph, to combine topic maps of higher similarity and the same themes according to certain rules, and make the merger of the sub-topic map in accordance with the previous model since the end of the front into a global topic map. This approach can be easily achieved the delete, modify of by the underlying data, and even change the structure of the database library.

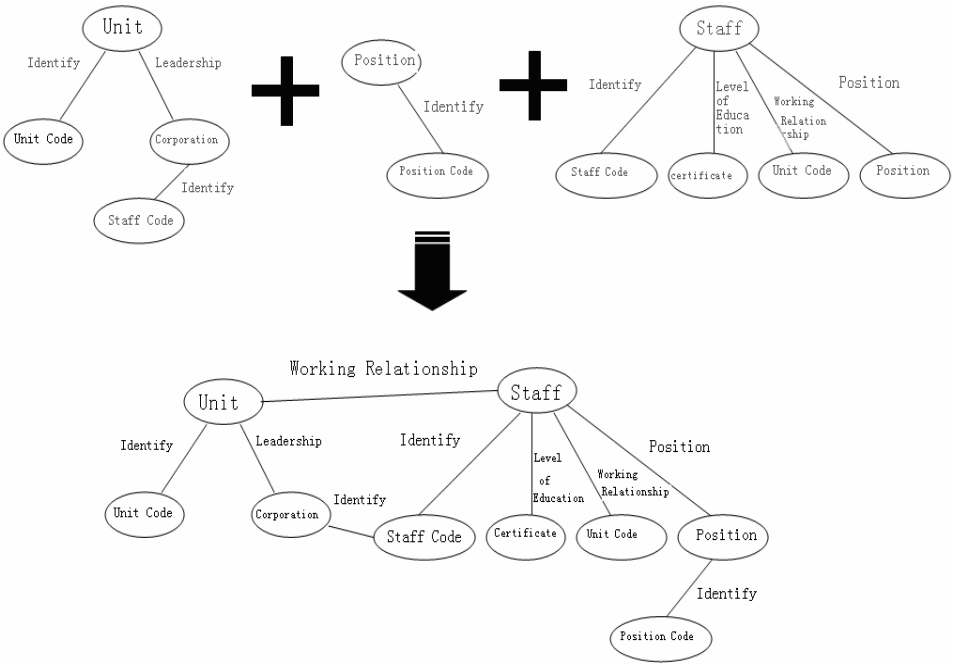


Figure 2. Conversion and Combination of Topic Maps.

3 Model Construction

The integration of data in multi-system is based on the unification of the heterogeneous database, while the interaction of data is the key of model design.

This model uses the OAI generate metadata aggregate, in order to achieve the standardization of heterogeneous data; by using reverse engineering in DBMS of government systems respectively, to make the combination of local sub-topic map based on the meta-data storage in accordance with the relevant rules, and then realize the interaction of data utilize the generated global topic map.

3.1 Model Overview

Figure 3 shows, the model designed as three modules: data processing module, topic map generation module, topic maps combination modules.

Data Processing Module. Mainly to parse the database, through the standardization of heterogeneous data in a database[6], database ER model extraction, metadata creation, metadata harvesting and other means to form a metadata warehousing, high-level application services.

Topic Map Generation Module. To use underlying formation meta-data storage, according to topic map templates and standard document that generate the topic map.

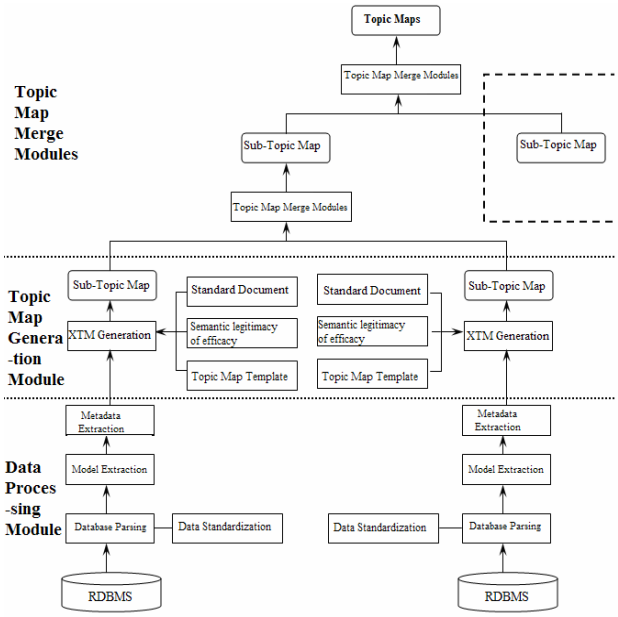


Figure 3. Data Integration Module Sructure

Topic Map Merge Modules. As the topic map has good scalability and reusability, global topic map can be divided into several thematic maps which can be con-

978-1-935068-76-1 © 2011 SciRes.

structed locally, and then use the bottom-and top-level-way merge, eventually constitute a general global topic map[7].

3.2 ER Model Extraction

The structure and semantic information of relational database are included in the database conceptual model, such as the ER (Entity-Relationship) model[8]. The key point of a relational database ER model extraction is the transition to the topic map. The use of semantic information in ER model guides the construction of topic map.

ER model extraction commonly use the database reverse engineering (DBRE, DataBase Reverse Engineering) method to complete[9]. Database reverse engineering means: access semantic information of the existing databases, and then transform relational model into a conceptual model (ER model), and finally use the easy-to-human understanding conceptual model to represent the results. In general, RDBMS data dictionary save the current "final" model state[10], this paper designed an ER Generator to extract ER model from the data dictionary, shown in Figure 4.

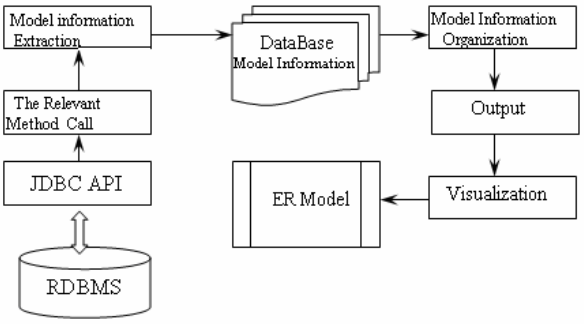


Figure 4. Entity Relationship Generator.

ER Generator use JDBC to connect to the RDBMS, get a Connection object which can be obtained a variety of RDBMS information. ResultSetMetaDataa class and DataBaseMetaData class show ways to implement the access of information. The main steps of using the ResultSetMetaData access the data sheet information and using DatabaseMetaData to obtain information in the database are as follows:

1)Establishing a Database Connection

When use ResultSetMetaData class and DataBaseMetaData class, need to use the JDBC API to connect the database. First, load the JDBC driver to connect the database and then call the the getConnection method of java.sql package in DriverManager class, to obtain a Connection object. The Connection object is an implementation of Connection in java. Sql package, indicating a connection with the database, the code is as follows:

```
Class.forName("oracle.jdbc.driver.OracleDriver");
String url="jdbc:oracle:thin:@202.114.37.152:1521:oemrep";
String user="usernm";
String password="huashi";
Connection conn=DriverManager.getConnection(url,user,password);
```

2) Obtain Entity's Attribute Information

The entity attributes of ER diagram of indicate by the column. It can get attribute information of the entity based on the read of column. ResultSetMetaData class provides getColumns method to get the result set of all the data columns information, the code is as follows:

```
Statement smt=conn.createStatement();
ResultSet rs=st.executeQuery("SELECT * FROM myTableName");
ResultSetMetaData rsMetaData=rs.getMetaData();
int numberOfColumns=rsMetaData.getColumnCount();
System.out.println("resultSet MetaData column Count="+ numberOfColumns);
for (int i=1;i<=numberOfColumns;i++) {
        System.out.println("column MetaData");
        System.out.println("column number"+i);
          System.out.println(rsMetaData.getTableName(i));
```

3) Access to entity-relationship information

The relationship in ER diagram to represent by FK, according to the reading to FK in procedures, you can get the relationship between entities. DatabaseMetaData class provides getImportedKeys methods get its references fields of other diagram. Through this interface can easily get the information related to FK, the code is as follows:

```
DatabaseMetaData dbMetaData=conn.getMetaData();
ResultSet rs=dbMetaData.getImportedKeys(conn.getCatalog(),null, "myTableName");
while(rs.next()){
  String fkTableName=rs.getString("FKTABLE_NAME");
  String fkColumnName=rs.getString("FKCOLUMN_NAME");
  int fkSequence=rs.getInt("KEY_SEQ");
  System.out.println("getImportedKeys():fkTableName="+fkTableName);
  System.out.println("getImportedKeys():fkColumnName="+fkColumnName);
  System.out.println("getImportedKeys():fkSequence="+fkSequence);
  }
```

Construction of ER-Generator in the laboratory used CA's CASE tools Erwin to support its operations.

3.3 OAI Metadata Harvesting Module

Because of the heterogeneous features of the data information in RDBMS in e-Government system, the OAI Metadata Harvesting Module uses the OAI Metadata Harvesting technology to extract the metadata.

OAI[11] Protocol for Metadata Harvesting is widely used in resource integration, cross-database searching, the subject information portals established, personalized service and other fields. The most important feature of OAI-PMH is that it uses the OAI-PMH which is relatively simple and independent of the application to make the convenient interoperability among the heterogeneous distribution of metadata becomes true.

This model use OAI technology to harvest the metadata though cross-database and the structure of this model is as shown in figure 5. This module consists of three steps:

1)At first all the data form the systemdatabase are transformed into digital ones and are stored in the digital repositories and the metadata database is established. Then this model does a structured meta-data organization for the metadata. After this procedure the data DP will be established and this DP provides meta-data information for the OAI Metadata Harvesting.

2)Metadata Collection This model uses meta-data collector to provide meta-data collection for data provider.

3) Standardization of Collected Metadata Processing:

a) Meta-data Filtering: The meta-data that are not meet the requirements would be filter.

b) Format Conversion of Metadata: Metadata ,after filtered, will transformed into a unified format.

c) Making Metadata Index to make metadata index for the processed metadata. After the formation of the standardized the metadata has been to collection of meta-data[12].

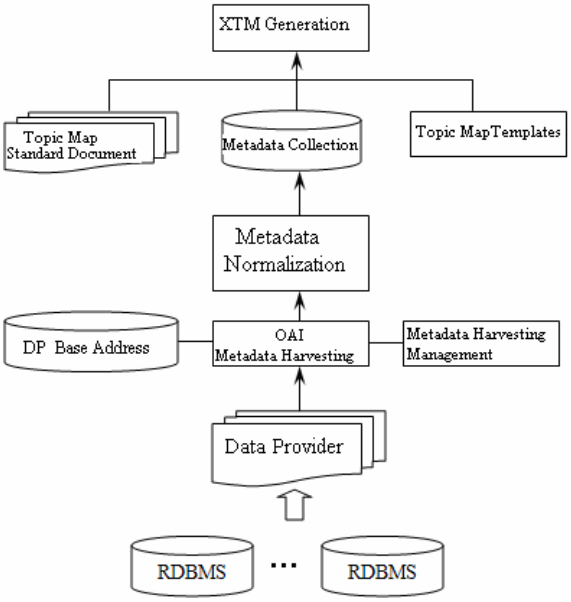


Figure 5. OAI Metadata Harvesting.

3.4 Generation and Integration of Topic Maps

1) Topic Maps Generation Module

As shown in figure 6, the topic maps generation module mainly completes the link of topic maps automatic generation; its core is that based on metadata warehouse in and according to the ER schema information, the topic maps generation module creates topic maps according to the standardization of the program, and then makes a check and dose visualization for the topic map. This module can be divided into three steps:

a)Topic maps preliminary generation

In this step this module uses OAI metadata harvesting module to harvest the metadata, and extracts data according to the standardization of data resources and data sets, and then generates the data warehouse. At last generates the preliminary topic maps according to certain standardization of topic maps and output it.

b)Calibration of Semantic Legitimacy

This step will check the semantic legitimacy for the preliminary topic maps according to the ER model and SCTM constraint language, and make it better[13].

c)Visualization and System Integration

After the topic maps generated, we should do the visualization for it. So that it can provide browse and search interface and interface which are based on topic maps, then express the topic maps in the way 2-d figure. At the same time we can use Java to develop it and make it can integrate to the e-Government system. So that the e-Government system index will be create.

Generally we can use Omnigator or StarTree[14] or other tools to do visualization for the topic maps. In order to make the topic maps interface more flexible, this model uses XML syntax and StarTree with XSLT[15] to realize visualization in the laboratory.

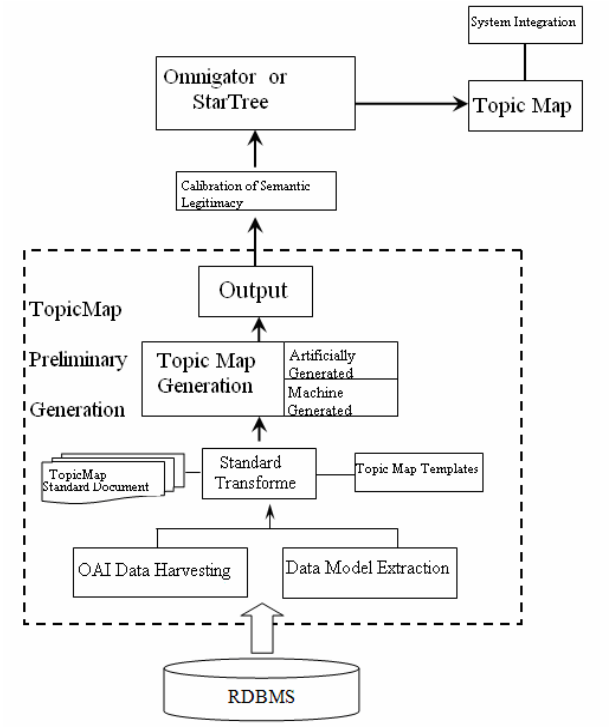


Figure 6. Topic Maps Generation.

2) Topic Maps merger Module

Because this model generates the topic maps by the bottom and top, expands gradually build, so the topic maps merger Module is the key technology.

Topic maps merger are the following principles:

a) when the merger of two themes are merged, representing the same topic and contact will be combined, and delete duplicate.

978-1-935068-76-1 © 2011 SciRes.

b) when the two themes are merged, the result is a topic; the characteristics of this topic are the udlte of two quondam elements[16].

When two themes are following conditions shall be regarded as one of the same concept description:

a)The two topic have one or more the same theme prompt.

b)The two topics have the same theme prompt in the same range defined .3 both of them are the specification of the same addressable things.

Based on the above principles, this paper uses class programming language to describe the three methods of merger in the topic maps:

a)Topic Merger：

```
IF (topic a in topic maps A)and (topic b in topic maps B)fit
THEN  generate (new topic c in topic maps C)
ELSE  copy(topic a in topic maps A)into topic maps C
      copy(topic b in topic maps B)into topic maps C
ENDIF
```

In this program the topic C should be the same as topic a and topic b, and the topic maps c is the merger of topic a and topic b.

b) Attribute Merger：

```
IF (attribute p of topic a)and (attribute q of topic b)fit
THEN generate(attribute p of topic C) OR  generate(attribute q of topic C)
ELSEIF (attribute p of topic a)and (attribute q of topic b) NOT fit
THEN generate(attribute p of topic C)AND generate(attribute q of topic C)
ENDIF
```

3) Relevance Merger:

```
IF (Relevance Ra(a1 ， a2)in topic maps A )AND (Relevance Rb(b1， b2)in topic maps B)
IF(topica1 in topic maps A)and (topicb1 in topic maps B) Merger
THEN  generate(topicc1 in topic maps C)
 generate(Ra(c1， a2) AND Rb(c1， b2))into topic maps C
ENDIF
ENDIF
```

4 Conclusion

This article is focus on the data integrate theme in Multi-E-Government systems. On the basis of deep analysis topic map and relational database, the article's writer raises to build a middle layer which called theme index layer between applied layer and data layer, and with the support of theme index layer to realize the heterogeneous database's data integrate. At the same time, this article constitute a model from relational database to topic map's generation and mergence by using technologies such as database reverse engineering, OAI technology and so on . This model can evade large calculation by heterogeneous database integration, and also overcome semantic information missing in database integration which has the advantages of low economic cost and powerful practicability.

In the future research, we could build a laboratory prototype system based on this model, and could do further research of the effective integrates in the model's referring topic map generation, mergence and e-Government system.

Acknowledgment

The paper is supported by National Natural Science Foundation of China "Research on the Knowledge Organization and Integration of Topic-Map-based E-Government Portal" (Project ID: 70873050) and Program for New Century Excellent Talents in University (Project ID: NCET-08-0788).

References

[1] CHEN Yifang, CHEN Qingkui and XU Fuyuan, Application Integration and Data Conformity Method in E-Government, Computer Engineering, 2008, 34(24), pp. 263-265.

[2] Hans Jochen Scholl, Electronic Government: Information Management Capacity, Organizational Capabilities, and the Sourcing Mix.Government Information Quarterly, 2006(23), pp.73-96.

[3] Abraham Silberschatz, Database system concept, 4th ed.,New York: McGraw-Hill Companies, 2002, pp. 921-1006.

[4] Steve Pepper, Navigating Haystacks and Discovering Needles: Introducing the New Topic Map Standard, http://www.ontopia.net/topicmaps/materials/mlangart.pdf, (Accessed Nov.11,2009).

[5] Martin S.Lacher and Stefan Decker, RDF, Topic Maps, and the Semantic Web, Markup Languages, Theory&Pratice, 2002,3(3), pp. 313-331.

[6] Wu Xiaofan, Zhang Lei, Hou Liang and Ding Qiulin, Heterogeneous Knowledge Integration Based on Topic Maps, New Technology of Library and Information Service, 2005(11), pp. 51-56.

[7] Steve Pepper and Graham Moore, XML Topic Maps (XTM) 1.0, http://www.topicmaps.org/xtm/ 1.0/ xtm1-20010806.html, (Accessed Nov. 23， 2009).

[8] Hainaut J-L, Research in Database Engineering at the University of Namur, SIGMOD Record, 2003, 32(4), pp. 124-128.

[9] Diego Calvanese, Maurizio Lenzerini, and Daniele Nardi, Unifying Class-Based Representation Formalisms, Jounal of Artificial Intelligence Research(JAIR), 1999 ,11, pp. 199-240.

[10] XU Zhuoming, and SU Wenping, Extraction of relational database schema information, Journal of Hohai University (Natural Sciences), 2005, 33(2), pp. 202-206.

[11] Carl Lagoze, Herbert Van de Sompel, Michael Nelson, and Simeon Warner, The Open Archives Initiative Protocol for Metadata Harvesting, http://www.openarchives.org/OAI/openarchives protocol.html, (Accessed Nov. 22, 2009).

[12] HU Minghui, Study on the OAI-based Cross-retrieval in Database, Document, Information & Knowledge, 2005(5), pp. 80-83.

[13] Mary Nishikawa, and Graham Moore, Topic Map Constraint Language (TMCL) Requirements and Use Cases, http://www.isotopicmaps.org/tmcl/requirements.html, (Accessed Nov. 22, 2009).

[14] http://highwire.stanford.edu/, (Accessed Nov. 1, 2009).

[15] James Clark, XSL Transformations (XSLT), http://www.w3.org/TR/xslt, (Accessed Nov. 12, 2009).

[16] Ma Jianxia, A Study of Application of Topic Map in Knowledge Organization, New Technology of Library and Information Service, 2004(7), pp. 11-16.

Hot Topics Detection of Web Public Opinion Based on Field-weighted Clustering Algorithm

Wei LU, Yi LIU, Rui MENG, Yingjie CHEN

The Center for the Studies of Information Resources of Wuhan University

Email: reedwhu@gmail.com, whu.louis@gmail.com, memray0@gmail.com, herochen04@gmail.com

Abstract: The research of information organization shows great significance when dealing with large amount of unordered web public opinion information. In this paper, we introduce a new organization method for web public opinion. We highlight the subject by weighting text fields which are more effective to express the theme, and then deal these fields with unsupervised clustering. By analyzing public opinion information, we realize the purpose of topic detection. In this method, the F-measure is more than 85%, which shows the effectiveness of letting clusters represent themes.

Keywords: web public opinion; field-weighted; hot topics detection; clustering algorithm

基于域加权聚类算法的网络舆情热点话题探测

陆伟，刘屹，孟睿，陈英傑

武汉大学信息资源研究中心

Email: reedwhu@gmail.com, whu.louis@gmail.com, memray0@gmail.com, herochen04@gmail.com

摘　要：面对自由无序的网络舆情信息，对舆情组织方式的研究体现出重要研究意义。本文提出一种网络舆情组织方法，以域加权的方式突出主题，通过聚类使得舆情信息以主题为依据组织起来，即对新闻主题或新闻事件有较强表达能力的域进行加权处理以突出该主题或事件,再以无监督自动化的方式对无序的网络舆情信息进行聚类。通过梳理舆情信息并发现热点话题，达到热点话题探测的目的。通过实验，类簇总体上都是基于主题或事件，可以表示为一个热点话题。本方法的 F-measure 值达到85%以上，表明了类簇代表主题的有效性。

关键词：网络舆情；域加权；热点话题发现；聚类算法

1 引言

中国互联网络信息中心(CNNIC)2010 年 7 月发布《第 26 次中国互联网络发展状况统计报告》[1]，《报告》显示，截至 2010 年 6 月底，我国网民规模达 4.2 亿人，互联网普及率持续上升增至 31.8%。网络正在成为人们获取与发布信息的主要渠道。

随着互联网的发展，网络逐渐成为舆情的最主要载体，通过天津社科院的刘毅[2,10]、王来华[8,11]和毕宏音[12]，华东师范大学的许鑫[3,9]等人对网络舆情理论以及应用前景的研究，笔者从舆情与舆论的概念上的区别[11]出发，从信息分析学的角度理解王来华教授关于舆情到舆论的转化过程[11]，可以认为舆论是对舆情进行信息组织整理后发现和得到的。

由于网络舆情的特殊性质[2]，以监督的手段进行组织是不符合实际情况的，所以笔者将在本文中介绍一种无监督聚类的方法实现对舆情的组织。

基金项目：本文为教育部人文社会科学规划项目“专家专长智能识别与检索系统实现研究”（项目编号：09yja870021）和教育部人文社科重点研究基地重大项目“面向细粒度的网络信息检索模型及框架构建研究”成果之一。

2 相关工作

关于网络舆论话题发现，国内外研究学者已经取得一定的研究成果。国外最具代表性的研究工作是话题检测与跟踪(Topic Detection and Tracking，简写为 TDT)。TDT 是一项面向新闻媒体信息流进行未知话题识别和已知话题跟踪的信息处理技术。James Allan[16] 和 Schultz[17]等人较早对话题探测与追踪问题进行研究，他们采用向量空间模型（简称为 VSM）描述报道的特征空间，根据特征在文本中的概率分布估计权重，利用余弦夹角衡量报道之间的相似性。此外，Leek[18] 和 Yamron[19]将参与检测的两篇报道分别看作一个话

978-1-935068-76-1

题和一篇报道，采用语言模型(简称为 LM)描述报道产生于话题的概率，并通过调换两篇报道的角色分别从两个方向估计它们的产生概率，最终的相关性则依据这两种概率分布。VSM 和 LM 存在的主要缺陷在于特征空间的数据稀疏性，通常解决这一问题的方法是数据平滑技术和特征扩展技术。

Kumaran[20]、James Allan[21]、Yiming Yang[22]和 Lam[23]等学者使用自然语言处理(NLP)技术辅助统计策略解决话题探测问题。其中最常用的 NL 技术是命名实体识别。比如 Kumaran[20]以 YimingYang 的分类方法为统计框架，将报道描述成三种向量空间，分别为全集特征向量、仅包含 NE 的特征向量和排除 NE 的特征向量。最终 Kumaran 对比了三种向量空间模型对新事件检测的影响，并验证 NE 极大地促进了事件之间的区分。

国内也有很多学者对热点话题探测与追踪问题进行研究，万小军[6]等提出了在线新闻主题探测方法，引入了创新阀值的概念，当文档与类簇向量间的相似度大于相似度阀值，但小于创新阀值时将该文档归入该类而不用更改类簇的中心向量。

中山大学黄晓斌与赵超[4]提出的一种通过文本挖掘处理网络舆情信息，并做分析预测的方法，通过代理软件抓取新闻数据然后人工筛选，再利用数据挖掘工具 Text Analyst 进行分析，该方法对人工的筛选需求比较严格。华东师范大学的王伟与徐鑫[5]提出一种在舆情分析中利用二次聚类的方法发现舆论热点，即通过先随机抽取网页样本集合进行聚类，选取关注的单个网页簇利用词的信息增益进行特征词抽取后，对全部网页使用抽取后的网页特征向量进行二次聚类，得到相关度较为纯粹的网络舆情网页集，从而用该网页代表一个舆论热点。南京大学的王昊与苏新宁[6]在研究中论证了条件随机场(Conditional Random Fields，CRFs)优于隐马尔可夫模型(Hidden Markov Model，HMM)和最大熵模型(Max Entropy Model，MEM)，并在此基础上提出一种基于 CRFs 的角色标注模型。利用该模型，对新闻或论坛讨论帖的标题进行角色标注，通过对人名出现次数的统计结合人名的背景进行舆情关注点的发现。

在聚类算法上，也有很多研究者提出了改进方案。文献[24]中提出了两层聚类的层次聚类算法，先找到每日热点簇，再利用增量聚类算法发现热点事件。文献[25][26]中结合基于密度的聚类算法和 K-means 算法的优点，改进了 K-means 算法中初始聚类中心选择的随机性问题。此外，基于主题的聚类方法也是热点新闻探测的一个研究方向，LDA、LSI 等主题模型方法在新闻聚类上的应用也是当前新闻主题发现的重要研究领域。

通过对研究现状的分析，笔者发现从信息技术的角度对网络舆情的研究较少，在已有的通过计算机技术的研究中人工参与的作用十分重要，虽然这样可以方便与用户交互，便于定制服务，但是自动化的全面的网络舆情监测与分析将会是未来研究的方向。笔者引入了 TDT 的思想，通过聚类的方式组织舆情信息，为使对舆情信息的组织基于主题，使用域加权的方式将主题相关域突显出来，最终达到发现网络舆情中热点话题的目的。

3 研究方法

3.1 基本聚类算法

考虑到在舆情分析应用过程中，需要对文本内容进行动态更新，本文采用 Single-pass 增量聚类算法[7]，即预设一个聚类阀值(Clustering Threshold)T，算法顺序处理输入的每篇文档，初始以第一篇文档为种子创建第一个类簇，对于每一篇输入的新文档，与以前生成的所有类簇进行相似比较，如果该文档与之前的某个类簇的相似度值大于聚类阈值时，那么该文档将属于该类簇；否则，将以该文档为种子创建一个新的主题类簇。

在计算相似度值时，本文主要采用在文本信息处理中常用的余弦相似度的公式：

$$\mathrm{Sim}(D_1, D_2) = \cos\theta = \frac{\sum_{k=1}^{n} W_{1k} \times W_{2k}}{\sqrt{\left(\sum_{k=1}^{n} W_{1k}^2\right)\left(\sum_{k=1}^{n} W_{2k}^2\right)}} \quad (1)$$

其中 W_{1k} 和 W_{2k} 表示两文本向量。

3.2 确定加权域

为使聚类的目标可以代表文本的主题或主要描述事件，需要确定待加权的域，使得这些域对主题或事件相关的词对文本相似性的贡献最大。从表达主题的关键词入手，笔者选择了标题域、全文本域和实体域三部分。

3.2.1 标题域

标题作为文本内容的直接体现，通常都包含主题

的关键词或者直接表明主题，并且标题的属性决定了标题的高度概括性与简短性。于是笔者仅以标题为对象，利用公式(1)对新闻数据聚类。

3.2.2 全文本域

笔者认为文本全文作为主题全部信息的载体，排除一些表述特别隐晦的文本，全文中必定会包含主题，这样可以较好的解决标题中信息不足的问题。

3.2.3 实体域

考虑到人们在表达舆情观点时，通常都是符合思维逻辑，即表述会包含 5W1H（Who 人物,When 时间,Why 目的, Where 地点, What 对象,How 方法）中的若干项，而通常目的、对象或方法即为表述的主题，所以通过对人物，地点和时间的关注可以发现较好的主题，比如对相关的主题事件的描述中会出现相同的人物或者地点等。文献[20]也说明了实体对于文本主题的作用。于是笔者以实体为对象，用公式(1)经行聚类。

3.3 域加权方案

通过加权的方式，计算相似程度时就具有了一定的容错能力，即使某一部分权值很低甚至为 0，但是通过其他部分的权值依然可以判断是否相似，是否属于同一主题或描述同一事件。并且通过对上述 3 种域的加权，突出了主题相关的词，使得聚类基于主题或主要描述事件。

在具体加权方式上，陆伟教授曾提到过 3 种计算文档权重的方法[13]，由于第一种方法过于简单，在本文 2.2 节确定加权域时分别进行了 3 次聚类就属于这种方式，这样的计算过于朴素，存在不少问题；而第三种域词频加权法，笔者认为该方法应该在监督环境下进行，对于本文需要无监督的聚类不太适用。所以本文采取第二种方案用各个域权重得分之和作为文档权重。

计算标题相似度、全文相似度以及实体相似度，再对 3 个相似度以一种权值组合的数值代表两文档的实际相似度，加权公式如下：

$$Sim = \alpha * Sim(title) + \beta * Sim(content) + \lambda * Sim(entity) \quad (2)$$

其中 Sim(title)表示标题相似度，Sim(content)表示全文相似度，Sim(entity)表示实体相似度。α，β，λ 分别表示其所占的权重，且 $\alpha + \beta + \lambda = 1.0$。

*Sim(title)*即标题的相似度，由于标题信息量有限，所以对标题向量化时单独处理。在去除停用词的基础上建立标题的动态词表，由此词表建立向量，在只考虑特征词共现程度的基础上用公式(1)计算得到 *Sim(title)*。

全文相似度 *Sim(content)*是由全文本域组成的向量经过公式(1)计算得出。

Sim(entity) 即实体相似度，实体信息是由 ICTCLAS 分词器[15]中实体标注功能得出，选取其中的 NR(人名)，NS(地名)，NT(机构名)作为实体特征。实体相似度计算基于词共现方法，当两文本具有相同的实体特征时，实体相似度权值 value 提升 1.0，最后用两篇文档中出现的实体最大数对求得的相似度进行归一化处理，即：

$$similar = \frac{value}{maxnum} \quad (3)$$

value 为最终累和得到的权值，maxnum 为两文档中共含有的实体数目。

在判断两文本相似时，本文认定 1.0 为完全相同，即全文相似度、标题相似度、实体相似度都为 1.0。

计算相似度的方法有多种，通常在文本向量计算中使用的是余弦相似度公式，在几何中常用的则是欧几里得距离。使用余弦相似度计算时不会放大数据对象重要部分的作用，而欧几里得距离则在一定程度上放大了较大元素误差在距离测度中的作用[28]。考虑到本文域加权的方式在热点事件发现中的突出主题作用，所以本文在总体聚类算法实现的基础上，仅换用距离计算公式，通过实验三验证欧式距离与余弦相似度在域加权聚类中产生的影响。

3.4 热点话题发现

将混乱的舆情信息用基于域加权的聚类算法组织之后，具有相同或相近特征的文章聚为一类，形成了若干类簇。由于在聚类过程中突出了新闻的主题特征，所以每一类都应当是由某一新闻主题或事件的相关新闻组成的簇。这样就可以从舆情研究的角度加以利用，对某一主题或事件的新闻数目在时间度量上的演化情况以及在信息来源（站点、博客、BBS 等）上的演化情况进行分析（对在时间轴上的演化再次投影在某个来源站点上细化研究）达到热点话题探测的目的。本文认定最后聚类结果中的每个类簇为一个话题，并根据类簇的大小确定热点话题。

3.5 其他细节处理

3.5.1 抽取特征词

978-1-935068-76-1 © 2011 SciRes.

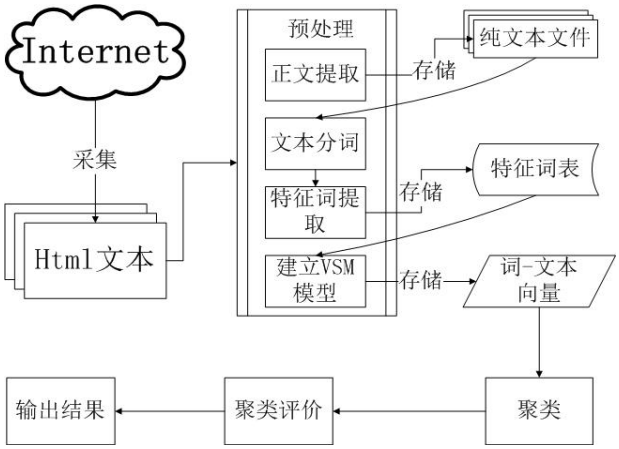


Figure 1. Performing workflow
图 1. 实现流程图

因为维度过高会导致“维灾难”[29]，所以要抽取特征词进行降维，缩小主题相关词的探查范围，在一定程度上避免因为向量过度稀疏或者冗余信息导致结果不好的问题。

考虑到实时抽取特征向量对聚类效率的影响，笔者通过对一个大语料库处理得到特征词列表，该特征词表存有特征词以及 IDF，该 IDF 的计算是由 Okapi BM25 排序函数中计算 IDF 的公式[27]给出，即：

$$\mathrm{IDF}(q_i)=\log\frac{N-n(q_i)+0.5}{n(q_i)+0.5} \tag{4}$$

其中，N 为数据集中的全部文档数，$n(q_i)$为数据集中包含特征词 q_i 的文档数目。

假设在生成该特征词表时使用的语料库足够大，则该特征词表可以包含所有汉语中具有语用或语义信息的词语，所以当有新数据集需要聚类时，增量的重新生成一个特征词表与原特征词表相比特征词数量的改变极其微小，特征词的 IDF 值的变动极小对结果的影响可以忽略不计。这样只需计算每篇文档中包含的特征词的 TF，再从特征词表中查得 IDF 值，以 TF*IDF 为权值生成文档向量。为了突出实体（人名，机构名等）在聚类中的影响，计算 TF_IDF 权值时根据 ICTCLAS 标注出的词性，给不同的词赋予不同的权值，例如：nr（人名）权值为 8、nt（机构）权值为 6。在实验二中将给出实验结果，以确定静态特征词表对聚类结果的影响。

3.5.2 建立 VSM 模型

根据特征词表生成词-文本向量：

$$Vector(i)=[V(1),V(2),\cdots V(m)] \tag{5}$$

其中 V(m)为第 m 个特征词的权值，V(m)的计算公式如下：

$$V(i)=tf(w_i)\times idf(w_i) \tag{6}$$

V(i)为第 i 维的权值，w_i 为第 i 维表示的特征词，$tf(w_i)$为词 w_i 在一篇文档中的词频，$idf(w_i)$为词 w_i 在一个文档集中的逆文档频。

3.5.3 聚类结果输出

聚类完成后，为了给每个聚类进行恰当的描述，从每个聚类中抽取出一组可以表征该聚类的关键词，对于该关键词的抽取，可以在聚类中共现度最高的词中选择 TF-IDF 权值最高的若干词。这样用这组关键词就可以大致描述该聚类所描述的事件。

3.5.4 聚类评价

完成上述整体过程后，为了对聚类的效果进行衡量，需要一个评价体系。本文采用外部评价法，首先抽取部分数据，手工标注出所属事件类别，根据评价结果判断聚类优劣。评价方法选用 F 度量值(F-measure)的评价方法[14]，该方法是通过聚类前的分类标记与聚类后的聚类标记建的查全率与查准率来评价聚类效果的。即首先给抽样中的文本按其所讲述不同事件标记不同的类别标号，在聚类之后，赋予新闻相应的聚类标号，统计聚类中某分类号的数目，根据信息检索中查准率(precision)与查全率(recall)的思想来进行评价。一个聚类 j 及与此相关的分类 i 的 precision 与 recall 定义为：

$$\mathrm{P}=\mathrm{precision}(\mathrm{i},\mathrm{j})=\frac{\mathrm{N}_{\mathrm{ij}}}{\mathrm{N}_{\mathrm{i}}} \tag{7}$$

$$\mathrm{R}=\mathrm{recall}(\mathrm{i},\mathrm{j})=\frac{\mathrm{N}_{\mathrm{ij}}}{\mathrm{N}_{\mathrm{j}}} \tag{8}$$

其中 N_{ij} 是在聚类 j 中分类 i 的数目；N_j 是聚类 j 中所有对象的数目；N_i 是分类 i 中所有对象的数目。分类 i 的 F-measure 定义为:

$$\mathrm{F}(\mathrm{i})=\frac{2\mathrm{P}\times\mathrm{R}}{(\mathrm{P}+\mathrm{R})} \tag{9}$$

对分类 i 而言，哪个聚类的 F-measure 值高，就认为该聚类代表分类 i 的映射。换句话说，F-measure 可看成分类 i 的评判分值。对聚类结果来说，其总 F-measure 可由每个分类 i 的 F-measure 加权平均得到：

$$\mathrm{F}=\frac{\sum[|\mathrm{i}|\times\mathrm{F}(\mathrm{i})]}{\sum|\mathrm{i}|} \tag{10}$$

其中|i|为分类 i 中的文本数。

4 实验描述与结果评价

本文测试实验所用数据是以“武汉大学”为搜索关键词在元搜索引擎中得到的结果进行采集。

4.1 实验一

通过对比实验，仅标题相似度、仅全文相似度、仅实体相似度、标题+全文相似度和标题+全文+实体相似度的聚类，通过聚类评价的 F-measure 值以及平均查全率及平均查准率来评估聚类。

首先在原始数据中挑选了 8 个新闻事件，其中有：1 蚁族报告 2 武大割喉案 3 台州官员子女高考加分事件 4 传统仪式过端午节 5 华科大根叔演讲 6 预科诈骗调查 7 武汉蓝藻爆发 8 汪晖学术剽窃事件 9 厅官妻子被打事件。根据事件抽取了新闻 326 篇，再向其中掺入无关新闻 54 篇，编号为 0 类，共组成共 380 篇文档的数据集，进行测试，结果如下表：

Table 1. Result of Contrast test for weighting scheme

表 1. 加权方案的效果对比试验结果

文本类别标号	文本数	标题			全文			实体			标题+全文			标题+全文+实体		
		查准率	查全率	f-measure	查准率	查全率	f-measure	查准率	查全率	f-measure	查准率	查全率	f-measure	查准率	查全率	f-measure
0	54	0.259	0.06	0.098	0.074	1	0.138	0.944	0.135	0.237	0.074	1	0.138	0.074	1	0.138
1	64	0.953	0.263	0.412	0.953	1	0.976	1	0.169	0.29	0.984	1	0.992	0.984	1	0.992
2	70	0.771	0.232	0.357	0.971	0.66	0.786	1	0.186	0.313	1	0.986	0.992	1	0.986	0.993
3	37	0.649	0.103	0.178	1	1	1	1	0.098	0.179	1	1	1	1	1	1
4	46	0.674	0.134	0.223	0.956	0.978	0.967	1	0.122	0.217	0.957	0.978	0.966	0.957	0.978	0.967
5	30	0.567	0.073	0.129	0.967	0.967	0.967	1	0.079	0.147	0.967	0.967	0.967	0.967	0.967	0.966
6	18	0.667	0.5	0.571	1	1	1	1	0.047	0.091	1	1	1	1	1	1
7	6	0.667	0.017	0.034	1	1	1	1	0.015	0.031	0.833	0.135	0.233	1	1	1
8	23	0.696	0.069	0.125	0.957	1	0.978	1	0.061	0.115	0.957	1	0.978	0.957	1	0.978
9	32	0.5	0.571	0.533	1	0.311	0.474	1	0.084	0.156	1	0.864	0.927	1	1	1
评价				0.283			0.782			0.221			0.849			0.867

标题列的结果并不理想，经过分析，认为问题可能在于：

1）标题不准确，页面提取以及文本信息预处理过程中识别不准确，标题完全错误，与主题无关。

2）标题信息量不足，标题本身概括程度太高，并且概括方式以能使读者理解为目的而形式(用词)多样，主题直接相关信息尤其是关键词较少。

3）分词器效果产生的影响，由于标题中关键词未识别或错误识别，导致标题的文本向量极其稀疏。

全文列的结果，通过与标题列结果对比，发现聚类结果有明显的提高。但是在检验聚类结果时依然发现了一些错误，笔者认为原因是：虽然全文是原始信息，但是不同的文本包含的信息量不同，可能包含有大量冗余信息。比如：仅包含主题关键信息的短文本与对同主题详细扩展描述的长文本计算相似度时，笔者希望相似度能大于阈值，划分为同一类，但是由于短文本生成的向量过于稀疏，导致关键词对主题的表达不能显现出来，从而相似度低于阈值。

该聚类结果表明标题、全文、实体对文本相似度有影响。但是由于是 Single-pass 增量聚类算法，所以当某文本包含 2 个主题时，会导致这两个主题的类边界模糊，甚至混合。仅标题和仅实体的情况下就含有这种情况。实验结果表明标题和实体对文本相似性确实有影响的，所以在全文的基础上加入标题加权后，F-measure 值明显提升，加入实体的影响后，F-measure 值又有所提升即是证明。但是加入实体后的 F-measure 值提升幅度比预计中要小。

4.2 实验二

通常涉及 VSM 模型都需要为目标数据集生成特征词表，在增量式聚类中加入新文本就需要重新生成词表。本实验特征词表的策略是静态特征词表，即由大量数据集生成一个静态的特征词表，固定 IDF 值，对任意数据集聚类时都利用该静态特征词表来生成文本向量。下列实验证明不同特征词表对结果的影响。

另标注 240 篇不同事件类别的文档，其中有：1 学

978-1-935068-76-1 © 2011 SciRes.

者证实曹操墓 2 方舟子炮轰刘维宁 3 白沙洲 4 曙光学子 5 曝光体检枪手 6 全球气温上涨 1.1 度 7 数字鸿沟 8 易中天与李泽厚 9 北极科考。

共 9 个事件类别，共 200 篇，另加入无关新闻 40 篇为噪声。

Table 2. Feasibility checking of fixed words IDF
表 2. 对固定词 IDF 的可行性验证

数据集数	特征库文本数	特征词数	f-measure
240	1416	10804	0.8503
240	5082	9741	0.8434
240	6497	44824	0.8503
380	1416	10804	0.8613
380	5082	9741	0.8613
380	6497	44824	0.8667

其中 380 篇文档的数据集是从 5082 篇文档的特征库中抽取出来标注的，而 240 篇文档的数据集是从 1416 篇文档的特征库中抽取出来标注的，其中 6497 维的特征库为两特征库的并集。

该实验表明在有足够大量文本后，得到的含有 IDF 值的特征词表，对增量聚类的效果影响很小，所以使用静态的特征词表对于聚类效果没有很大影响，这样对聚类效率的提升就很明显。

4.3 实验三

将文本转化为词-文本向量，计算向量的相似度来表达词之间的共现的程度，以此表明文档的相似程度。但是对于文档向量间相似度的度量方式有很多种，在信息处理中较常用的是余弦相似度以及欧几里得距离。虽然根据余弦定理等数学公式可以推导得到余弦相似度与欧几里得距离对应关系，但是张宇等[28]认为欧几里得距离在一定程度上放大了较大元素误差在距离测度中的作用。而余弦相似度规范化了向量的长度，这意味着在计算相似度时，不会放大数据对象重要部分的作用[30]。众多相似度度量的方法在侧重上区别较明显，此处我们只对比余弦相似度以及归一化的欧几里得距离的出的聚类结果。

Table 3. The impact of Cosine distance and Euclidean distance on Domain weighting Algorithmic
表 3. 余弦距离与欧式距离对域加权算法的影响

文本类	文本数	余弦相似度			欧几里得距离		
		查准率	查全率	f-measure	查准率	查全率	f-measure
噪声杂类	54	0.074	1.0	0.138	0.056	1.0	0.105
蚁族报告	64	0.984	1.0	0.992	0.484	1.0	0.652
武大割喉	70	1.0	0.986	0.993	1.0	0.986	0.992
高考加分	37	1.0	1.0	1.0	1.0	1.0	1.0
端午节	46	0.957	0.978	0.967	0.891	0.976	0.932
根叔演讲	30	0.967	0.967	0.966	0.567	0.944	0.708
预科诈骗	18	1.0	1.0	1.0	0.889	1.0	0.941
蓝藻暴发	6	1.0	1.0	1.0	1.0	1.0	1.0
学术剽窃	23	0.957	1.0	0.978	0.957	1.0	0.978
厅官妻被打	32	1.0	1.0	1.0	1.0	1.0	1.0
评价				0.867			0.778

其中文档 Wi 与 Wj 的归一化欧几里得距离为：

$$\text{Euclidean} = \frac{\sqrt{\sum\left(\frac{w_{ik}}{|w_i|}-\frac{w_{jk}}{|w_j|}\right)^2}}{\sqrt{2}} \qquad (11)$$

其中 W_{ik} 为文档 W_i 的第 k 个维度的值，这样计算得到的距离之是一个在 0-1 之间的数，0 代表完全相似，1 代表完全不相似。

从上述结果可以看到欧几里得距离的聚类结果明显比余弦相似度的 F-measure 值小，虽然在某些类别上查准率、召回率以及 F-measure 是完全不变的，但是在某些类上明显缩小。这就是对数据重要部分或者说权值较大的词的作用是否放大产生的效果，欧式距离把一些权值较大的关键部分的误差放大了，所以导致聚类时特征词在事件上的投影很粗糙甚至不明显，聚出的类别相比于事件类别，更像是领域类别，而余弦相似度可以保证词的存在不被影响到，基于词的共现，这样能更好的表现出主题事件的相似。

5 结束语

通过 3.1 中的对比实验，表明这种域加权的聚类方式是有效的，在未标注数据集测试，通过人工对聚类后的文档内容识别分析，可以认为本文方法实现了对新闻事件的聚类。

但是，本方法仍然存在些缺陷：首先从 3.1 中可以看出虽然考虑到实体的影响因素，但是对于聚类评价的 F-measure 值的提升并不大，所以现阶段对实体处理方式后续需要改进。其次，本文在文本信息处理方面使用了分词的技术，利用 ICTCLAS。虽然 ICTCLAS 的准确率在通常文本中表现已经很好了，但是在一些专业的领域或者面对专有的词语效果不好，并且 ICTCLAS 的分词在很大程度上还是需要词表的支持，虽然可以用户自定义扩展，但是为达到聚类效果的稳定，需要克服这些问题。最后，本方法在超大规模的数据集上效率是否在可接受范围内，还需要后续实验检验。

References (参考文献)

[1] 26th China Internet Development Statistics Report. http://research.cnnic.cn/html/1279171593d2348.html

[2] LiuYi. On the concept, features, expression and communication of network public opinion [J]. *Theory world*,2007(1).

[3] Xu Xin, Zhang Chengzhi, Li Wenjing. Domestic Review and Prospect on Network Public Opinion Research[J]. *Review and-Comment*.2009,32(3):23-25.

[4] Huang Xiaobin, ZhaoChao．Applicationof Text Mining on the Network Public Opinion analysis[J]． *Information Science*, 2009(1).

[5] Wang Wei, Xu Xin．Online Public Opinion Hotspot Detection and Analysis Based on Document Clustering[J]．*New Technology of Library and Information Service*，2009(3).

[6] Wang Hao, Su Xinning. Model for Person Name Recognition Based on Role Labeling Using CRFs and Its Application to Web Opinion Analysis[J]. *Journal of the China Society For Scientific And Technical Information*，2009(2).

[7] R Papka. On-line new event detection,clustering,and tracing:[PhD dissertation]．MA： University of Massachusetts Amherst, 1999.

[8] Wang Lai Hua. Public Opinion Research: theory, method and practical hotspot[M]. Tian Jin: Tianjin academy of social science press ,2006,6.
王来华.舆情研究概论：理论、方法和现实热点[M].天津：天津社会科学院出版社,2006.6.

[9] XU Xin, ZHANG Cheng-zhi. Research on Application and Analysis of Internet- Mediated Public Sentient[J]. *Information Science*,2008(8):1194-1200,1204.
许鑫,章成志．互联网舆情分析及应用研究[J].情报科学,2008(8) :1194-1200,1204.

[10] LIU Yi. Application of Content Analysis Method in Analysis of Net-Mediated Public Sentiment[J]. *Journal of Tianjin University (Social Sciences)*,2006(7):307-310.
刘毅.内容分析法在网络舆情信息分析中的应用[J].天津大学学报:社会科学版，2006(7) :307-310.

[11] Wang Lai Hua. The Conversion and Influence of internet public sentiment and public opinion [J]. *Tianjin Social Sciences*, 2008 (4): 66-69.
王来华.论网络舆情与舆论的转换及其影响[J].天津社会科学,2008(4):66-69.

[12] Bi Hong Yin. Main Features of Netizens Network Public Opinion [J]. *Guangxi Social Sciences*, 2008(7):166-169.
毕宏音.网民的网络舆情主体特征研究[J].广西社会科学,2008(7) :166-169.

[13] Lu Wei, Xia Li Xin. Realization of XML Information Retrieval Based on Okapi[J]. *The Journal of the Library Science in China*, 2006(4): 60-64.
陆伟，夏立新.基于 okapi 的 XML 信息检索实现研究[J].中国图书馆学报，2006(4):60-64.

[14] Yang Yan, Jin Fan, KAMEL Mohamed. Survey of clustering validity evaluation [J]. *Application Research of Computers*, 2008, 25(6): 1630-1632,1638.
杨燕，靳蕃，KAMEL Mohamed. 聚类有效性评价综述[J].计算机应用研究，2008,25(6): 1630-1632, 1638.

[15] ICTCLAS Tokenizer. http://ictclas.org/ictclas_introduction.html

[16] J Allan, V Lavrenko, and R Swan. Explorations within topic tracking and detection [A]. In: Topic Detection and Tracking: Event-based Information Organization [C]. Kluwer Academic Massachusetts, 2002, 197-224.

[17] J M Schultz and M Y Liberman. Towards an universal dictionary for multi-language IR applications [A] . In: Topic Detection and Tracking : Event2based Information Organization [C]. Kluwer Academic: Massachusetts , 2002 , 225-241

[18] J Yamron,L Gillick,P van Mulbregt , and S Knecht .Statistical models of topical content [A] . In: Topic Detection and Tracking : Event2based Information Organization [C] . Kluwer Academic: Massachusetts , 2002 ,115-134.

[19] Leek T,Schwartz R M.,and Sista S. Probabilistic approaches to topic detection and tracking [A] . In: Topic Detection and Tracking: Event-based Information Organization [C]. Kluwer Academic : Massachusetts ,2002 , 67-83.

[20] G Kumaran and J Allan. Text classification and named entities for new event detection [A]. In: Proceedings of the SIGIR Conference on Research and Development in Information Retrieval [C]. Sheffield,South Yorkshire : ACM , 2004 , 297-304.

[21] J. Allan, H J in, M Rajman, C Wayne, GD , L V , R Hoberman, and D Caputo. Topic-based novelty detection [A]. In: Proceedings of the Johns HopkinsSummer Workshop [C]. CLSP, Baltimore, 1999.

[22] Y Yang , J Carbonell , C J in. Topic-conditioned novelty detection [A]. In: Hand D, et al. Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining[C]. New York: ACM Press, 2002, 688-693.

[23] W Lam, H Meng, K Wong , and J Yen. Using contextual analysis for news event detection [J]. *International Journal on Intelligent Systems*, 2001, 16(4):525-546.

[24] Liu Xing-xing, He Ting-ting, GONG Hai-jun, CHEN Long. Design of Hot Web Event Detection System [J]. *Journal of Chinese Information Processing*, 2008(6):80-85.
刘星星,何婷婷,龚海军,陈龙. 网络热点事件发现系统的设计[J]. 中文信息学报,2008(6):80-85.

[25] LI Ruo-peng, Li Xiang, Lin Xiang, Li Jian-hua. Discovering Information Hotspots on Initiative over Internet Based on DK Clustering Algorithm [J]. *Computer Technology and Development*,2008(9)1-4.
李若鹏, 李翔, 林祥, 李建华. 基于 DK 算法的互联网热点主动发现研究与实现[J].计算机技术与发展.2008.18(9):1-4.

[26] Huang Yu-dong, Li Xiang, Lin Xiang. Research and Application on Active Discovery Technique of Internet Media Information Hotspots[J]. *Computer Technology and Development*, 2009.19(5): 1-4,187.
黄宇栋,李翔,林祥.互联网媒体信息热点主动发现技术研究与应用[J]. 计算机技术与发展. 2009.19(5): 1-4,187.

[27] Robertson,S.E.and Steve Walker.Some simple effee—tive approximations to the 2-Poisson model for probabilisticweighted retrieval.Proceedings of the 17th annualinternational ACM SIGIR conference on Researeh anddevelopment in information retrieval.1994, 345-354.

[28] Zhang yu, Liu yudong, Ji zhao. Vector similarity measurement method [J]. *Technical Acoustics*.2009.28(4).

[29] Introduction to Data Mining.[DB/OL]. http://book.csdn.net/bookfiles/327

[30] TIAN Runtao,XIE Peishan. Study on the standardization of sinai-larity evaluation method of chromatographic fingerprints (PartI) [J]. Traditional Chinese Drug Research & Clinical Pharmacology, 2006, 17(1): 40-42.

978-1-935068-76-1 © 2011 SciRes.

Bisecting Tensor Decomposition to Discover Theme Changes over Time

Pradeep DHANANJAY, Weijia XU*, Maria ESTEVA, Victor EIJKHOUT
Texas Advanced Computing Center
The University of Texas at Austin, Austin, Texas, USA
Email: {prd329, xwj, maria, eijkhout}@tacc.utexas.edu

Abstract: With the goal of helping archivists analyze document collections we developed a bisecting tensor decomposition method to detect theme changes and relationships across time and provenance automatically. The results show that the method highlights relevant features in the collection and traces their evolution over time. This is consistent with the way in which archivists describe collections for access purposes.

Keywords: tensor analysis; digital archives; text mining

1 Introduction

Archives descriptions or finding aids, provide comprehensive panoramas of collections and function as guides for access. The process of analyzing and describing collections (both may go in tandem) are traditionally done by hand, following professional standards. In these descriptions, archivists highlight relevant components in a sequence that goes from describing the whole collection to its parts and, less frequently, the individual documents.

Descriptions of groups of documents are made in relation to their author or originating agency (provenance), to time, and to the collection's structure. The latter is the criteria used by the author to group his or her documents, which informs about functional and content relationships that may exist between documents and allows navigating the collection.

And yet, the native directory structure of a digital collection may or may not bear cohesiveness in themes, provenance, or time. Moreover, when the collection is large and complex it is practically difficult to uncover variations or relationships between groups of documents manually. We developed a bisecting tensor decomposition, to automatically uncover relationships and differences between groups of documents in context with time and provenance. We evaluated the results and explain how they may help archivists in their analysis and descriptive tasks.

Tensor analysis encompasses methodologies that help uncover differences and relations in multidimensional datasets. Tensors are multidimensional arrays, each of which represents a different aspect of the data. Since the tensor is able to accommodate multiple information dimensions, the data can be analyzed in relation to each of these dimensions independently. For example, using tensors to analyze a collection we may find themes associated with provenance, and follow these relationships across time and across different groups of documents.

2 Related Work

Tensor decomposition is a mathematical problem known as factorization. There are two common methods for tensor decomposition, parallel factor analysis (PARAFAC) [1] and Tucker Decomposition [2]. Tensor analysis has been applied in the fields of signal processing, chemoinformatics and bioinformatics. For a recent overview and a survey of tensor decomposition methods and applications, the readers can refer to [3,4].

The application of tensors to text mining is relatively new [5-9]. Tensor analysis is analogous to the singular value decomposition (SVD) method, which is used extensively for text mining. While SVD is an effective method to uncover the main feature in a text corpus, it operates on two-dimensional (term-document) matrices and cannot utilize other available information in the documents such as timestamps, provenance, names, places etc. In contrast, tensor analysis can factor additional information dimensions into the analysis.

In [5], a user, keyword, and time tensor was constructed to identify separate conversations in an online chat room. Similar three-dimensional tensors were applied to identify discussion threads in the Enron email collection [7]. Both to personalize web searches [9], and to improve page ranking in web searches [8], tensors were constructed using different combinations of query terms, anchor text, and webpages. Tensor analysis was also used for cross language information retrieval, where the third dimension was based on multilingual terms [10]. The tensor model was also used to extend the vector space model, showing considerable improvement for analyzing large amounts of documents [11].

In turn, the use of automated methods specifically for archival analysis and description tasks is not frequent. Previously, the authors studied personal electronic record-keeping practices using visualization and RDBMS [12] and investigated text alignment methods to find stories as points of access to chaotic document collections [13]. In this project, we investigate the application of

PARAFAC decomposition in a bisecting fashion. To our best knowledge this is a novel implementation of the PARFAC method.

3 Method

The number of information dimensions in a tensor gives the order of the tensor. A vector can be treated as a first order tensor and a matrix as a second order tensor. In this research we build a third-order tensor (term-document-time) and perform bisecting PARAFAC to predict relationships and changes among these three dimensions. We first explain PARAFAC decomposition and later how to implement it iteratively in a bisecting fashion.

3.1 Standard PARAFAC

PARAFAC is a multi-way tensor decomposition also known as CANDECOMP-PARAFAC (CP) model [14]. Given a third-order tensor **X** of size $(m \times n \times p)$ and a desired approximation rank r, the PARAFAC model approximates **X** as a sum of r rank-1 tensors formed by the outer (tensor) product of three vectors. It is convenient to group each set of r vectors together as matrices **A**, **B** and **C**, which we call factor matrices. The PARAFAC model may be extended to N-way data. All columns of the factor matrices are normalized to unit length. The accumulated weight is stored in a vector $\boldsymbol{\lambda}$. Following is a general form of the decomposition [7]:

$$x \approx \sum_{l=1}^{r} \lambda_l (a_l \circ b_l \circ c_l)$$

A common approach to fitting the PARAFAC model to data is an alternating least squares (ALS) algorithm.

In this research we are interested in exploring the thematic changes and commonalities that may exist in a collection of documents over time. Therefore, each triad $\{\mathbf{a}_j, \mathbf{b}_j, \mathbf{c}_j\}$, for $j = 1,...r$, defines relevance scores for each dimension: terms, document, and time in a decomposed tensor as shown in Figure 1. The value of the document dimension of each decomposed rank-1 tensor can be considered as a *theme* (cluster of documents) separated out from others, and the two remaining vectors indicate where these themes are prominent along the two dimensions. [1]

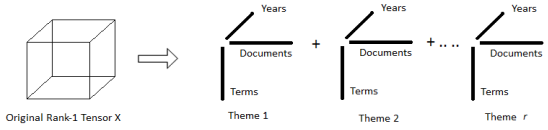


Figure 1. Illustration of tensor decomposition with approximation rank *r*.

[1] Even though each decomposed tensor can be treated as a mathematical separation of a subset of documents based on a set of word distributions, a decomposed tensor may not bear a cohesive content.

3.2 Bisecting Tensor Decomposition:

With PARAFAC, r, the number of rank-1 tensors that the decomposition should produce is an input parameter to the decomposition. Because there is no proven model to help decide this number, the value of r is chosen empirically based on the dataset and the information that needs to be extracted.

Bisecting tensor decomposition does not require a priori knowledge of r. First, the original tensor is broken down to two tensors. Bisecting decomposition is then applied recursively to the resulting tensors. At each step of the decomposition, metrics are evaluated to decide if further decomposition is required (Figure 2). We used two metrics, each with a different objective.

1. Sparseness of a tensor $S(M)=\frac{M_0}{M_0+M_1}$
   Where M_0 and M_1 are respectively the number of occurrences of zero values and non-zero values in the tensor.
2. Overlap between two tensors

$$O(M_a, M_b)= \frac{M_{11}}{M_{11}+M_{10}+M_{01}}$$

Where M11 represents the number of elements that have non-zero values in both tensors. And M10 (M01) the number of elements that have non-zero value in Ma (Mb), and zero value in the Mb (Ma).

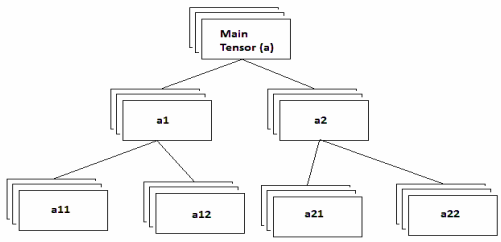


Figure 2. Graph of the top-down bisecting tensor decomposition

The sparseness of a tensor indicates the amount of information it carries. The sparser the tensor, the less information it contains. Since the information contained in the original tensor is split, we expect the resultant decomposed tensors to be even sparser. At the same time, we expect the information contained within each decomposed tensor to be more specific. If the sparseness falls below a given threshold, we stop the decomposition of that tensor.

The overlap between the non-zero terms in the two tensors is calculated after each decomposition. A high value between the parent tensor and the child tensor shows that the *theme* doesn't change much. A low value indicates when a new *theme* is identified. A selective decomposition can be carried out based on the value of the overlap.

978-1-935068-76-1 © 2011 SciRes.

4 Test-bed Collection

As a test-bed we selected a set of the USA Government Manuals. Created to inform about the functions of the different Departments in the USA Federal Government, these manuals are issued on a yearly basis for each administrative year. While from one year to the next they may contain repetitive information, they do reflect administrative and functional changes that occurred during that period. We chose a subset of 17 departments, one per year for 11 years totaling 187 documents in html format with a total of over 14 million characters and 1,780,000 words. The length of each document ranges from ~2,600 to ~20,000 words.

The manuals are arranged in a directory structure by year (1997-98 to 2007-08) and by the department. To construct the tensor we used year and department information from the directory labels. In this particular case, the name of the department is equivalent to provenance and also corresponds to the document dimension. Each of the yearly term-document matrix (of order 14,990 x 17) was generated based on the TFIDF model after stopword removal [15] and used to construct the term-document-years tensor. This generated a sparse tensor X of size 14,990 x 17 x 11. The 14,990 terms were parsed from the 187 manuals using a master dictionary of 25,978 terms. Since not all of these terms appear in every document, most of the elements in the term-document matrix are zeros. In this case, the number of zero entries in this tensor was 2,556,429 and the number of non-zero elements was 246,701. The tensor decomposition was computed using the MATLAB Tensor Toolbox [16].

5 Results Analysis

Using this method we intend to trace theme changes and overlaps across documents in relation to time and to departments (provenance). The assessment of the quality of clustering is done quantitatively and qualitatively. Quantitatively we compare the sparseness and overlap metrics resultant of the bisecting clustering with those from the original PARAFAC decomposition using the r value as the decomposition parameter. The qualitative evaluation entails manually checking the decomposition results against the document contents.

5.1 Comparison between Bisecting Tensor Decomposition and PARAFAC Decomposition

Table 1. Comparison of tensor decomposition results

method	Number of decomposed tensors	Number of tensors with distinct theme
Bisecting-1	15	3
Bisecting-2	13	4
Standard	14	8

Table 1 shows the results of three different tensor decompositions. “Bisecting-1” is the method where the decomposition was stopped when the difference in sparseness before and after the decomposition was greater than 10%. “Bisecting-2” is the method where the decomposition was stopped when the overlap was 0.1 or less. “Standard” is when we applied PARAFAC with r=14. The number of decomposed tensors resultant from each method is shown in the middle column. The results were manually checked to determine how many of those tensors contained distinct themes.

Using the direct PARAFAC method, the number of distinct themes is bigger than using the bisecting tensor decomposition. A close examination indicates that in the former case, each tensor contains almost all the documents corresponding to a unique department without variation over the time dimension. While this result is accurate, it does not point to the differences or relationships that may exist between departments across time. Differently, the bisecting decomposition produces less number of tensors, showing that there are thematic relationships between the departments. In the following section we present the results and their interpretation.

5.2 Results interpretation

Measuring the relevance of terms across any given dimension in the cluster allows us to identify its salient features. Relevance is the weight of each vector computed during the decomposition. Through graphics generated after the decompositions we visually detect the salient features in the clusters. We explain the process of observation and interpretation with an example.

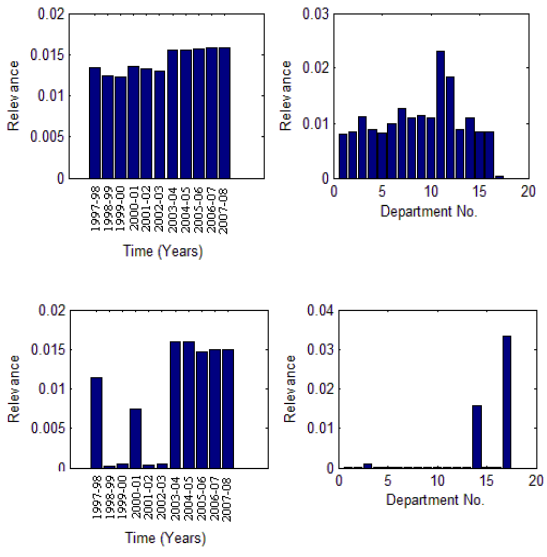


Figure 3. Relevance distribution along provenance (department) and time dimensions for two tensors (clusters) decomposed from the original tensor

Figure 3 shows the results of two tensors named, X1 (top) and X2 (bottom), which resulted after the first decomposition of the original tensor (all the documents). In the left side, the graph shows the relevance of the time dimension for X1 and X2. We can observe that the distribution of relevance across time is even in X1 but not in X2. The graphs to the right show the relevance along the provenance (departments) dimension. In X2, two departments show significant relevance while others hardly show any. A combined view of the graphs leads us to interpret that tensor X2 has the following characteristics:

1. Includes documents largely from departments no. 17 and no. 14.
2. Includes very few documents from years 1998-2000 and from year 2001-2003.

This decomposition indicates a binary split of the original document set into two clusters of documents. One cluster contains documents from the department no.14 (Department of Navy) and the department no. 17 (Department of Veterans Affairs) for the years from 1997-1998, 2000-2001 and 2003-2008. The rest of the documents are in the other cluster. The separation is due to the characteristics embedded in the mathematical model representation.

The next step is to identify what makes these departments salient at this point of the decomposition and what does that mean across time. For each department we generate a list of words that appear with high frequencies for each of the departments in one set but not in the other. We use these words as search terms in the corresponding yearly documents and read the context in which they appear. The narratives reveal administrative changes due to merges and or creation of new units as well as changes in personnel. Within some variations, for each department these changes remain stable during the years indicated for each set. [2] Evaluated in the historical context, the changes roughly coincide with two different presidential periods in the US and their correspondent administrative

After having identified the changes within the departments over time, we wanted to identify the similarities that may exist between them. For this, we generate lists of words shared by both in each of the two sets. Again, we use those as search terms. It has to be noted that in both cases, the lists of words include many terms that are of common use in these types of manuals such as: director, secretary, government, agency, and program. We do not use those words as search terms but instead we select names or nouns that point to functions such as: medical, counseling, cemetery, Pensacola, and dental. Reading the text around those words we found that for the given years, the departments share similar or complementary functions (both have medical services, one offers legal counseling and the other offers psychological counseling), and that they have related facilities in similar sites (an army cemetery within a navy camp in Pensacola). The words act as access points that the archivist can use to evaluate the nature, extent and scope of the changes and relationships.

[2] The variations reflected in the years 1997-1998 and 2000-2001, whose relevance is less than that one shown for the rest of the years in that cluster reflect minor changes in those manuals.

6 Conclusion and Ongoing Work

We investigated the use of tensor analysis with bisecting decomposition in text documents, and developed metrics to evaluate the efficiency of the decomposition and when to stop decomposing. The method can be used to find changes and relations along different information dimensions that were not evident from the normal PARAFAC.

Tensor analysis allows detecting relevant features in document collections automatically instead of manually, as archivists normally do it. Moreover, the flexibility provided by the tensors enhances the possibility of presenting different layers of information to users. Our next project involves developing a user interface to show the tensor analysis results and allow users to interact with collections for interpretation and verification.

References

[1] R.A. Harshman, "Foundations of the PARAFAC procedure: models and conditions," *UCLA Working Papers in Phonetics*, no. 16, pp. 1-84, 1970.

[2] L.R. Tucker, "Some mathematical notes on three-mode factor analysis," *Psychometrika*, no. 31, pp. 279-311, 1966.

[3] T. G. Kolda and B. W. Bader., "Tensor Decompositions and Applications," *SIAM Review*, vol. 51, no. 3, pp. 455-500, September 2009.

[4] M. Mørup, "Applications of tensor (multiway array) factorizations and decompositions in data mining," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, no. 1, pp. 24–40, 2011.

[5] by Evrim Acar, Seyit A. Çamtepe, Mukkai S. Krishnamoorthy, and Bülent Yener, "Modeling and Multiway Analysis of Chatroom," in *IEEE International Conference on Intelligence and Security Informatics (ISI05)*, New York, 2005, pp. 256-268.

[6] M.W. Berry and M. Browne, "Email surveillance using non-negative matrix factorization," *Computational & Mathematical Organization Theory*, no. 11, pp. 249-264, 2005.

[7] Brett W. Bader, Michael W. Berry, and Murray Browne, "Discussion tracking in Enron email using PARAFAC," in *Survey of Text Mining II: Clustering, Classification, and Retrieval*, Michael W. Berry and Malu Castellanos, Eds. London, UK: Springer, 2008, ch. 8, pp. 147-163.

[8] Tamara G. Kolda, Brett W. Bader, and Joseph P. Kenny, "Higher-order web link analysis using multilinear algebra," in *IEEE International Conference on Data Mining*, Houston, TX, 2005, pp. 242-249.

[9] Jian-tao Sun, Hua-Jun Zeng, Huan Liu, and Yuchang Lu, "CubeSVD: A Novel Approach to Personalized Web Search," in *14th International Conference on World Wide Web*, 2005, pp. 382-390.

[10] P. A. Chew, B. W. Bader, T. G. Kolda, and A. Abdelali, "Cross-

978-1-935068-76-1 © 2011 SciRes.

language information retrieval using PARAFAC2," in *13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2007, pp. 143-152.

[11] Deng Cai, Xiaofei He, and Jiawei Han, "Tensor space model for document analysis," in *the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, 2006.

[12] Weijia Xu, Maria Esteva, and Suyog Dott Jain, "Visualizing personal digital collections," in *IEEE/ACM joint conference on Digital Libraries*, Gold Coast, Austrlia, 2010, pp. 169-172.

[13] M. Esteva and W. Xu, "Finding stories in the archive through paragraph alignment," in *Digital Humanities 2010 (DH2010)* , London, UK, July 7 to 10 2010.

[14] J.D. Carrol and J.J. Chang, "Analysis of individual differences in multidimensional scaling via an N-way generalization of 'Eckart-Young' decomposition," *Psychometrika*, no. 35, pp. 283-319, 1970.

[15] Gerard Salton and Michael J. McGill, *Introduction to Modern Information Retrieval*. New York, NY, USA: McGraw-Hill, Inc., 1986.

[16] Brett W. Bader and Tamara G. Kolda. (2010, March) MATLAB Tensor Toolbox Version 2.4. http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox/.

The Methodology of Improving the Efficiency and Quality of Providing Services on Information Science and Technology

Wei SUN

Institute of Scientific & Technical Information of China, Beijing, China

Digital Publishing Center, Science Press Ltd., Beijing, China

Email: sunwei@mail.scoencep.com

Abstract: At the time of internet, massive information is a gift as well as a headache to the educators, researchers and scientists. On one hand, the kinks of scientific document available now on internet are much more than what they can get in a traditional library; on the other hand, it is also a challenge to reduce the noise which is increased along with the quantity of information so as to achieve a balance between efficiency and quality. Library and information institutes in China use the methodology of information sequential organization to manage and organize the massive resources. The method of scientific information analysis is also adopted in accordance with the characteristics of the educators, researchers and scientists to make the most effective integration of information resources in full consideration of the Chinese situation. In addition, technology is used to raise the efficiency and quality of application of the scientific information resources. The combination of all above mentioned methods is the best option for the Chinese educators, researchers and scientists to maximize the utility of scientific document in their work and research.

Keywords: document; knowledge organization; knowledge management; information processing technology

提高科技信息服务效率和质量的方法

孙卫

中国科学技术信息研究所，北京，中国，100038

科学出版社，北京，中国，100717

Email: sunwei@mail.scoencep.com

摘　要：互联网传播时代的海量信息处理，对于教育、研究和科技工作者来说既是福音又是苦恼。一方面，当前科技文献的种类比传统图书馆能够提供的服务更加丰富；而另一方面，如何剔除随信息量增加的噪声信息，以求获取效率和质量上的平衡则是一个需要攻克的课题。图情服务机构利用图情有序组织的方法，针对海量的资源进行序化的组织和管理；结合教育、研究科技工作者的特点，利用科技信息分析的方法，在充分考虑我国国情的前提下最有效率地整合资源，并结合技术手段来提高科技信息资源使用的效率和质量。这对于中国教育、研究、科技工作者最大化利用科技文献来开展研究和工作，是最好的选择之一。

关键词：文献；知识组织；知识管理；信息处理技术

1 问题的提出

上个世纪九十年代中期开始，互联网与移动通信网进入了公众的视野，在不到 20 年的时间里已经改变了人们的生活、工作、学习习惯而成为当今最主要的传媒方式。

尽管 1994 年美国自然科学基金（NSF）开始数字图书馆的研究计划到 1998 年开始的第二期研究计划[1]，都期望利用数字图书馆的研究，解决互联网的信息孤岛问题、解决互联网的信息组织问题、改变互联网信息质量良莠不齐的问题。但是，效果并没有达到：①人们在网络上检索和判断有用的结果时依然花费很多时间；②由于自动摘要技术的限制和资源本身质量良莠不齐，很难保证找到的资源的准确性、有价值；③由于使用爬虫发现与获取技术，很多链接关系随时间的变化或者网站商业价值变化而成为无效链接，所以造成了检索到了但

978-1-935068-76-1

是无法获得，或者点击后才知道是无效链接等。

爱斯维尔科技（Elsevier Science）公司和谷歌公司（Google）的学术搜索宗旨都是为教育和科研提供学术资源信息服务。而爱斯维尔公司是以控制世界上权威的 STM 期刊、书籍等资源作为服务的基础，谷歌公司则主要是利用爬虫技术发现互联网的内容和与各个出版合作获取的期刊和书籍的相关元数据信息作为搜索服务的基础。

Table 1. Comprehensive contrasts between Elesvier Science and Google Scholar
表 1. 爱斯维尔科技和谷歌学术搜索综合对比

	爱斯维尔科技[2]	**谷歌学术搜索**
资源	STM 科技期刊和书籍	各类学术信息元数据及其链接关系
检索	有	有
写作	有	无
评价	强	弱
链接可靠	可靠	不可靠
核心影响力	高	低
使用者	最好的大学和科研机构	自由人群为主
使用效率	高	低
资源质量	高	低
使用成本	高	低

在表 1 中，谷歌公司的学术搜索引擎和爱斯维尔科技的学术研究与服务是不同的，最主要不同是：使用成本，爱斯维尔科技的使用成本高。使用效率和资源质量表明爱斯维尔科技提供的信息服务是是物有所值，这个就是今天出版机构和互联网信息服务公司依然在进行博弈的价值所在。

通过以上对比，看到两个事实：①互联网信息服务中的噪声，无法帮助教育与科研工作者进行有效率的，高质量的科研活动。②简单的书籍、刊物已经无法满足教育与科研工作的需要，STM 出版资源也遇到质高价高而变成不得已时才被使用的资源。

那么对于科技文献共享，是否有另外一个方法，即降低使用者的成本支出，又可以保证资源的质量和利用的效率呢？

2 资源的种类和组织方式

互联网信息服务系统与出版机构以外的，可以利用对文献的组织和管理的方法、经验、平台，向教育和科研机构提供服务的是图书馆和科技信息服务机构。

表 2，表明了图书馆科技情报服务机构既有出版机构产生的资源，也有互联网上对应的资源，还有长期的馆藏加工出来的资源。从资源优势上，可以确定资源比通过互联网检索到的资源要集中、资源的质量优于互联网的发现的大部分资源、资源种类多于出版机构的资源。

Table 2. Types and sources of resources
表 2. 资源的种类与来源

资源类型	来源	利用率
书、刊、报（含电子版）	购买、交换、共享	物理利用率低，网络电子版利用率高
中外专利数据	购买、互联网获取	分析利用高，直接利用低
学位论文	缴存、购买	物理和电子利用率高
OA 资源	互联网捕获	结合主题、收藏利用率不同
科技报告	收藏、缴存	战略、战术型的报告利用率不同
外文连续出版物数据	利用馆藏加工	检索利用率很高，物理出版物利用率低，电子内容链接或者传递
机构数据	利用馆藏加工	
企业数据	利用互联网捕获、行业协会	
主题词表	利用馆藏加工	
叙词表	利用馆藏加工	

根据表 3，在图书馆与情报科学单位的绝大部分文献资源是进行过再组织、再加工的，这样的资源的质量优于互联网各个网站处理过的资源，劣于出版机构权威性编辑处理过的资源。

根据表 2 和表 3 的对比，是否可以通过图情机构为教育、科研、专业工作者提供更有效的科技信息呢？

问世 100 多年的图书馆最主要的目标是为了读者使用方便而进行的文献有序化组织和管理，另一个目标是保存于传承文化。在上个世纪 90 年代初期，美国开始数字图书馆项目研究的目的也是利用图书馆对于文献资源的序化的方法与经验，改进互联网资源的杂乱无序的状态；利用图书馆的互操作的标准和数据交换的方法与经验， 改变互联网的信息孤岛的现象。所以，我们相信利用图书馆与科技信息服务机构对于资源的序化整理和一定的技术服务手段的支持是可以改进科技信息资源发现与使用的效率和提高服务的质量的。

Table 3. Resources Organization and common features
表 3. 资源组织与共性

资源类型	主要组织特点	处理方式	备注
书、刊、报等	书目数据（MARC）	人工	规范知识
中外专利数据	专利规范组织	人工	规范知识
学位论文	书目数据（MARC）	人工	规范知识
OA 资源	简单数据（DC）	半自动	简单/规范知识
科技报告	书目数据	人工	规范知识
外文连续出版物	部分书目数据	人工	规范知识
机构数据	基于文献	人工	非规范
企业数据	基于文献、互联网、工商注册	人工	非规范、非实时
主题词表	基于一类文献统计特征词汇	半自动	非规范、非实时
叙词表	基于文献抽取、购买等	半自动	新词发现不够、规范不够

3 网络传播与资源利用的关系

“清华大学图书馆读者利用图书馆方式的调查”[3]一文中：（1）在利用图书馆网络资源（不同时间段累计），教师占 80.69%，研究生占 98.7%，本科生占 93.96%。（2）在没有在图书馆找到资源时，首选互联网，老师占 58.12%，研究生占 62.73，说明互联网资源成为图书馆资源的补充。（3）获得所需文献的相关信息途径，79.06%的老师和 72.45%的研究生利用文后参考文献，然后才利用文摘索引、搜索引擎。这组调查数据清晰地说明了对于教育和科研工作，图情机构组织和提供的服务是主要的，互联网与搜索引擎是辅助的。而在资源获取上，图书馆科技服务机构的优势更强，因为这些机构是提供科技信息为教育和科研服务的公益单位之一。在学术资源这个前提下，内容丰富互联网和图书馆处于相当的水平，而在质量可靠上，首选利用上，图书馆组织和管理的资源明显占有优势，在查找方便上，图情的服务是优先的，互联网的服务是互补的。

“高校研究人员学术信息资源利用及信息查寻行为调查与分析”[5]一文中：（1）在使用过的电子资源类型中，学位资源占 37%，电子期刊资源占 91%，数据库资源占 61%，电子图书资源占 31%。由此可见，这四种类型的学术资源很难在互联网上免费获得的。（2）对于专业学术资源的熟悉程度中，很熟悉的占 28%，比较熟悉的占 56%，一般占 16%。由此可以看出，在利用专业资源上，针对性强是提高科技信息利用效率和质量的一个因素。（3）获取资源的途径上看，图书馆网站、专业网站排在 2 和 3 位，而图书馆培训排在最后。说明图情人员想按照图书情报专业的思路提高使用者的信息素养是一件很困难事情。简单对资源序化还不能彻底的提高使用的效率，需要符合自然人的行为习惯与文献组织序化原理之间的收敛，才可能找到提高效率的方法。

提高效率和改进质量的关键是如何把自然人的行为习惯与图书情报专业对于资源的组织和管理的方式找到收敛和匹配的方法。

4 收敛与行为匹配的方法

4.1 用户检索词与资源关键词分析计算收敛方法

在谷歌上是利用词表对于资源进行切分处理，对检索词汇与这个词表匹配进行检索和定位的。图书情报的资源组织是按照主题词表的方法进行分析处理的，而主题词表[5]的用、代、属、分、参是可以把词汇之间进行关联。但是，传统的这个主题词表收集与整理的时效性是比较弱的，主题词表现的是一类文献的共性词汇。所以，可以利用主题词原理进行资源的自动分类处理，基于国家图书馆的分类主题词表已经实现的资源组织管理过程的自动分类，这个极大的提高了文献资源组织中知识处理的效率。针对每篇文献给出的关键词，一般这个关键词代表了本篇文章的重点的概念。

基本原理如下（见图 1）：

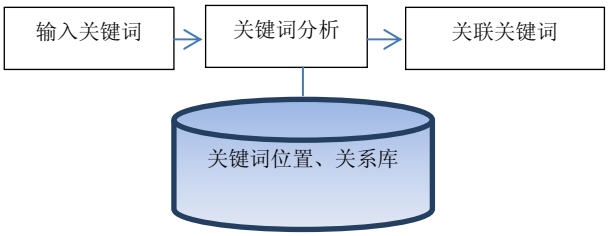


Figure 1. Get a set of keywords through analyzing links among keywords of literature resources
图 1. 建立文献资源关键词关联分析取得关键词团

在新的方法里，建立关键词位置、关系库。利用输入的检索词，根据关键词库，计算出这个检索词与其他关键词同时出现在资源中频度，同时出现频度最高的关联关系最强，依次类推，从而根据设定的关键词群团的关键词的数量，构成关键词团。再把这组关键词团发到检索系统中去进行检索，就可以把单个关键词与多个资源关联的关键词上进行资源组织关键词的收敛，然后再排序上，先处理与的关系，在处理或的关系。达到组合检索的目的，又不需要使用者一定

978-1-935068-76-1 © 2011 SciRes.

要学检索理论。这样检索命中效果、按照先“与”后“或”的方式排序，可以提高符合使用者愿望的结果按照资源组织的特点排在前面的位置上。在中国科学技术信息研究所 2010 年完成的国家科技基础支撑项目的科技评价分析系统，已经实现了关键词分析，下一步解决与检索系统有机结合和自动处理。这个收敛的方法，明显不同于完全个性化的，基于不同角度的相同的词汇利用的点击统计，而是利用资源组织的关键词，利用关键词的关联分析，把绝对个性化的输入检索词与资源关键词进行关联，加速针对资源组织的收敛关系，使得检索的命中既符合读者的个性又满足资源组织的共性。

4.2 针对学术资源特点把排序选择

在不同的检索结果排序中，有很多算法，典型的是按照字母、偏旁部首笔画、Page Rank 等。谷歌的搜索引擎采用对于组合检索中，“与”在前，“或”在后，然后是 Page Rank 的组合排序法。而与、或是检索方法，Page Rank 是文献链接方法，没有办法证明这两个方法和学术专业的相关度更强，特别是对于研究者的兴趣的关联关系。而对于研究、科技、学习者，可能关心某个领域资源的权威性，权威最主要的特点是经典性为主，而经典性主要在时间轴和被引次数累计两个维度上表现出来的（传统的引文分析，只有引用率）。对于研究者和一定专业的经验者，可能关注是不是核心期刊，或者关心是不是新的为主，那么表现的特点是重要性和实时性。那么，基于研究、科技、学习者在利用资源上，在检索结果排序原则上出现了时间、是否是核心、被引三个维度的组合。在北京万方数据的检索体系[6]里，可以让使用者选择按照时间与引用、按照引用、按照核心等条件进行排序选择。如果使用者不选择，就给一个三维加权不同的综合排序方法。目的就是使得命中的检索结果尽可能的出现在现实的结果集的第一页上。这个算法有效的提高了用户期望的检索结果出现在第一页面的概率。

4.3 提供服务结果的服务方式

在传统的信息服务商，一般是检索，原文链接，点击阅读的模式。在面对海量信息检索服务中，对于命中的结果集的分析判断是专业用户面最为头疼的事情。把科技情报查新查证、分析、汇总、报告这些流程组合成资源元数据、检索工具、分类汇总、生成报告过程的工具，最后提供报告，是一种面向开题、查新查证等的科技创新辅助决策系统尝试的新的服务模式。对于提高服务的效率和质量是一次有益的探索。

4.3.1 海量的资源

在中国科学技术信息研究所的“科技创新辅助决策系统[7]”中已经有 2 亿条以上的文摘和元数据，大部分可以连接到原始资源上。

中文期刊元数据、西文期刊元数据、中文会议元数据、外文会议元数据、中文学位元数据、中国专利元数据、国外专利元数据、科技成果元数据、企业信息、词知识库、作者知识库等构成这个服务系统海量的科技资源相关信息。

4.3.2 五要素关联分析

基于人物、机构、项目、主题、分类五大要素对于检索结果集进行汇聚、统计分析、引文率加权的处理，实现了多种异构资源信息的统一处理的方法。根据表 3 可以得知，各种资源的组织模式是不同的，那么如果检索平铺，就很难聚合、统计、分析。中国科学技术信息研究所的五要素法是在原有资源组织基础之上先进行汇聚处理。科技活动中最主要的核心都是围绕这五个要素表现的，再利用情报计量学，引文分析等理论对汇聚结果进行统计，达到了发现、汇聚、统计分析、排序的目的。

4.3.3 按照报告框架构成报告内容的主题

传统的科研人员查新查证后的结果是分散的，要分别写对于各个数据库查询的结构然后汇总形成报告。而该系统在特定框架下，把汇聚分析的内容直接填入对应的部分，并给出可视化的分析图，形成主的报告结构和内容，极大地方便研究者书写报告和对于主报告进行润色的工作。

该平台建立了利用各种资源（只是利用资源的摘要、元数据）、利用各种工具（检索、五要素聚合、引文分析、报告、可视化），提供信息服务的成果（报告）的崭新信息服务模式，帮助研究、科技、学习者可以高效率和高质量的进行海量科技信息的查新查证、主题分析、人物分析、项目分析、论文写作分析等。如果使用者需要，资源可以链接到原始资源供应商、图书馆、科技情报服务机构等提供原始资源的服务。

4.4 利用工具判断论文质量

研究成果的重要表现形式之一是科技论文。传统的论文主要靠编辑、专家、导师对于作者论文的审读

来保证其学术质量。在没有互联网的时代，由于纸媒传播的周期长，数量有限，很少有作者能阅读到非常全面的各种科技文献。但是，在互联网时代通过搜索引擎作者可以很方便发现海量的科技文献，由于计算机大量使用，方便写作的工具大量涌现，可以很方便的摘抄、转译、修改并利用。同时，不少作者利用名人进行第二、第三署名，来提高论文被核心期刊录用的概率。所以，把握研究成果之一的科技论文的质量，提高科研单位的声誉，保障名人的荣誉成为打击学术不端行为的任务之一。目前中国有三种类型的相似性检测系统，清华同方知网公司和北京万方数据公司为编辑部、教育与研究机构的研究生办公室都开始提供检测服务，这两个相似性检测系统主要利用中文科技文献对于学位论文和期刊论文进行相似性检测，由于这两个系统的推广使用，已经减少了相当一批简单抄袭的作品发表在核心期刊和研究生毕业论文。

另一个系统是武汉大学基于数十万互联网资源的相似性检测，对于利用互联网写作的相似性检出率很高，对于基于海量的科技文献很难检出，加上互联网测试源的快速增长，获取数百亿资源形成困难，这个系统基本无法发挥作用。

但是，利用工具来辅助提高研究成果的质量是得到了编辑部和研究生办公室认可。

中国科学技术信息研究所正在研究检测相似性论文的指标体系，期望可以更科学对不同类型的论文、论文的不同结构、标点符号、参考文献、引文进行更规范的检测指南。同时，转译、修改等类型的相似性的算法研究也得了社科基金的支持。

5 结束

通过对问题的引出、分析了资源的类型、组织方式，分析了互联网传播与资源利用的关系，对于科技信息服务机构利用方法，技术提高科技信息资源的使用效率和保证利用的质量，通过中国科学技术信息研究所和北京万方数据、清华同方知网公司的研究、实验、提供服务的效果，已经证明，在中国这种体制下，利用图情服务机构可以为中国的教育、科学技术研究、创新等提供更有效率，质量更高的科技信息服务。

未来的中国科技信息服务需要在数据挖掘、文本挖掘分析上，提供更多的应用平台；在在线翻译、半自动翻译上为科技人员利用俄文、日文、英文科技资源提供帮助；在多角度的评价指标上帮助科技人员更好的对待科技文献资源的取舍。

References (参考文献)

[1] NSF. Digital Libraries Inittiative –Phase 2. http://www.nsf.gov/pubs/1998/nsf9863/nsf9863.htm.

[2] Elesvier Science. http://www.hub.sciverse.com/action/home/proceed.

[3] Yi YANG. Investigation of Behavior Mode of Information Demand of University Library Readers [J]. *Digital Library Forum*, 2009, 1: 1-24(Ch).
杨毅. 清华大学图书馆读者利用图书馆行为方式的调查[J]. 数字图书馆论坛，2009, 1: 1-24.

[4] Xiao Dong LI. Survey and Analysis on Library Use and Information-seeking Behaviors of Researchers in Academic Institutions [J]. *Digital Library Forum*, 2009, 1: 25-42(Ch).
李晓东. 高校研究人员学术信息资源利用及信息查寻行为的调查与分析——经北京大学图书馆用户调查为例[J]. 数字图书馆论坛，2009, 1: 25-42.

[5] Chinese thesaurus. ISTIC, NLC. Scientific and Technical Documents Publishing House, 1980.
汉语主题词表. 中国科学技术信息研究所,国家图书馆主编. 中国科学技术文献出版社，1980.

[6] WANFANGDATA CO., Ltd. http://www.wanfangdata.com.cn/Help/index7.html.
北京万方数据股份有限公司. http://www.wanfangdata.com.cn/Help/index7.html.

[7] WANFANGDATA CO., Ltd. http://stads.wanfangdata.com.cn.
北京万方数据股份有限公司. http://stads.wanfangdata.com.cn.

978-1-935068-76-1 © 2011 SciRes.

Ontology Mapping Model of Mining Literature Knowledge Element Object

Youkui WEN[1], Hao WEN[2], Bing SHEN[1]
[1]*School of Economy and Management, Xidian University, Xi'an, China*
[2]*School of Information and Control Engineering, Xi'an University of Architecture Technology , Xi'an, China*
Email: wykui123@126.com, smczg@126.com, shenbing9599@126.com

Abstract: The model to express knowledge element object on the semantics relationship unit existing within knowledge fragments of the literature is proposed, and ontology mapping models between the ontology of literature topic factors and the ontology of knowledge elements is explored in this paper. To solve the current literature-based knowledge discovery methods difficult to overcome two key problems: 1) Reasoning object is not reflected in the semantic relations unit of knowledge fragments, loosing the true semantic relationships. 2) Reasoning object selection limited only the title or abstracts in the literature which result in the loss of a large number of objects of reasoning and knowledge discovery opportunities. The illustration conducted indicates that ontology mapping model of mining literature knowledge element object is an effective method.

Keywords: factors of literature topics; knowledge element; ontology mapping method

1 Introduction

Facing the most serious challenging problem that finding the ways knowledge associated with the interdisciplinary cooperation during the study [1], information scientists proposed a new theory to deepen the knowledge service units from literature to the "knowledge element" [2]. This theory is considered the key breakthrough to new information environment in Information Science [3]. Theory of knowledge element was successfully validated and applied by Swanson in the type of knowledge fragments [4]. Swanson was also developed a software supporting query Arrowsmith with the help of database. Knowledge discovery method based on the literature has attracted the attention of many researchers including medicine, biology, information science and computer science researchers. However, after Swanson, the improvement of the method can not be widely promoted because of inefficient and heavy workload for experts to identify the correlation between literatures. Intelligence experts proposed that how a particular way of organizing information or knowledge with visually can be presented to the user has been a problem to be solved for knowledge management system. This article explores a knowledge element structure of breaking knowledge element fragments in literature down into semantic triples, researches into a new method of literature knowledge discovery based on the knowledge element for the object, solves the problems of inefficient in the current methods of non-relevant knowledge discovery , and will realize the goal of literature knowledge discovery ,which is based on element knowledge for the object, direct application of knowledge and dynamic semantic inference, easy-to-wide promotion, and high efficiency. Literature-based Knowledge Discovery will be shown in Section II, Subject Elements and Meta-object triples semantic ontology mapping model of knowledge in Section III, and conclusion and outlook in Section IV.

This work is partially supported by the Ministry of Education, Humanities and Social Sciences Planning Fund No.09YJA630127.

2 Knowledge Discovery Based on the Literature

2.1 Swanson' Knowledge Discovery and Improvement

(1) Two modes of knowledge discovery

In 1986, the University of Chicago Professor Don R. Swanson research into division phenomenon in knowledge, and created two modes of "literature-based knowledge discovery": open and closed.

Open knowledge discovery process is the one that by the literature A searching for the middle of words (or literature) B, then by the middle of the literature B searching literature C. Open path of knowledge discovery can be expressed as: A → B → C. Closed knowledge discovery process is taken A and C as the starting point to find common intermediate term B. Closed path of knowledge discovery can be expressed as: A → B ← C.

Whether open or closed knowledge discovery based on non-relevant documents, the key is to determine the middle literature set B.

(2) Improvement of middle literature set B

Since Swanson, most of the researchers mainly made a study of the efficiency of knowledge discovery in non-relevant literature from the view of the determination of associated word B and sorting techniques and methods, which is shown from aspects of the object in the literature mining, text analysis unit, the results of evaluation of mining methods and degree of automation. 1993, Z. Chen [5] proposed the knowledge integration method which is based on logical interconnection of knowledge fragments

dispersed within the literature. In 1999, Gordon and others[6], extended text mining from the title object to abstract, from term to phrases. 2001, Hristovski introduced association rule mining to literature-based knowledge discovery. Weeber and others [7] thought that Swanson and other researches studied and tested based on unit words, binary and ternary words, which is not suitable for the current increasingly common semantic analysis environment, because a number of semantic can not be generally and accurately expressed in just units word, binary word or three words. 2004, 2006, Illhoi Yoo's doctoral thesis [8] proposed a distinction between text data mining and text knowledge mining, introduced a concept of semantic text mining and semantic reasoning hypothesis, and tested ontology mapping of the free text and MeSH terms biological concepts, but it can not solve problems of the association of independent semantic.

2.2 Research on Knowledge Discovery in China

Since 2002, Chinese Information Science workers have done a large number of introduction and evaluation to Swanson methods[9], such as the development of the Arrowsmith web Swanson open software testing in Medline database, as well as Chinese literature experiment. Knowledge Discovery of non-relation literature is also suitable for Chinese literature. At the same time, data processing in Arrowsmith will produce a lot of middle words caused great difficulties for the user identification, which is the one of the limited factors promoting the use of Arrowsmith.

Chinese researchers generally improve the experimental with words cutting, word frequency, weight calculations, threshold value determination, the appropriate intermediate literature B selection, Document version and abstracts for the objects. They consider that the shortcoming of knowledge discovery of non-relation literature is that it can not objectively reflect the degree of subject association of B, propose that cohesion literature weighted ranking method B-word can improve the higher detection rate than cohesion literature weighted ranking method, and figure out the biggest shortcoming of using simple keywords previously is resulting in a large number of redundant pairing.

2.3 Two key Issues of Literature Knowledge Discovery

We see that Swanson knowledge discovery of non-relevant documents consists of two key components: First, a "semantic triple knowledge element" object based on a causal relationship; the second one is through semantic relation link with action and the reaction between the two objects finding a reasoning process of new knowledge element, we will use Figure 1 showing the principle. For example, two knowledge elements: knowledge element 1), A (migraine) and B (the proliferation of cortical depression) → (relevant); knowledge element 2), C (Mg) → (reduce) B (the proliferation of cortical depression). It can deduce that there may be a new knowledge element, C (Mg) → (helps reduce) A (migraine).

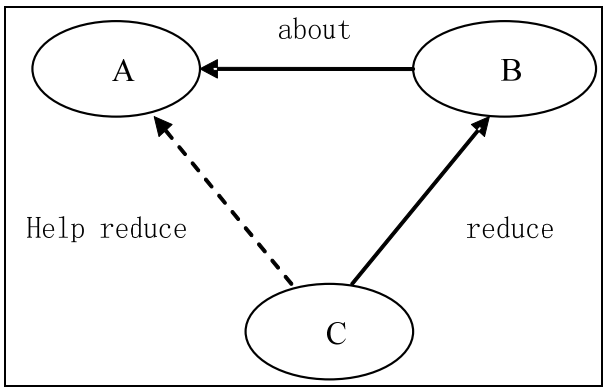


Figure 1 the structure of semantics triples knowledge element of containing within the knowledge fragment of literature

The literature used in knowledge discovery approach has two shortcomings difficulty to overcome[10,11]: (1) Reasoning object is not a semantic relations unit reflecting knowledge fragments, loosing the true semantic relationships. (2) The choosing range of reasoning object limited only in the title or abstract of the literature, which result in the loss of a large number of reasoning objects and knowledge discovery opportunities. These are the key issues leading to the current low efficiency. So this paper pays attention on knowledge per unit of semantic object literature inherent fragments, new methods based on the knowledge element of the dynamic semantics of reasoning between objects of knowledge, with theoretical innovation and broad application prospects.

3 Ontology Mapping Model between Subject Factors and Knowledge Element

3.1 The Model of Semantic Triple Knowledge Element Object

We classified and analyzed 11 semantic relations found by Swanson "Mg affects migraine", then we can draw three conclusions: (1) there is the structure of semantic triples knowledge element < the main, semantic relation, the object >,which means there is a causal semantic relationship between(A) and (B), a reaction semantic relation between (C) and (B); (2) semantic transfer relation,)there is a semantic transfer relation between (C) and (A); (3) essence of literature knowledge discovery is a process finding a new knowledge element based on semantic links of action and reaction between a pair of "semantic triple knowledge element".

Semantic triples knowledge element structure can be defined as BNF:

978-1-935068-76-1 © 2011 SciRes.

Semantic triple knowledge element (Semantic Triple Knowledge Element, STKE):={(subject, object), semantic relations}, subject: = event sponsors; object: = bearer of event; semantic relations: = the intermediary between subject and object, which can be action, characteristic, relation etc..

3.2 Subject Elements and Knowledge Element Ontology Mapping Conceptual Model

Extraction method of semantic triple knowledge element is a process mapping literature knowledge fragment into semantic triples according to rules of ontology database of documentary subject elements

(1) Ontology mapping conceptual model

Ontology model warehouse is built using ontology technology to guide and explain text mining process of semantics of text units, to achieve ontology mapping from the text to the knowledge element. The subject elements and knowledge element ontology mapping model is shown in Figure 2.

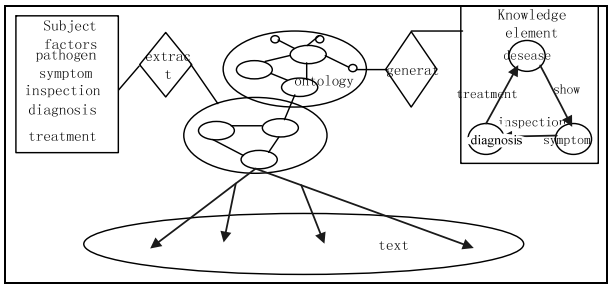


Figure 2. Conceptual model of ontology mapping of knowledge element

(2) Subject factors ontology model

The establishment of thematic elements in the six-level ontology model: we will further vary and reduce factor in the literature the theme of the ontological structure to six levels.

Thing A1 = {(Type B), (Part C), (material D)},

The thing itself A2 = {(process E), (phenomenon F), (characteristic G), (rule H)},

The relationship between things A3 = {(effect I), (application J), (reaction K), (relation L),

The operation to A4 = {(measures M), (process N), (equipment, O), (institution P), (purpose Q), (evaluation R)},

The objective environment A5 = {(site S), (time T), (condition U)},

The exterior features A6 = {(exterior features of literature Z)}

(3) Subject factors semantic relations network model

Establishing semantic network in the literature subject factors ontology model: we take the first layer A1 to indicate the vocabulary set of things' names, and the second layer A2 to indicate the semantics set within things. After further semantic analysis of the local ontology view, different ontology view elements can logically form a "semantic network of associations". The "node" of the semantic association network represents ontology views, and "edge" represents semantic relationships between elements connecting ontology view.

For example, we would like to express semantic relation networking between the two things A11 and A12, and it can be expressed as: {A11, I, J, K, L, A12}, which structure is shown in Figure 3.

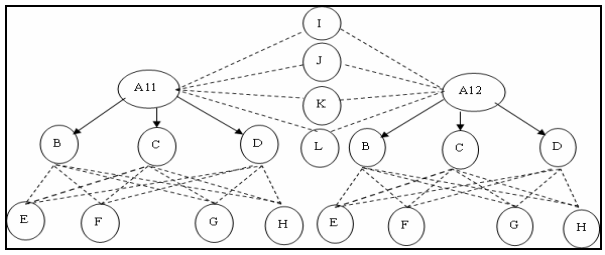


Figure 3. The triple semantic network between A11 and A12

4 Ontology Mapping Steps from Documents to the Knowledge Element

4.1 Semantic Mapping of Knowledge Element

Definition 1: Document knowledge element is a fragment composed of a set of lexical semantic knowledge. It constitutes a specific semantic scene, and completes a particular pragmatic knowledge.

Definition 2: Document knowledge element ontology structure can be regarded as sub-ontology of literature theme ontology. Knowledge element body structure literature can consist of the name of things, a collection vocabulary, an agreed set of special semantic relation pragmatic components.

We propose that the semantic mapping of knowledge element of literature theme is the mapping one ontology to another, which can divided into a direct ontology, mapping, an including mapping, a portfolio mapping and a decomposition mapping, etc. Then we put the semantic mapping of literature knowledge element expressed as three levels of mapping:

The first layer: structure mapping. The ontology mapping document structure to knowledge element structure;

The second layer: A collection vocabulary mapping. It is simplified as "literature structure", words set W to knowledge element structure W1;

The third layer: the theme mapping. Further simplify to the mapping of the subject literature.

We can see document ontology as a global view, knowledge element as partial view. After semantic analysis of global view elements, different ontology views elements form a "semantic network of associations". in the logic. Its "node" represents the view of ontology knowledge element, and "edge" represents semantic relationships between elements connecting ontology view.

On this basis, we can calculate degrees of semantic association between concepts, according to which we can cluster the local ontology concepts, and then generate conceptual clustering as the basis of building knowledge element ontology view.

4.2 Ontology Mapping Implementation

According to the definition we give of literature and knowledge element mapping, we take this paper "Automatic indexing and automatic segmentation" as an example to verify the method literature to the knowledge element of ontology mapping proposed in this paper. At the same time the mapping steps have been given that establishing a "document structure" in the literature, which are shown as follows:

(1) Ontology mapping

Step one: extract the sentence with the characteristic "structure" words from chapter titles of the literature. Software is the knowledge element extraction software text (TKEE) developed by ourselves.

Step two: extract a collection vocabulary about "document structure" W {A1, A2}.

A1 = {basic structure, subsidiary structure}

A2 = {section title, introduction, title, classification, keywords, summary, conclusions, the core word of the literature theme, reference lists, drawings, schedule}

Step three: extract semantic relations A3 among "literature structure" words.

A3 = {divided into, including, present in}

Step four: Construct subject factor ontology (SCO) structure of "literature structure".

$$SCO = \{A1,\ \ A2,\ \ A3\}$$

(2) Information unit structure of ontology knowledge element

Step Five: Build "literature structure" information unit of knowledge element ontology framework (KEO)

"Document structure" knowledge element: O, R

O: Object

R: Relations in objects

(3) Ontology mapping

Step Six: mapping from SCO to KEO is shown as follows:

$$SCO(\mathrm{A1},\mathrm{A2},\mathrm{A3}) \xrightarrow{\mathrm{Map}} KEO\{\mathrm{O},\ R\} \qquad (4\text{-}1)$$

Where, A1, A2 map O, A3 maps R.

Through the above mapping of documentary subject elements and knowledge element ontology, we map the subject element "document structure" to the corresponding semantic knowledge element. The information unit structure of semantic knowledge element model of "document structure" is shown in Figure 4.

5 Conclusion

Contributions of this paper are breaking internal document knowledge fragments down into the knowledge element structure of the semantic triples element relationship, making a study on a new method based on the literature of the new knowledge element as object, and solving current inefficient because of the semantic loss in knowledge discovery. Using documentary subject elements and knowledge element ontology mapping, we can realize extracting semantic triples knowledge object model from literature knowledge fragments, which can build the theory and methods for direct application of knowledge and dynamic semantic inference.

References

[1] Lan Foster, Carl Kesselman. The Grid Blueprint for a New Computing Infrastructure [M]. Amsterdam, Boston Elsevier Morgan Kaufmanns, 2004.

[2] Xu rujing. Development of knowledge resources, development of knowledge industries, services, knowledge-based economy [J]. Modern Library and Information Technology, 2002 (s1).

[3] Ma Feicheng. Information Science Progress and Deepening [J]. Information Technology, 1996,15 (5) :337-343.

[4] Swanson Don R. Undiscovered Public Knowledge[J].The Library Quarterly, 1986,56(2):116.

[5] Lindsay R K, Gordon M D. Literature-based discovery by lexical statistics [J].Journal of American Society for information Science,1999,50(7):574-587.

[6] Lindsay R K，Gordon M D. Literature-based discovery by lexical statistics [J].Journal of American Society for information Science,1999,50(7):574-587.

[7] Weeber M, Klein H,Lolkje T W, et al. Using Concepts in Literature-Based Discovery: Simulating Swanson's Raynaud-Fish Oil and Migraine-Magnesium Discoveries [J]. Journal of the American Society for Information Science and Technology, 2001, 52(7):548~557.

[8] Illlhia Yoo. Semantic Text Mining and its Application in Biomedical Domain [D].Drexel University, 2006.

[9] Rong Yihong, Liang zhanping. Findings based on the literature [J]. Intelligence, 2002,21 (4) :386-390.

[10] Wen Youkui, Cheng Peng. Implicit association between elements of knowledge-based knowledge discovery [J]. Intelligence, 2007,26 (5) :653-658.

[11] Wen Youkui, Wen Hao. Semantic Text Deep Mining Based on knowledge element[C]. 2010 The 3rd International Conference on Computational Intelligence and Industry Application (2010 PAIICA), December 4-5, 2010, 426-429.

978-1-935068-76-1

Research Profiling Based on Semantic Web Mining

Zhixiong ZHANG[1], Na HONG[1], Ying DING[2], Jianhua LIU[1], Jian XU[1], Daiqing YANG[1]

[1]*National Science Library, Chinese Academy of Science, Beijing, China*

[2]*School of Library and Information Science, Indiana University, Bloomington, US*

Email: {zhangzhx, hongn, liujh, xujian, yangdaiqing}@mail.las.ac.cn, {dingying}@indiana.edu

Abstract: Research profiling is an analysis method could be used to broadly scan the contextual literature information and depict the research context and research efforts wisely. The authors of this article now carry out a project which needs to extract intelligence from web resources especially for scientific research analysis. Research profiling based on semantic web mining is the heart of the project. Semantic web mining of the web resource include the two processes which are semantic knowledge extraction and semantic mining. The former focuses on how to automatically extract research terms, research objects and relationship between those objects from the web resources. The latter focuses on how to mine semantic knowledge from a large volume of the structured semantic data. Detailed solutions to semantic mining of web resources are discussed in this article. And the authors also implement a research profiling experiment in Artificial Intelligence area to show the effectiveness of the research profiling based on semantic web mining.

Keywords: research profiling; semantic web mining; knowledge object extraction; visualization

1 Introduction

Research profiling is an analysis method based on bibliometrics and text extraction tools, which could be used to broadly scan the contextual literature information and depict the research context and research efforts wisely [1] [2]. With the help of Research Profiling tools, one can easily observe related topics within the research domain, gain a "big picture" perspective on the research activity, understand the research community, gain insight into how innovation is progressing, and map (graphically represent) topical interrelationships for a whole research area.

Much work about Research Profiling has been carried out, such as [1] [3] [4]. The data they used in analysis are generally structured data such as projects lists (e.g., NSF projects), abstract databases (e.g., Chemical Abstracts), citations indexes, patents abstracts, and formal scientific publications such as periodical papers, conference proceedings. But in real world, we get lots of information firstly from informal web resources. Compared with the high-quality literatures, those informal Web resources are open to access, rich in content and large in volume, could be distributed immediately, and dynamically reflect the changes and the progresses of scientific researches. Those Web resources are good resources for monitor and depict the development of scientific researches. But, on the other hand, those informal web resources are unstructured, poor in semantic meaning, and sometimes unreliable and unstable. So it is a great challenge to do Research Profiling based on those Web resources.

Some attempts have been made in Research Profiling based on those Web resources. For example, Alan Porter and his team have tried to mine the internet for competitive technical intelligence (CTI). They tried to bring Research Profiling into Web resources mining to discover competitive intelligence in commercial area through Google Soap Search API [6]. But it focuses on statistics of search results rather than deep analysis of text contents.

Nowadays, the authors of this paper are undertaking a project named Science Monitoring and Evaluation based on Scientific Web Resources, which is funded by National Key Technology R&D Program in the 11th Five Year Plan of China. The main goal of this project is to form a comprehensive methodology on automatically (or semi-automatically) extracting intelligence from web resources especially for scientific research analysis. Developing technologies to detect the scientific research activities, to monitor the progress of one research area, to track the evolution of one research topic or a research community, and to profile the key research unit is the heart of the project. Compared with Alan porter's research, we tries to do research profiling based on semantic web mining, which means we need to use various integrated knowledge technologies such as information crawl, information extraction, semantic annotation, clustering and visualization etc., to achieve the deep content analysis.

This paper is organized as follows. Section 1 introduces some background of this paper. Section 2 overviews the whole idea and the proposed framework of our research. Section 3 presents the solutions of semantic mining of web resources which is the key problems we need to overcome. Section 4 gives an implemented use case to evaluate the effective of our semantic web mining methods in research profiling and section 5 concludes the paper.

2 Overall Ideas and Framework

In order to do Research profiling based on semantic web mining, we need to: (1) collect and clean different

kinds of scientific web resources, so that we can get the data used for analysis; (2) extract semantic research objects (such as the researcher of a research project and terms of special subject) from those web resources; (3) construct knowledge repository to store the extracted knowledge objects for the future mining and analysis; (4) perform semantic mining based on the semantic data stored in knowledge repository, try to discovery such as hot topic, novelty things that hidden in the resources; (5) do visualization analysis to profile the targeted object. Fig. 1 shows the overall idea for Research Profiling based on semantic web mining.

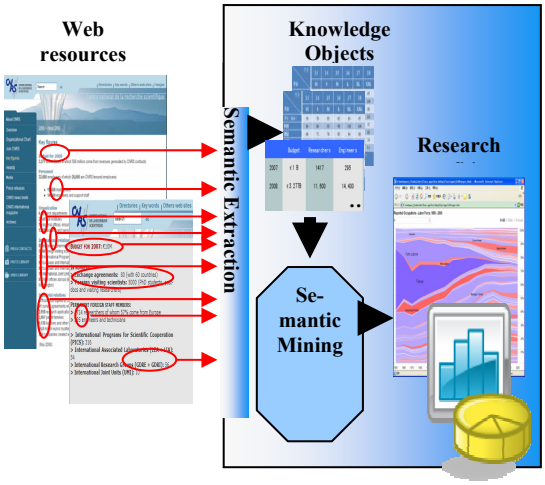


Figure 1. Overall idea for Research Profiling based on semantic web mining

Base on the upper ideas, we bring forth the whole framework of research profiling based on semantic web mining, as depicted in Fig. 2.

Five main functions have been included in this framework, which will be detailed below.

1) Web resource collecting and cleaning: We collect important institution websites, news websites and newsgroups of related research area, RSS from related website, OAI repository of one institution, personal homepages and blogs of one researcher from Internet

2) Semantic knowledge extraction: We named these extracted research terms and research objects as knowledge objects. Based on a Research Ontology designed for a specific research area, we combined existing methods; various integrated approaches and improved approaches to achieve more refined extraction results. The knowledge objects extracted here are all important terms and entities related to a specific research area. Besides, we extract relations between two research objects.

3) Knowledge repository construction: Knowledge repository is composed of a series of extracted knowledge objects with timestamps, for example, an instance of structure "class, research object, harvest time" is "research project, Science Monitoring and Evaluation based on scientific web resources, 2009-01-01"; Based on this computable data structure, knowledge repository is more clear and effective for future analysis.

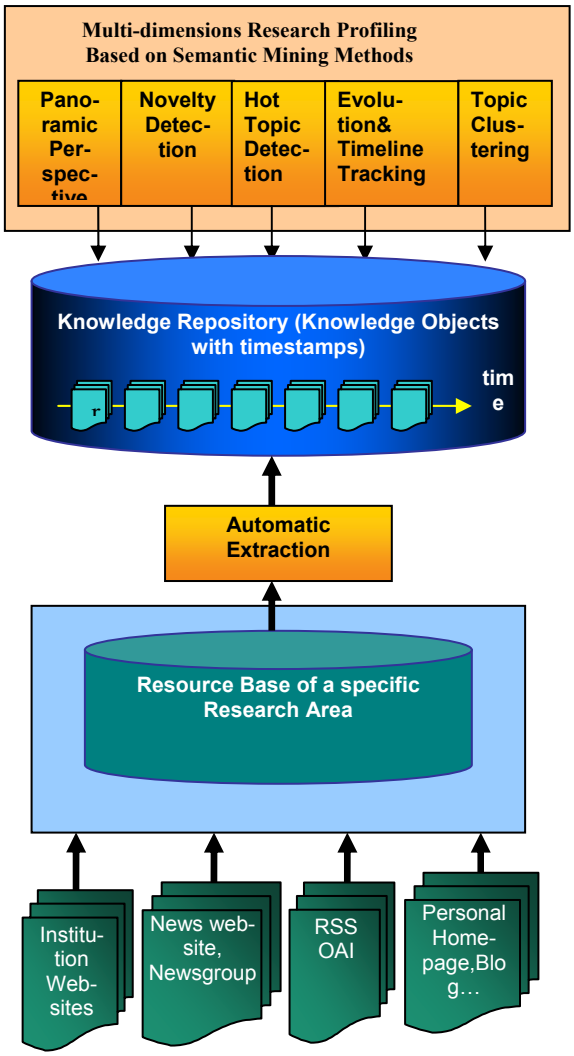


Figure 2. Framework of Research Profiling based semantic web mining

4) Semantic mining: By using a set of bibliometrics and semantic mining methods, we perform semantic mining based on the semantic data stored in knowledge repository, try to form panoramic perspective of a specific research area, perform novelty detection, hot topic detection, timeline tracking and topic clustering to discovery knowledge hidden behind the web resources.

5) Research profiling: In this process, we perform visualization analysis to profile the targeted research area. We detect the scientific research activities in one research area,, figure out the key components in the area, depict the relation between those components, monitor the progress of one research area, track the evolution of one research topic or a research community.

978-1-935068-76-1 © 2011 SciRes.

3 Semantic Web Mining of the Web Resource

As mentioned earlier, semantic web mining of the web resource is the heart of the project. From the framework we brought forth in last section. Semantic web mining of the web resource include the two processes which are semantic knowledge extraction and semantic mining. The former focuses on how to automatically extract research terms, research objects and relationship between those objects from the web resources. The latter focuses on how to mine semantic knowledge from a large volume of the structured semantic data.

In this section we present the solutions we taken for semantic mining of web resources in our project.

3.1 Semantic Knowledge Extraction from Web Content

When we implement a research profiling for a research area, one thing we need to depict is the key research institution, researcher, research activity, conference, research outcome etc in this research area and the relationship between them in the research area, the other thing we need to depict is the structure and evolution history of topics in this research area. This gives us the hint that the knowledge objects we need to extract from the web content include two kinds. One kinds of the knowledge is research object such as researcher, research institution, research project, and research foundation in a research area which tell about the research activities that take place inside the research area. The other kind of the knowledge is researches terms that tell about the topics and subjects of the research area. In order to guide the extraction action and organize extracted knowledge objects, the research ontology is proposed.

1）Research Ontology

In the Research Ontology we proposed, the top classes include Research Activity, Research Outcome, Research Organization and Person, Research Facilities and Basic Concepts. These classes can be classified further into more specific subclasses. For example, Research Activity includes Project, Conference, Lecture, Research Award, Experiment, Investigate and Train. In addition to hierarchical structure of the concepts, the Research Ontology also describes relationships between the research objects. For example, we could define has_attendees relation between Research Activity and Person, supports relation between Foundation and Project, etc.

Part of the Research Ontology is presented in Fig. 1. In this figure, circles denote top classes while ellipses with grey shadow denote classes belonging to top one. The other ellipses with broken line denote classes belonging to grey ones. Besides, thin arrows denote relations between classes and thick ones denote class hierarchy relations.

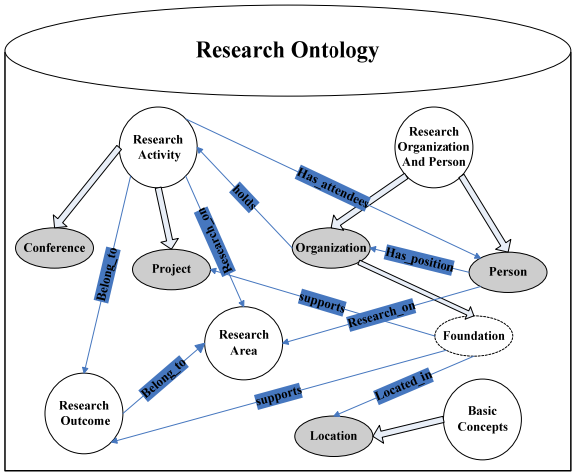


Figure 3. Structure of the Research Ontology

All instances of the classes defined in the Research Ontology need to be extracted from the collected web resources. Next we will introduce the solution to extract knowledge objects from the web resources.

2）Knowledge Objects Extraction

As we mentioned earlier, the knowledge objects which should be extracted are research objects, research term and the relationship between objects. To extract these different knowledge units, we design different method based on some open source systems.

a) Research objects extraction: Research objects extraction is quite different from the general Named Entity Recognition (NER) tasks, such as Person, Location, Organization, Data and Numerical [9]. Because the research objects in our Ontology are often short of distinct indication, have more complicated format and often characterized with complex variability and ambiguity. It is difficult for some traditional information extraction approaches (such as rule based approach and context-based statistical approach) to extract them through the indication words. The approach we choose to use is a integrated one which includes such as lexical-pattern approach and statistical approach. In fact, we choose some matured open source software such as GATE (General Architecture for Text Engineering) 1 and Stanford Parser2 to provide basic NLP support, then we use some machine learning technologies to form objects extraction rule set semi-automatically, and also calculate the similarity of object's context to improve precise. The whole technical framework is presented in Fig. 4.

In Fig. 4, the blue modules are implemented using plug-ins of GATE directly which contain Tokenize, Sentence splitter and part-of-speech. These modules will provide basic NLP information of words such as kind,

[1] GATE Home. http://gate.ac.uk/

[2] The Stanford Parser: A statistical parser. http://nlp.stanford.edu/software/lex-parser.shtml

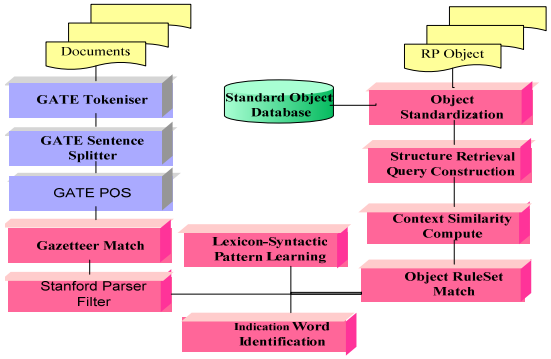


Figure 4. Framework of research object extraction

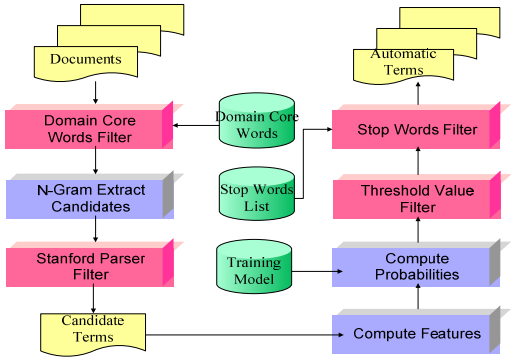


Figure 5. Framework of term extraction

orthographic, etc. The red ones are developed based on resources of GATE or developed directly by ourselves, while the green one denotes the out referring database. In this framework, we mainly use Stanford Parser to obtain the complete NP in one sentence without verb or verb phrase. Semi-learn the construction rules such as POS, kind and so on to form object rule set and some special instance expression Gazetteer via some corpus. All the Gazetteer and rule set will play important role in object recognition. Besides the special expression dictionary, we construct Standard-General expression map dictionary, stop word dictionary to identify variability and ambiguity. All the dictionaries will be added automatically based on the extraction results. More detail could be found in our related research papers.

b) Relation extraction: To mining more hidden information in document, the relation between objects is very important. According to the analysis of different sentence type which contains relation, we propose relation triple construction based on pattern and non-pattern methods and all relation triples constructed here are limited in the same sentence for reducing extraction difficulty.

To some relation triples, their elements and relation marks are fixed relatively such as the position in sentence and semantic information. Take employ relation between organization and person for example. It usually has some fixed expressions as follows:

"<research person>, <position> of <organization>"
"<position>< research person >of<organization>"
"< research person> (<organization>'s <position>)"
"<organization>'s <position>< research person>"

In such situation, we could expand some relation rule set based on Hearst pattern for extraction. But In addition to these fixed expression, there are more relations have no rules. So we design an algorithm named RelaPair for relation triples construction. The main idea of this algorithm is that each relation triples contains two objects and a relation mark word. So we could filter the unsuitable sentences which are short of the two factors. And then solve the semantic transfer problem through sentence syntax parser.

c) Research Term Extraction: To extract research term from the irregular web resources, we also choose to use an integrated approach which includes machine learning, statistic and dictionary-based method. The main features of our approach are it is based on TF*IDF feature, using first occurrence feature and take advantage of lots of NLP information. We use the KEA, an algorithm for extraction key phrases, as our basic research term extraction tool, but do lots of work to improve it. Fig. 5 shows our Improved term extraction method (Blue modules are original KEA components while red ones are new components designed ourselves). In this framework, we mainly provide mechanism to acquire output terms based on domain probability score threshold; integrate Stanford Parser to obtain the key phrase with nominal component to argument N-gram methods that KEA used; filter input documents using domain words list and use stop words list to filter extract terms.

3) Knowledge Repository Construction:

To manage and use the extracted knowledge objects effectively, we should construct a knowledge repository. Based on the Research Ontology mentioned above, we construct a knowledge repository containing all the classes and the relationship between them defined in the Ontology. After instances of research object, research term, and the relationship are extracted, they will be attached to classes defined in repository. In our knowledge repository system, it could reflect all the knowledge objects extracted. The knowledge repository provides information retrieval, inference, and statistics interface, so that it could easily be used to semantic mining and reasoning.

3.2 Semantic Mining in Knowledge Repository

Semantic mining is an effective route to achieve our research profiling goals. The core task of semantic mining is in-depth analysis and mining valuable information in our constructed knowledge repository. We try to perform it from four aspects: novelty detection, hot topic detection, timeline tracking and topic clustering.

978-1-935068-76-1 © 2011 SciRes.

The target of novelty detection task [9] in our study is identification of new or unknown research terms and research objects which our knowledge repository is not aware of before, in addition, we are trying to discover those research terms and research objects emerged recently with abrupt grownup.

1) Novelty detection

In the novelty detection task, the algorithm of detecting novelty research terms or research objects is the major challenging issue. In this paper, based on statistics, our approach to compute the novelty score of knowledge objects is an integrated method which combined with timestamp information, word frequency and domain relevance degree together. We paid attention to three points: time score, word frequency score and domain relevance score. The NoveltyScore（w）is the general evaluation of a knowledge object. As shown in formula (1) below.

$$\text{NoveltyScore（W）} = \text{TimeScore（W）} + s.\text{FrqScore（W）} + k.\text{DomainScore（W）} \quad (1)$$

Where W is an extracted research term or research objects, and W is already stored into our knowledge repository.

TimeScore（W）is a liner value with the object W's appear time, TimeScore（W）is higher if the object appear time is nearer current time.

FrqScore（W）is a comprehensive score of W which is composed of word frequency from website, RSS, Newsgroup and OAI, and we made respective frequency adjustment to a certain degree according to different data set scale, then we integrated them together.

DomainScore（W）is a comprehensive score of W which is representation of domain relative degree from website, RSS, Newsgroup and OAI. It is difficult to compute an object word how closely relative with a research area directly; however, we can compute an article or a text segment how closely relative with a research area. In our algorithm, we used a domain wordlist to help us judge an article or a text segment how relative with a research area. We considered that an article or a text segment is closer relative with a research area if it contains more words in the domain wordlist. We chose a certain number of articles or text segments randomly from four kinds of web resource respectively, and cutoff long text to avoid unfair computing, subsequently, we compute the domain relative degree of these articles or text segments, then integrated them together into a value which reflect the domain relative degree of a knowledge object.

In order to balance three different values above, we use parameter s to normalize FrqScore（w）and use parameter k to normalize DomainScore（W）.

2) Hot topic detection

Web is the sources contain so much information where media and people study on various research topics. Some topics are hot while others are unpopular. It's rather a hard work to find out hot research topics by manual way. We adopt a way which can find hot topic on web automatically, and discover topics and also judge their hot center and hotness. Terms or objects are firstly calculated based on their frequency and coverage, then a word cluster algorithm is used to judge which topic these words belong to. According to a cluster's activity, its overall frequency as well as its overall coverage, we classify topics into different hotness.

We found that a hot topic always has several characteristics, so we character them in different aspects as follows:

a) Word frequency: Word frequency is a basic and important statistics value to reflect the hot degree of a word in a specific data sets and a specific period of time. Fl(W) denotes the frequency of word W. We have conducted regular statistic about each word frequency in our basic period of time.

b) Document frequency: Document frequency is another important statistics value to reflect the coverage of a word in a specific data sets and a specific period of time. The Dl（W）is the general evaluation of a knowledge object W. In our study, it is regarded that a word with different importance when it comes from different document source. As shown in formula (2) below.

$$Dl(W) = \sum_{i=1}^{n} Li(pagerank) \quad (2)$$

Where Li denotes the source where word W comes from. The words come from authoritative web pages are allocated a larger weight, and a smaller weight will be allocated to the words which come from normal web pages. So, the Pagerank [11] value of a webpage is used for reference to evaluate the authority of a webpage. The value of Li is in direct proportion to the value of Pagerank.

c) Hot topic with a hot center: To begin, we computed the intersection of the highest word frequency and the highest document frequency word sets and selected the first 2000 for further analysis. Interestingly, there was a rather low correlation among the word frequency sets and the document frequency sets. The co-word analysis and cluster analysis was conducted for those intersection sets. Subsequently, the word with the highest comprehensive score of word frequency and document frequency in the cluster comes to be the hot center.

d) Hotness of a topic: A topic was generated after word cluster process and it contains uncertain numbers of words. In our study, we use the mean value of word frequency and document frequency each word in the cluster denotes the overall hotness of a topic. As shown in formula (3) below.

$$Hotness(topic) = \sum_{i=1}^{n} (Fl(Wi) + Dl(Wi)) \Big/ n \quad (3)$$

Besides, we have done some theoretical study and algorithm design about burst topic detection [11][12]. In further study, we plan to introduce this method and conduct more careful and refined experiments.

3) Evolution & Timeline Tracking:

The main purpose of this part is identifying what kinds of main research terms, research objects and research topics exist on timeline and how they evolve with timeline. Specifically we consider the following two tasks:

a) Term timeline tracking and object timeline tracking: Here, we tracking the evolution of each term or object extracted from text. We can observe how they developed through statistic value in different aspects, such as word frequency changing, document frequency changing, word frequency rate of increase, document frequency rate of increase, and burst period and so on.

For this task, we use word frequency rate of increase and document frequency rate of increase respectively for tracking the changing of term and object.

b) Topic timeline tracking: As we all know, any term or object is not exist independent, they are exist within some specific context or driven by some specific events. Therefore, we are studying on how to recognize topic grow and dynamic evolution [13]; besides, terms or objects are main elements of a topic inner structure, we also tracking the inner structure changing in a series of typical topics.

For this task, we proposed a series of methods mainly focus on the following three aspects: a) The dynamic evolution trend of topic scale and hotness; b) Tracking the inner structure changing in a series of typical topics; c) Discovering terms and objects which close relative with hot topic, and tracking them evolution for timely prediction.

4) Topic clustering:

We hope to discover the relations of the terms and objects by cluster extracted terms and objects. A typical method is generating the term or object co-occurrence matrix and then clustering it. We take the following cluster algorithm steps.

a) Storing the term-web document pairs into the database, and distinct them (to avoid recounting the times of terms in one web document), then select the pairs meeting our time range and term frequency requirements.

b) Sorting the selected term-document pairs by document ID. Initial a temporary list variable “temp” to store the terms belonging to the same web document, and use a hash table variable “ht” to store each term pairs’ co-occurrence frequency.

c) Generating the co-occurrence matrix file from the result in database. We consider the clustering toolkit: CLUTO3 It is a software package for clustering low or high dimensional datasets and for analyzing the characteristics of the various clusters.

The clustering method we used is k-means clustering via repeated bisections [14].In this method, a k-means solution is obtained by first bisecting the whole data. Then, one of the two Clusters is selected and further bisected, leading to a total of three clusters. The process of selecting and bisecting a particular cluster continues until k clusters are obtained. Each of these bisections is performed so that the resulting two-means clustering solution optimizes a particular criterion function. [15]

$$\text{maximize } \mathcal{I}_2 = \sum_{r=1}^{k} \sum_{d_i \in S_r} \cos(d_i, C_r). \quad (4)$$

In practice, we transfer the co-occurrence matrix into the Input format file and invoke the CLUTO in java like this:

V cluster filename k (5)

4 Implementation and Evaluation

4.1 Web Resource Collection

At present, we choose "Artificial Intelligence" domain as our monitor target. We have obtained 444 web site seed and 166 RSS harvest seed related to Artificial Intelligence domain. In Web resource harvest aspect, we have harvest above web site twice a week and harvest 20 batch data until now, the harvested number of pages reached 1,273,037. After the filter steps, the number of unrepeated new web pages reaches 134368. In RSS harvest aspect, we harvest RSS seed site every day. Now we get totally 18,591 RSS records. In Newsgroup harvest aspect, we harvest 535 news server seed site every day and till now we get totally 6,467,448 articles.

4.2 Knowledge Extraction and Knowledge Repository Construction

After filtering and cleaning of each batch harvested data, knowledge objects are extracted. We designed a multi-machine multi-process extraction deployment for scalability. The extraction statistics are shown in the table1. And all extracted knowledge objects are stored into our knowledge repository with timestamps and organized according to Research Ontology defined above. We can retrieve the knowledge repository and use extracted objects for semantic mining and reasoning.

Table 1. The statistics of knowledge objects extracted from repository

	Document Count	Research Objects	Research Object Instances	Research Terms	Research Term Instances
Web	134,368	273,240	3,311,923	107,928	3,311,923
RSS	18,591	19,836	69,997	11,404	69,997
Total	152,959	293,076	3,381,920	119,332	3,381,920

[3] CLUTO - Family of Data Clustering SoftwareTools. http://glaros.dtc.umn.edu/gkhome/views/cluto/

978-1-935068-76-1 © 2011 SciRes.

4.3 Semantic Mining and Profiling

For Panoramic Perspective Profiling task, through analyzing and scoring changes of each research object along the time axis, we can find out the most important research objects in Artificial Intelligence domain, these important research objects help us to know AI domain from overall perspective as Fig. 6 depicts.

Figure 6. Panoramic perspective profiling in "Artificial Intelligence" domain

For evolution and timeline tracking task, we use curve figures to help intuitively understand the historical development of specific knowledge object, and predict the future development trends of a knowledge object. In Fig. 7, it is the frequency change curve with timeline of Knowledge Object "Japanese Robotics Institute".

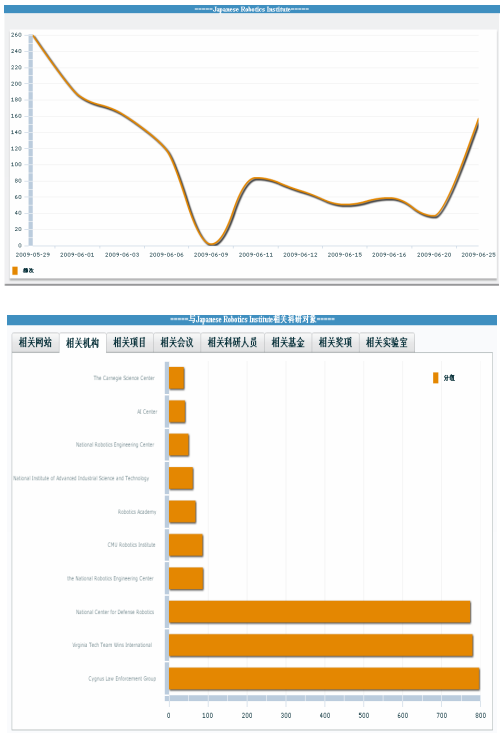


Figure 7. Knowledge object timeline tracking and relevant knowledge object

After the analysis of relationship among extracted knowledge objects, we connect each knowledge object into relation networks. Through this processing, the profiling of individual knowledge object is no longer isolated. For example, the following figure shows relevant research organization statistical figure of organization object "Japanese Robotics Institute". In addition, we also can obtain the statistical figure of relevant websites, relevant projects, related conferences, relevant research personnel, related funds, related awards and related laboratories.

For hot topic detection task, in order to show the hot topic structure of Artificial Intelligence domain, we have clustered the extracted research terms with top 2000 terms with high word frequency and document frequency, and cluster twice at two levels, each level is clustered 10 partitions using above method. The following table shows the result after the first clustering.

The clustering results are visualized as showed in Fig. 8. From the Fig. 8, we can clearly see the hot center of each topic and the scale of each topic.

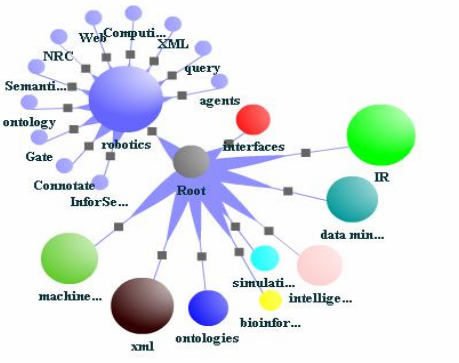


Figure 8. Hot topic clustering result

As we can see from Fig. 8, based on our collected web resources, the first level hot research topics in AI domain is robotics, machine learning, xml, ontologies, bioinformatics, intelligence, simulation, data mining, IR and interface. In these hot topics, Robotics has the highest hotness and the second level hot topics contained in it as shown from Fig. 8.

5 Conclusion

In our research profiling experiment we can get a comprehensive view of the overall development of AI domain. We can see the evolution history of a specific object in timeline; find related objects to a specific knowledge object; and learn hot topics in AI domain. We can also discover novel terms and objects, hot terms and objects in AI domain. However, there are still many problems need for further explore. For example, now we only calculated whole period clustering result based on research terms. In future, with the continuous accumulation of data, we hope to cluster the research terms periodically

and tracking timeline of topic, and find topic changes through comparing every clustering result. Another important and hard-working task is to try to refine the extraction result to improve the performance.

According to our investigating on Artificial Intelligence research area, it is clear to see that new research interest and advanced technology are possible became the driver of AI research development. Emerged new terms or objects could represent a valuable signal to emerging research trends, and hot topics reflect the research focus and research interest. Our experiment result conforms to the development status and trends of AI domain. So, we believe that research profiling based on semantic web mining is an important and effective method to continuously monitor a research area.

6 Acknowledgements

This paper is supported by a project of National Key Technology R&D Program in the 11th Five year Plan of China named Science Monitoring and Evaluation based on Scientific Web Resources (2006BAH03B05)

References

[1] Bragge, J., Sami R., et al. Enriching Literature Reviews with Computer-Assisted Research Extraction. Case: Profiling Group Support Systems Research. Proceedings of the 40th Annual Hawaii International Conference on System Sciences, Hawaii 2007.

[2] Porter, A. L., A. Kongthon, et al. "Research profiling: Improving the literature review." Scientometrics, Vol.53.No.3, 2002, 351-370.

[3] Bragge, J., Jan S. Profiling Academic Research on Digital Games Using Text Extraction Tools. Proceedings of DiGRA Conference, Japan, 2007.

[4] Hicks, D., G. Atlanta. "Global Research Competition Affects US Output". School of Public Policy, Georgia Institute of Technology, Atlanta, 2004.

[5] J. Webster and R.T. Watson, "Analyzing the Past to Prepare for the Future: Writing a Literature Review", MIS Quarterly, Vol. 26, No. 2, 2002.

[6] Alan L. Porter, David J. Schoeneck, et al. Mining the Internet for Competitive Technical Intelligence. Competitive Intelligence Magazine, Vol. 10, No. 5 , 2007, pp25-28.

[7] Tim Berners-Lee, James Hendler, Ora Lassila. The Semantic Web. Scientific American, No. 5, 2001.

[8] Nenadic, G., Irena S., et al. Automatic discovery of term similarities using pattern mining. http://acl.ldc.upenn.edu/coling2002/workshops/data/w05/w05-08.pdf.

[9] Markou, M. and S. Singh, Novelty detection: a review—part 1: statistical approaches. Signal Processing, Vol. 83, No. 12, 2003, pp2481-2497.

[10] Langville, A.N., C.D. Meyer, and P. Fern ndez, Google's PageRank and beyond: thescience of search engine rankings. The Mathematical Intelligencer, Vol. 30, No. 1, 2008, pp 68-69.

[11] Jon Kleinberg. Bursty and Hierarchical Structure in Streams. Data Mining and Knowledge Discovery, Vol. 7, No. 4, 2003, pp373-397.

[12] Qi He, Kuiyu Chang, Ee-Peng Lim.Using Burstiness to Improve Clustering of Topics in NewsStreams. ftp://ftp.computer.org/press/outgoing/proceedings/icdm07/Data/3018a493.pdf.

[13] Rudy Prabowo, M. Thelwall, Mikhail Alexandrov. Generating overview timelines for major events in an RSS corpus. Journal of Informetrics, No. 1 ,2007, pp131–144.

[14] Zhao, Y., G. Karypis, and U. Fayyad, Hierarchical clustering algorithms for document datasets. Data Mining and Knowledge Discovery, Vol. 10, No. 2 ,2005, pp141-168.

[15] CLUTO—A Clustering Toolkit. http://glaros.dtc.umn.edu/gkhome/fetch/sw/cluto/manual.pdf.

The Review of Acupuncture Research Development Based on SCI Records

Weixiang SONG[1], Jia JIA[2], Na WANG[3]

[1]*Beijing City University, Beijing, China*

[2]*Department of Information Management, Peking University, Beijing, China*

[3]*Institute of Scientific and Technical Information of China, Beijing, China*

Email:jane00800d@163.com

Abstract: By analyzing the data based on SCI records, this article details the situation of acupuncture on the aspects of publication year, author's country, author's affiliation and the high-cited papers etc. It provides the objective foundations to the study of acupuncture.

Keywords: SCI; acupuncture; bibliometrics; review

基于 SCI 数据的针灸研究发展态势综述

宋维翔 [1]，贾佳 [2]，王娜 [3]

[1] 北京城市学院，北京，中国，100083

[2] 北京大学信息管理系，北京，中国，100871

[3] 中国科学技术信息研究所，北京，中国，100038

Email: jane00800d@163.com

摘　要：本文采用 SCI 数据，分别分析了 1921 年至 2009 年以来以“针灸”为主题的论文发表年代、作者国别、所属机构以及高被引论文的相关情况，为了解针灸领域的研究发展态势提供了客观依据。

关键词：SCI；针灸；文献计量学；综述

1 引言

针灸疗法是祖国医学遗产的一部分，也是我国特有的一种民族医疗方法。千百年来，对保卫健康，繁衍民族，有过卓越的贡献，直到现在，仍然担当着这个任务，为广大群众所信仰。新中国成立后，由于党的中医政策的实施，带来了针灸事业的复兴与繁荣。全国各地先后成立了中医学院、中医院，设置了针灸专业和专科，并建立了专门研究机构，使针灸在教学、医疗和科研等方面都获得了巨大的成就，在全国各报刊发表的针灸论文资料多达几万篇，大大丰富了针灸医学的内容。随着我国在世界范围影响力的扩大，西方各国都对针灸产生了浓厚的兴趣，逐渐认可并投入越来越多的资源来研究该领域相关知识。

本文利用文献计量法统计了 1921 年至 2009 年 3 月 SCI 数据库中针灸研究领域发表的论文,试图通过回溯分析，揭示该课题研究的发展轨迹、研究路线、学术成果及存在相关问题，并对其发展做出展望。

2 数据采集和处理

为获得全面的数据，笔者查阅各种文献，选取了 Acupuncture、Moxibustion、Electro-Acupuncture 等几个表达“针灸”含义的关键词，在 Web of Science 数据库中进行检索并作去重处理，有效文献共计 7908 篇（含所有文献类型）。随后，笔者使用美国 Thomson 公司开发的专利分析工具 Thomson Data Analyzer (TDA)，对数据进行深度挖掘并提供可视化分析。

3 SCI 数据分析

3.1 论文年代分析

Table 1. Time distribution of acupuncture field papers between 1921 and 2009

表 1. 1921-2009 年针灸领域论文年代分布情况

年份	国际论文数量	国内论文数量
1921-1950	5	0
1950-1960	3	0
1960-1970	13	0
1970-1980	1022	27
1980-1990	1235	37
1991-2000	1891	104
2000-2009	3738	492

表 1 数据表明，1921－2009 年，针灸领域的 SCI 论文共计 7908 篇。中国于 1973 年发表了该领域的第一篇 SCI 论文。1970 年以后，全世界关于针灸领域的论文数量均呈现激增趋势。这一数据充分表明，对于针灸领域的理论及技术研究一直在不断地发展和成熟。但我国关于针灸领域的 SCI 论文数量较少，并且增长速度缓慢，直到 2000 年后才有明显提升。

笔者查阅各种资料，发现 1970 年这个论文数量转折点的产生可能有以下两个原因：第一，1971 年，新华社首次向全世界宣布了针刺麻醉这一伟大成就，这一消息的发布，不仅推动了国内的针刺麻醉研究和应用热潮，也第一次使中国的传统医学在国外产生了重要、深远的影响。第二，中美关系开始缓和后，尼克松总统访华之前的 1971 年 7 月，美国著名记者詹姆斯·赖斯顿被派往中国采访，在访问时不幸患了急性阑尾炎，在协和医院接受了阑尾切除手术治疗，术中使用的是常规药物麻醉，术后感到腹胀不适，接受了针灸治疗，之后于 1971 年 7 月 26 日在纽约时报上发表了那篇著名的纪实报道：“现在让我告诉你们我在北京的阑尾切除手术”。由于这位资深记者的不凡经历和纽约时报在新闻界的地位，在一般美国人心目中，这篇文章的可信度是极高的，被认为是美国大众传媒第一次向美国公众介绍中国针灸疗法，也是针灸传入美国的历史性标志。

3.2 针灸领域 SCI 论文国别分析

全部论文共涉及 79 个国家和地区，而论文数量最多的 TOP20 国占了全部论文的 72%(见表 2)，可见，针灸领域的研究工作主要集中在 TOP20 国中。尤其是美国凭借其强大的科研实力，发表的论文数占世界论文总数的 22.6%，中国虽位居第二，但文章数量距美国还相差甚远。

值得一提的是，中国台湾地区的论文数量位居第九，原因在于 7O 年代台湾各大专院校掀起一股针灸热，纷纷成立中医针灸相关社团，在各大学校园内带领风骚。不只是医学院，连一些文、法学院为主的大学，如辅仁大学等也都有相关的社团成立，一时蔚然成风，也造就了不少中医针灸人才。目前，中国医药学院附设的医院中医部、长庚大学中医分院及公私立各大医院中医科均有针灸治疗服务，成为培养针灸临床医师的重要场所。针灸治疗从此正式纳入医疗教育体系中，显示台湾针灸教育发展的丰硕成果[1]。

Table 2. National distribution of acupuncture field papers between 1921 and 2009

表 2. 1921-2009 年针灸领域论文国别分布情况

排名	国别	文章数
1	美 国	1786
2	中 国	660
3	英 国	547
4	德 国	436
5	韩 国	342
6	日 本	320
7	加拿大	218
8	瑞 典	215
9	中国台湾	204
10	奥地利	155
11	澳大利亚	151
12	意大利	104
13	法 国	94
14	瑞 士	78
15	巴 西	74
16	挪 威	67
17	西班牙	66
18	以色列	57
19	丹 麦	56
20	荷 兰	56

3.3 针灸领域 SCI 论文机构分析

3.3.1 国际机构分析

从针灸领域 SCI 论文数量的世界 TOP20 科研机构（见表 3）分布来看，韩国的庆熙大学摘取了世界第一的桂冠，美国有 8 所大学入围 TOP20 机构，为所有国家中数量最多者，中国位居第二，有 4 所大学入围。然而 TOP20 的科研院所全部来自于论文发表数 TOP10 的国家，仍然维持了针灸领域研究强势国家统领大局的模式。

排名第一的庆熙大学在韩国学研究方面成绩突出，在韩医学方面的研究不仅在韩国，而且在国际上名列前茅。庆熙大学东西医学研究所是世界卫生组织传统医学在西太平洋地区的合作中心之一。排名第二的埃克斯特大学士成立于 19 世纪中叶的英国一流大学，其与普利茅斯大学（University of Plymouth）合建的半岛医学院（Peninsula Medical School）在英国 Maurice Laing 基金的资助下于 1993 年成立了一个补充医学研究小组，这是第一个英国大学设立的补充医学研究机构。现在它已经发展成为一个国际认可的补充替代医学（complementary and alternative medicine, CAM）科学研究中心[2]。排名第三的马里兰大学，其医学院的补充医学中心成立于 1991 年，是美国在西医学院成立的第一个以研究中医药为主的综合性补充和替代医学中心。从 1995 年起连续被美国国家健康研究

978-1-935068-76-1 © 2011 SciRes.

院评为全国中医研究的杰出中心，研究主要集中在评估中医药、针灸和身心疗法。中心近年来还增加了针灸和中医药研究在国际间的交流合作，特别与中国的香港中文大学、上海中医药大学和其他中医药大学进行了合作[3]。

Table 3. World TOP 20 research institutes (Sorting by SCI publications between 1900 and 2009)
表 3. 世界 TOP 20 科研院所（按 1900--2009 年 SCI 论文数量排序）

排 名	机构名称	国 别	论文数
1	庆熙大学	韩 国	77
2	埃克斯特大学	英 国	75
3	马里兰大学	美 国	57
4	华盛顿大学	美 国	49
5	加州大学洛杉矶分校	美 国	46
6	德克萨斯大学	美 国	46
7	哈佛大学	美 国	42
8	维也纳大学	奥地利	42
9	首尔国立大学	韩 国	38
10	北京大学	中 国	37
11	复旦大学	中 国	36
12	哥德堡大学	瑞 典	36
13	香港大学	中 国	35
14	宾夕法尼亚大学	美 国	35
15	卡罗林斯卡学院	瑞 典	33
16	中国医科大学	中 国	29
17	耶鲁大学	美 国	28
18	哥伦比亚大学	美 国	27
19	格拉茨大学	奥地利	27
20	慕尼黑理工大学	德 国	27

3.3.2 国内机构分析

国内的 TOP20 科研机构（见表 4）论文数量差距较大，北京大学以 37 篇位居第一，排名 10 至 20 位的科研机构论文数量较少，差异不大。说明中国的针灸领域研究与世界先进水平的距离还很远。虽然针灸是中华民族的传统医学，但在该领域的研究却有待提升理论起点，并应在进一步扩大数量规模的基础上，向该领域研究的纵深地带挺进。

Table 4. Domestic TOP 20 research institutes (Sorting by SCI publications between 1900 and 2009)
表 4. 国内 TOP 20 科研院所（按 1900--2009 年 SCI 论文数量排序）

排 名	机构名称	论文数（篇）
1	北京大学	37
2	复旦大学	36
3	香港大学	35
4	中国医科大学	29
5	香港理工大学	18
6	上海医科大学	17
7	中国科学院	17
8	四川大学	10
9	天津中医药大学	9
10	中国中医科学院	8
11	广东邦民制药有限公司	8
12	中国医学科学院	7
13	天津大学	7
14	武汉大学	7
15	南京大学	6
16	上海交通大学	6
17	燕山大学	6
18	浙江大学	6
19	第四军医大学	5
20	华中科技大学	5

3.4 针灸领域 SCI 期刊源统计分析

本文所分析的 7908 篇文章共发表在 1259 种期刊上，排名前 20 的期刊（见表 5）共 2622 篇，占总数的 33.16%。发表文章最多的期刊是《美国针刺疗法杂志》，该杂志主要刊载针刺和相关的物理疗法以及中西结合针刺治疗方式及其应用的理论与临床研究论文、病例报告、札记和文摘。共发文 590 篇。另外 1239 种期刊共发表 5286 篇论文，其中 592 种期刊只发表过 1 篇论文。这些论文大多数刊登在医药类刊物上，另外一些论文零散地分布在大学学报、其他专业类期刊上。中国发文数最多的期刊是《中华医学杂志》，该杂志创刊于 1915 年，于 1920 年被美国 Chemical Abstract 收录，1927 年起就被美国医学会 Quarterly Cumulative Index Medicus 收录，是中国最早被国外医学索引收录的中文版医学期刊。

Table 5. Journal distribution of SCI papers
表 5. SCI 论文期刊分布情况

排名	论文数	期刊
1	590	Am. J. Acupunct
2	332	Am. J. Chin. Med
3	300	Acupunct. Electro-Ther. Res
4	275	Journal of Alternative and Complementary Medicine
5	112	Pain
6	106	Complementary therapies in medicine
7	103	British Medical Journal
8	94	Neuroscience Letters
9	78	Medical Journal of Australia
10	76	Journal of the American Medical Association
11	72	Anesth. Analg
12	68	Lancet
13	61	New Zealand Medical Journal
14	60	Chinese Medical Journal
15	54	Altern. Ther. Health Med
16	51	CLINICAL JOURNAL OF PAIN
17	49	Brain Research
18	48	Zhurnal Nevropatol. Psikhiatrii Im. S S Korsakova
19	47	Ann. Intern. Med
20	46	Arch. Phys. Med. Rehabil

3.5 高被引论文情况

对于一个国家、一个学科或者一位研究人员来说，仅有论文数量还不足以体现科研实力，更要看论文质量，而被引频次是论文质量比较直接的量化体现。笔者分别选取了论文数量排名前十国家的被引频次 Top100 的论文，并分别得出了每个国家高被引论文的

平均被引频次（见表 6）。结果显示，美国以每篇平均被引 80 次遥遥领先位居第一，英国、加拿大和瑞典分列其后，中国虽名列第五，但被引频次仅为 16 次，距离前几位的国家差距还比较大。

Table 6. National analysis of high cited papers
表 6. 高被引论文国别分析

排　名	国　别	被引 Top100 论文平均被引频次
1	美国	80
2	英国	36
3	加拿大	22
4	瑞典	21
5	中国	16
6	德国	15
7	中国台湾	9
8	日本	8
9	韩国	7
10	奥地利	4

4 小结

我国有关针灸领域的科技论文数量自 1970 年起有明显增长，但具备国际影响力的 SCI 论文数量较少，在世界排名第二，与排名第一的美国论文数量相比仍有较大差距。世界范围内对于针灸领域研究成果较多的前 20 位机构中，我国共有 4 所高校入围，最靠前的是北京大学，以 37 篇论文排名第 10。我国在平均被引频次国际排名第 5，前三名分别为美国、英国和加拿大。数据表明，在针灸研究领域，我国不仅要增加论文数量，更要提高论文质量，扩大影响力范围。

References（参考文献）

[1] Zhaoming Chen. The Review of Traditional Chinese Medicine Acupuncture Development [J]. Ancient Medicine Knowledge, 2005,(2):40.
陈昭明. 台湾中医针灸发展概况[J]. 医古文知识, 2005,(2):40.

[2] Junying Fu. The Main Institute of Complementary Medicine in UK(1)-- The University of Exeter and The University of Southampton[J]. Journal of Chinese Integrative Medicine, 2009，(3)：291.
傅俊英.英国补充替代医学研究的主要机构（一）--埃克斯特大学和南安普敦大学[J].中西医结合学报，2009，(3)：291.

[3] Lixing LAO, Brian M.BERMAN. The Center for Integrative Medicine at the University of Maryland:the first complementary and alternative medicine center in a US medical school [J]. Journal of Chinese Integrative Medicine, 2008, 6(11): 1205.
劳力行，布赖恩·伯曼. 马里兰大学医学院结合医学中心—美国医学院最早成立的替代医学中心[J]. 中西医结合学报，2008，6(11)：1205.

978-1-935068-76-1 © 2011 SciRes.

A Query Translation Disambiguation Method for Cross-Language Information Retrieval Based on Hybrid Corpora

Yingfan GAO, Yanqing HE, Hongjiao XU, Huilin WANG
Institute of Scientific and Technical Information of China, Beijing, China, 100038

Abstract: Disambiguation between multiple translation choices is very important in dictionary-based cross-language information retrieval. A hybrid corpora-based method for translation disambiguation is proposed in the paper. Provided an English-Chinese bilingual dictionary in scientific and technological fields, parallel corpus was employed to append translation probability value to bilingual dictionary. Moreover, parallel corpus could contribute to solve the problem of OOV words with the help of toolkits (such as GIZA++). Estimating the CLIR system in experiment corpora is another important question. To establish test collections in small-scale document set, the user queries' coverage rate and hit rate about document titles are proposed and it's shown that the selection of the two parameters could improve detective rate of information retrieval systems greatly. At present, the test collection constructed has been used in testing CLIR systems successfully, and the experimental results demonstrate the methods of query translation and translation disambiguation for CLIR in the paper is feasible.

Keywords: translation disambiguation; translation probability; parallel corpus; test collections

1 Introduction

The basic idea of Cross-Language Information Retrieval(CLIR) is to bridge the language boundary by providing access in one language(the source language) to documents written in another language(the target language), by using query translation from the source language into the target language and/or document translation from the target language. The main strategies for query translation are based on three different methods: dictionary-based methods with specific relevance to (bilingual) translation dictionaries, corpus-based methods, and machine translation[1][2]. Dictionary-based translation has been widely used for CLIR, especially in cases where a high- performance machine translation (MT) system is not available. In general, most retrieval systems are still based on so called "bag-of-words" architectures, in which both query statements and document texts are decomposed into a set of words(or phrases) through a process of indexing. Thus we can translate a query easily by replacing each query term with its translation equivalents appearing in a bilingual dictionary or a bilingual term list[3]. Ballesteros[4] pointed out problems with this method as follows:

- Specialized vocabulary not contained in the dictionary will not be translated.
- Dictionary translation are inherently ambiguous and add extraneous information.
- Failure to translate multiterm concepts such as phrases reduces effectiveness.

The questions above result in the poor performance of CLIR system than mono-language information retrieval (MLIR) system. Researchers concentrate on reducing the ambiguity from dictionary-based methods[6]. Hull[5]'s experiments prove that translating the phrase that is composed of multi-nouns can improve the system performance effectively. Boughanem[7] proposed to use bi-directional translation-based strategies to solve the problems of English-French query translation provided by a bilingual dictionary. Dong Zhou[8] marries a graph-based model and a pattern-based method for dictionary-based query translation suitable for English-Chinese CLIR. Employing many kinds of translation resources provides new ideas for dictionary-based query translation and would be the basis of this paper.

This work has been supported by Postdoctoral Science Foundation of China (20090450465), Key Task Project in 2010 of ISTIC(ZD2010-3-3) and Subject Building Project in 2010 of ISTIC(XK2010-6).

2 Identifying OOV Words Using Parallel Corpora

The bilingual dictionary available is in the scientific and technological fields and there are many multi-term phrases in it. So, identifying and translating multi-term concepts of documents and matching them with dictionary would be the main solution to the question (3) at section 1. The method we used is Maximum Forward Matching. Before matching, some preprocessing steps have been done, including lowercase transforming, stopwords filtering, stemming, and lemmatizing, et al. To another two questions, a method based on hybrid corpora is proposed in the paper.

Given a finite vocabulary, the presence of out-of-vocabulary (OOV) words is inevitable. OOV words constitute a major source of error in query translation. The vocabularies in bilingual dictionary are limited with respect to magnanimity literature. To solve the question,

we use large-scale parallel corpus to complement information.

"Translation" here means mapping term statistics from one language to another, not simply replacing the terms themselves. Translation probabilities can be estimated from parallel corpora. With sentence-aligned parallel corpora, the freely available GIZA++ toolkit[9] can be used to train translation models. GIZA++ produces a representation of a sparse translation matrix and many words, forms, or expressions are incorrect or unacceptable. So we remove these words first according to some prearranged rules. For example, if a digit and a word are linked together without space or words with different language are linked together without space, there would be wrong. Then, after data cleaning, each word's translation probabilities would be recomputed according to the principle of normalization. Moreover, to further reduce the translation noise, we could set threshold of translation probabilities, such as 0.3.

Coming from parallel corpora, translation probability dictionary could solve the OOV problem to a certain extent and would be beneficial supplement to the bilingual dictionary in scientific and technological fields.

3 Improving Bilingual Dictionary with probability value

There are multi-translation entries for a word (or a phrase) in the bilingual dictionary usually and it's difficult to distinguish these different translations. Section2.2 tells us how to use the parallel corpora to produce a translation probability dictionary. If the words in bilingual dictionary are in accordance with the words in translation probability dictionary, the former's translation entries would be appended with probability values coming from the latter's translation matrix. When the translation entries in the bilingual dictionary are not absolutely the same as the translation probability dictionary, the former's probability value would be recomputed according to the principle of normalization.

We couldn't use the method above to improve the bilingual dictionary available completely. In the bilingual dictionary that we use is in scientific and technological fields, there are many multi-word phrases in it and each multi-word phrase might be of multi-translation entries. The method in section 2.2 use GIZA++ toolkit and could produce translation probability only at word level, not at phrase level. We have to reuse the parallel corpora to make the phrases' probability value. The new method could be described as follow:

Input: Sentence-aligned parallel corpora; English-Chinese bilingual dictionary

Output: Bilingual Dictionary with probability values

Steps:

- Take an English term *ET* from the English-Chinese bilingual dictionary, and simultaneously take one or multi-Chinese translation entries CT_i(i=1,2,…k) corresponding to *ET* where k is the number of CT_i to *ET*.
- According to sentence-aligned parallel corpora, we get the number $n_{CT_i|ET}$ which is the occurrence times between CT_i and *ET* in English-Chinese parallel sentence pairs.
- When *ET* is translated to CT_i , we could compute the translation probability value according to formula as follows:

$$P(CT_i \mid ET) = \frac{n_{CT_i|ET}}{\sum_{i=1}^{k} n_{CT_i|ET}} \quad (1)$$

- In the Chinese-English direction, the method is like steps above.

We consider, to a word or a phrase in dictionary, the different probability value of translation entry reflects its translation habit in reality. The bigger the translation probability value, the higher the possibility that the word or the phrase is translated to the translation entry. And these appended information all comes from the parallel corpora and could help us reduce the ambiguity of translation.

4 Test for Cross-language Information Retrieval

4.1 Constructing Test Collection

- Document Set

The background of this paper is in the scientific and technological fields, not in the journalism field. So the test collections for the famous CLIR test conferences, such as NTCIR,TREC et al., would not practicable for our experiments. Moreover, because of lacking test documents of immense amounts, the study in this paper emphasized on feasibility of constructing the test collection for CLIR system based only on small-scale corpus. We have got the test documents from WanFang DATA company (www.wanfangdata.com.cn) and all documents are dissertation abstracts in the scientific and technological fields. The sum is in close proximity to 100,000, which includes almost 50,000 English documents and 50,000 Chinese documents .

- Query Set for CLIR

We got the users' query log during 2009 from WanFang Data company too. By data cleaning and random sampling, about 1000 queries were reserved. The main problem here is the large differences in coverage between these queries and document set. The coverage of the former is large and plain, but the latter is small and profound. If improper queries are selected, the search engine would return 0 results by searching the document set above. We proposed to use two parameters to solve the problem. They are the user queries' coverage rate *COV* and hit rate about document titles *HIT*.

$$Cov_i = N_{Q_{HIT}>2} \quad (2)$$

978-1-935068-76-1 © 2011 SciRes.

$$HIT_i = \frac{Q_{HIT_i}}{Q_i} \quad (Q_{HIT_i>2}) \tag{3}$$

where Q_i represents the number of words in query i , Q_{HIT_i} represents the number of words of query i which appears in title of document, and $N_{Q_{HIT}>2}$ represents the number of documents whose title contains two or more words in query i.

Compared to *COV*, *HIT* could reflect the co-occurrence of words in query and documents' title better. So *HIT* would be the main parameter in selecting query set, and *COV* are the assistant parameter. According to these principles, we extract query-needed in 1000 candidate queries.

- Related Documents

The most popular method for looking for related documents is called *Pooling*. It means that the forward section of the search results of participants would be extracted and merged into a documents *pool* and this *pool* could be viewed as the possible related documents candidate. For our experiment system, there are not many participants for testing. In order to assure fair test, we extract searching results from multi-searching engine and put them in *pool*. The annotation specifications of related documents are just like NTCIR and TREC. Bilingual topic statement and their human translation have been made and professional annotators have annotated about 2000 documents coming from documents set.

4.2 Experiments and Analysis

According to test collection in section4.1, we select 10 Chinese queries from query set for CLIR experiment. The testing index in the paper would be the cross-language search efficiencyη, which is the ratio of average precision of CLIR system to MIR system.

$$\eta = \frac{\eta_{CLIR}}{\eta_{MIR}} \tag{4}$$

whereη_{CLIR} is the average precision of CLIR system andη_{MIR} is the average precision of MIR system. The precision *P* of information retrieval system could be estimated as follows:

$$P = \frac{N_r}{N_r + N_{nr}} \tag{5}$$

where N_r is the number of related documents in searching results and N_{nr} is the number of non-related documents in searching results. For simplicity, we count the number on the basis of the first 10 search results. The result of experiments could be shown in table1.

In accordance with in formula (3) and (4) and data in table1, the results are as follows:

η_{MIR}=38%

η_{CLIR} =25%

η=66.3%

The result of experiments in table1 can be described in terms of the following categories.

- Inadequate corpora for experiments results in the 0 precision. For example, when Chinese query is "有限元分析", the translation result by method in this paper is just the same as the human translation. Because of no related documents in corpora, no search result returned. The result couldn't become the evidence that the performance of CLIR system is poor than MLIR system.
- The coverage of Chinese queries translated by method in the paper, such as "ARM 微处理器","高分辨率", are larger than human translation. By rights, the result of experiment of the former should be better, but it didn't turn out as we intended. Again, inadequate corpora became the killer.
- The problem of OOV. Whatever corpus was used by us, OOV words are inevitable. For example, the "无线" in "无线家庭网络" in table1.

There are 10 results of experiments in table1, 7 of which indicate the precision of CLIR reached over 75% that of MIR. The average precision of CLIR reached 66.3% that of MIR. The experimental results demonstrate the methods of query translation and translation disambiguation for CLIR in the paper is feasible.

5 Conclusion

With the development of multilingual language information resource, the user requirement for Cross language information retrieval has gone up continuously. Disambiguation between multiple translation choices is very important in dictionary-based cross-language information retrieval. A hybrid corpora-based method for translation disambiguation is proposed in the paper. Provided an English-Chinese bilingual dictionary in scientific and technological fields, parallel corpus was employed to append translation probability value to bilingual dictionary. To a word or a phrase in dictionary, the different probability value of translation entry reflects its translation habit in reality. The bigger the translation probability value, the higher the possibility that the word or the phrase is translated to the translation entry. Moreover, parallel corpus could contribute to solve the problem of OOV words with the help of toolkits (such as GIZA++).

Estimating the CLIR system in experiment corpora is another important question in the paper. To establish test collections in small-scale document set, the user queries' coverage rate and hit rate about document titles are proposed and it's shown that the selection of the two parameters could improve detective rate of information retrieval systems greatly. At present, the test collection constructed has been used in testing CLIR systems successfully, and the experimental results demonstrate the methods of query translation and translation disambiguation for CLIR in the paper is feasible.

Table 1. Result of experiment for CLIR

Chinese Query	Human Translation	Translation by method in this paper	Precision in MIR	Precision in CLIR	CLIR/MIR
高分辨率	High Resolution	high definition ;hi res ; high resolution ; high resolution capacity	0.484536082	0.422680412	87.23%
无线家庭网络	Wireless Home Network	home network	0.3571	0.1607	45%
有限元分析	Finite element analysis	finite element analysis	0	0	/
ARM 微处理器	ARM Microprocessor	arm micro processor ; arm micro multiprocessor ; arm micro handler	0.8889	0.1111	12.5%
PWM 整流器	PWM Rectifier	pwm rectifier ; pwm electric commutators	0.229166667	0.208333333	90.9%
存储区域网 SAN	Storage Area Network	store regional net san	0.4706	0.3529	75%
网络体系结构	Network Architecture	network architecture	0.1579	0.1579	100%
商业银行风险	Commercial bank risk	commercial bank risk	0.2778	0.2778	100%
网络营销模式	Internet Marketing Model	net marketing model	0.2222	0.2222	100%
企业库存管理	Inventory Management	corporations inventory control	0.7188	0.6094	85%

References

[1] Oard, D. W., & Diekema, A. R.. Cross-language information retrieval. Annual Review of Information Science and Technology, 1998, 33: 223-256.

[2] Pirkola, A., Hedlund, T., Keskustalo, H., Javelin, K. Dictionary-based cross-language information retrieval: problems, methods, and research findings. Information Retrieval, 2001,4: 209–230.

[3] Kazuaki Kishida. Technical issues of cross-language information retrieval: a review. Information Processing and Management, 2005, 41: 433-455.

[4] Ballesteros, L., & Croft, W. B. Resolving ambiguity for cross-language retrieval. Proceedings of the 21st ACM SIGIR conference on research and development in information retrieval ,1998:64-71.

[5] Hull, D. A., Grefenstette, G. Querying across languages: a dictionary-based approach to multilingual information retrieval. Proceedings of the 19th ACM SIGIR conference on research and development in information retrieval .1996: 49-57.

[6] Kazuaki Kishida, Emi Ishita. Translation disambiguation for cross-language information retrieval using context-based translation probability. Journal of Information Science, 2009,35(4):481-495.

[7] Boughanem, M., Chrisment, C., & Nassr, N. (2002). Investigation on disambiguation in CLIR: aligned corpus and bi-directional translation-based strategies. In C. Peters, et al. (Eds.), Evaluation of cross-language information retrieval systems. LNCS (2406,pp. 158-168).

[8] Dong Zhou,Mark Truran, Tim Brailsford, et al. A Hybrid Technique for English-Chinese Cross Language Information Retrieval. ACM Transactions on Asian Language Information Processing. Vol 7, No 2, Article 5. June2008.

[9] Och, F. and Ney, H.. Improving statistical alignment models. ACL, Hong Kong, 2000: 440-447.

978-1-935068-76-1 © 2011 SciRes.

Mapping and Implementation from Chinese Natural Language Queries to nRQL

Qian GENG, Ming ZHANG, Tao YIN
Management School, Beijing Normal University, Beijing, China
Email: gengdaqian@263.net

Abstract: This paper mainly discusses the grammar of ontology query language nRQL and the conversion from users' Chinese natural language queries to nRQL. Furthermore, we design a series of question regulations based on limited natural language for the query processing experiment. Based on searching for the query target and query condition, this research processes some of grammatical phenomenon in Chinese through semantic understanding and completes the conversion from query to nRQL successfully.

Keywords: Chinese natural language queries; nRQL; queries mapping; information retrieval

1 Process of Natural Language Retrieval Research

Natural language retrieval is defined as the retrieval behavior conducted in natural language index or queries environment, whose basic feature is that users ask questions by means of natural Language, rather than single keyword or retrieval logic formed from that. The form of natural language queries can be a phrase, sentence or a paragraph, which can express the user's real needs more directly. Compared with controlled retrieval, it has a lower requirement for the index process background and frees professionals from indexing work.

Consequently, it is more suitable for users' retrieval under current Web environment, without any expert guidance. The flexibility of its allowed form improves user-friendly at the same time. However, natural language retrieval transfers some manual work ---- including requirements analysis, specification control of search terms, etc. ---- to computer, which will require more on the processing ability of machine---- especially the ability of natural language processing. The system needs to analysis and understand the content of users' queries by means of algorithm design, and implements a good control on the searching process.

The earliest research work on natural language retrieval is to apply NLP application to areas of information retrieval, which can be found from the early research of automatic retrieval in 20th century 60's to 70's. The emergence of natural language retrieval has a compact relation with the development of NLP. After years' research, there have been some achievements in this field. However, the processing approach of natural language retrieval has been used to through analysis of user queries and matching the index. It adopted the route of lexical, syntactic and semantic processing, while the semantic processing also usually use traditional approach. Although it has been realized that natural language retrieval is in essence concept retrieval, which needs concept-oriented controlling on processing and concept reasoning based on KB. However, in practice, attempt in this area is still rare.

It is an important approach for natural language retrieval on semantic level to complete concept control with the help of ontology and reasoning on it. Ontology provides explicit interpretation for concept and describes the relationship between different concepts in the domain, which can provide knowledge base for the retrieval. Based on tools that support ontology reasoning, this research proposes to process users' natural language queries and convert them into language form that can be reasoned on ontology. For implementations, we will convert user's natural language queries into nRQL, which can reason and retrieve on ontology. The key work is how to establish an interface between natural language queries sentence and nRQL.

nRQL (new Racer Query Language) is an ABox query language for Racer system. Racer（Renamed ABox and Concept Expression Reasoner） is an expressive reasoned based on description logic. It is first developed by Ralf Möller from University of Hamburg and Volker Haarslev from Concordiia Univ. in 1999. Racer is a Semantic Web reasoning and knowledge representation system, which is based on the description logic SHIQ. Racer is also a mixture of description logic and specific relational algebra. Not only can it retrieve the relation of ABox assertion, but also it can retrieve special and temporal relations in the restriction related reasoning network.

The research of converting natural language into specific formal language started in the late seventies of the 20th century. The main work focused on converting user query into SQL statements and completing retrieval the structured data on relational database. In the domestic, a number of experimental system were developed, including SPS&ZPS by Tsinghua University, E-RVT by Huazhong University of Science and Technology and LIGC by Shanghai industrial university. Related query language includes: Chinese database query language chiqu(developed by Chinese University of HongKong and Renmin University of China) , CQI(written by professor

LiuDaxin from Harbin Unstitute of Technology), etc. However, research on establishing interface between natural language queries and ontology query language is still rare. Abroad, Abraham Bernstein studied retrieval based on limited natural language. This research based on limited natural language and introduced middle logic layer for representation. On realization, it was similar with most natural language queries interface to database. The system didn't consider the semantic mapping problem between keywords and ontology entities. When users changed their expressions, it performed poorly. There is another similar system named AquaLog. This portable system can accept natural language queries and make use of user's pre-defined ontology to give answers. But the processing ability of this system is not satisfactory. The main reason is that it has a strict requirement for processing and lots of sentences are beyond that.

This paper targets on natural language query and explores a appropriate way to complete the conversion from users' natural language queries to ontology query language nRQL. After retrieve on ontology with nRQL, the related knowledge and reasoning tools provided by ontology can guarantee concept control on semantic level, which assurance the effectiveness and improve the recall and precision ratio. As for experiment, we have chosen the field of library knowledge services and designed a series of question regulations based on limited natural language. Based on searching for the query target and query condition, this research processes some of grammatical phenomenon in Chinese through semantic understanding and completes the conversion from query to nRQL instructions successfully. Figure 1 shows the background, while emphasis is on the mapping to nRQL.

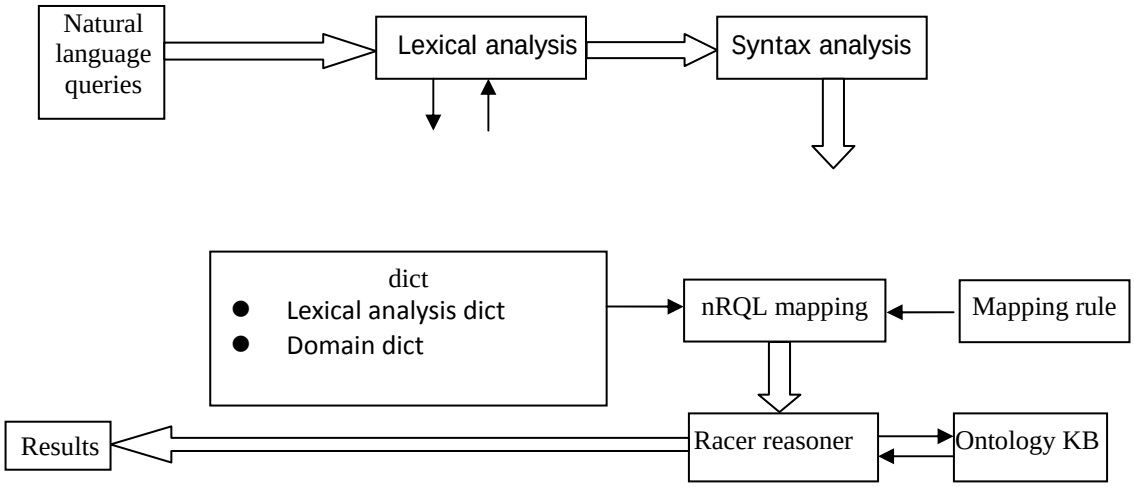


Figure 1. The retrieval process of natural language queries with nRQL

2 Grammar and Query of nRQL

2.1 Features and Basic Structure of nRQL

As an expressive ABox query language for Racer-Pro, nRQL can be used to inquiry KB(knowledge base) based on description logic, as well as RDF/OWL files. Similar to database query statement, it can be used as interactive command, as well as string form embedded into client applications, which can submit query request and deal with the result.

The features of the nRQL language can be summarized as follows:

1) nRQL has a well-defined syntax and semantics.

2) nRQL offers a variety of query atoms, including concept query atoms, role query atoms, constraint query atoms, and SAME-AS query atoms.

3) Special support for querying the concrete domain part of an ABox.

4) So-called complex TBox queries are available.

5) nRQL is also a powerful OWL and RDF(S) query languages.

6) The language offers extensibility and flexibility by means of a simple expression language called MiniLisp.

The top level syntax of nRQL is defined as follows: retrieve <query head> < query body>

The body of a query specifies the retrieval conditions, and the head specifies the format of the query result/ answer tuples. The head may also contain individuals. However, the set of objects mentioned in the query head must be a subset of the set of objects used in the query body. An empty query head is permitted as well; this will result in a boolean query (which only returns TRUE or FALSE).

The basic expressions of the nRQL languages are so called query atoms. Complex query body can be constructed with the help of query body constructors.

2.2 Query Atoms

Atoms are either unary or binary. A unary atom references one object, and binary atom references two objects.

An object is either an individupal, or a variable. In

978-1-935068-76-1 © 2011 SciRes.

case of an ABox atom, this object is thus either an ABox individual or a ABox variable. For example, ?x and $?x are variables, and betty is an individual. Variables are bound to those ABox individuals that satisfy the query expression.

1) Concept Query Atoms

Concept query atoms are unary atoms. A concept query atom is used to retrieve the instances of a concept or an OWL(or RDF(S)) class, for example, all of the instances belonging to the concept woman. Now, we retrieve the instance of the class woman using a query whose body consists of a single concept query atom, (?x woman):

? (retrieve (?x) (?x woman))
> (((?x betty)) ((?x eve)) ((?x doris)) ((?x alice)))

2) Role Query Atoms

Role query atoms are binary atoms. Role query atoms are used to retrieve role fillers from an ABox, or OWL/RDF(S) individuals which are related via a certain OWL object property.

For example, we can retrieve all individuals in the ABox using the following query, whose body consists of a single role query atom (?x ?y has-child):

? (retrieve (?x ?y) (?x ?y has-child))

Results are all individual pairs in the ABox with the role has-child。What is more, individuals can be used at any position where a variable is accepted − to retrieve the children of betty:

? (retrieve (?child-of-betty) (betty ?child-of-betty has-child))

3) Constraint Query Atoms

Constraint query are binary atoms and address the concrete domain part of an ABox. Like role query atoms they are used to retrieve pairs of individuals which are in a certain relationship to one another. However, this relationship is not a role, but specified with a binary concrete domain predicate, like =. For example, we can retrieve those pairs of individuals that have the same age. Age is a so-called concrete domain attribute of type cardinal.

? (retrieve (?x ?y) (?x ?y (constraint age age <)))

2.3 Operators

1) Query Head Projection Operators

We can use so-called head projection operators in the head of a query. A head projection operator is simply a function which is applied to the current binding of the variable, and the operator result is included in the answer tuple.

Head projection operators can be used to retrieve concrete domain values, like age attribute of individuals in the ABox:

? (retrieve (?x (told-value-if-exists (age ?x))) (?x (>= age 18)))

In RacerPro, OWL KBs may contain datatype fillers of datatype properities. With the help of the datatype-fillers head projection operator it is possible to retrieve these data fillers.

Syntax as follows:

? (retrieve
(?x
(datatype-fillers (#!name ?x)))
(?x (and #!person (> #!age 30))))

2) Query Body Constructors

There are four kinds of query body constructors:

a) and : an n-ary constructor, which is used for the formulation of conjunctive queries.

b) union : an n-ary constructor, which computes the union of the query answers of its disjuncts.

c) neg : a unary constructor, the negation as failure (NAF) negation.

d) project-to : a unary constructor. This projection operator is needed in combination with neg. The NAF-negation operator neg is designed to compute the n-dimensional complement set of its n-dimensional input argument set; while project-to operator can reduce the dimensionality of the argument.

For example, suppose we are looking for the woman which do not have a male child:

? (retrieve(?x)
(neg (project-to (?x) (and (?x woman) (?x ?y has-child) (?y man)))))

More information about nRQL's syntax, refer to nRQL users' guide[1]。

3 Convert from Natural Language Queriess to nRQL

3.1 The Basic Approach

Currently, the approaches for database natural language interface can be summarized as follows: one way is to create a so-called Intermediate language as middle logical layer, which is a kind of class relational algebra expression; another is sentence pattern matching, which will divide user's natural language queries sentences into several classes and describe them with a specific form -- such as syntax tree[2]. For each kind of pattern, it gives the corresponding way for converting. Besides, there is an approach called lexical-syntactic analysis. This method will identify database-related words (such as entity, property, property value) in lexical analysis stage, extract query target and condition, and complete the conversion to corresponding database instructions. Ma Xiaona, Yang Chenglei divide the content in database query sentence into three parts: condition, target and noise, and design an algorithm which obtain the target and condition field by means of searching keywords[3]. Xu Haiyun, Qian Bin accomplish similar effects through the design of data dictionary, which stores the information of syntax label, table, etc.[4].

Studies have shown that when users express their que-

ries with natural language, they tends to take the model of specific domain as basis[5]. Therefore, this research also takes the most general approach for converting natural language queries into SQL statement[6] – syntax-semantics analysis. The method does not consider the POS(Part-of-Speech), but targets on obtain the query object and target directly. This method is consistent with the syntax of nRQL, making it a effective and feasible approach for converting.

For ease of system processing, we refer to the traditional approach[7] and finally divide the words in users' queries into the following parts of speech: C—Class, I—individual, V—property value, P—property name, R—relation between individuals, O-- Relationship operator (Such as equals, larger than, less than), Z—auxiliary, L--Logical operators (such as and, or, and etc.),W-- Question word, N—numerals, Q—quantifiers, B-- Quantitative restrictions (such as above, at least), A—unknown.

3.2 User Query Sentence

Studies[8] have shown that six sentence patterns are used most frequently when users query the database with natural language , they are: who-questions, questions started with other question word, questions with question word in mid of sentence, imperative sentences started with query verb, sentences started with "Ba" and sentences ended with question words. At the same time, Statistics show that most questions are in form of imperative or special questions and imperative is used most frequently sentence patterns. Because "Ba" sentences can be transferred into imperative sentences started with query verb easily, we don't take this pattern into consider in this research. For imperative sentences started with query verb, system grammar are defined as follows:

<S query>::=<query verb><query condition section> [Z]<query target section>

<query verb>::=find out|query|inquire|retrieve|count|add up……

<query condition section>::=<query condition>{<L><query condition>}

<query condition>::=IZR|POV|PONQ|RNQBC|RC|BNQR|C

<query target section>::=<query target>{[L]<query target>}

<query target>::=P|C

For question sentences, we mainly consider about the situations when question words appear at the start or end of the sentences. Sometimes two words make a question phrase, which is also in our consideration—like "how old", "which year", "Which annual" ,etc. We will not consider the question sentences that question words appear in the middle, and define the grammar as follows:

<S question>::=< query target section >< conj >< query condition section >|< query condition section >< conj >< query target section>

< conj >::=unknown|Z

<query condition section>::=< query condition >{<L>< query condition >}

<query condition>::=IZR|POV|PONQ|RNQBC|RC|BNQR|C

< query target section >::=<pre-word><question target><suffix-word>

< query target>::= W|WQ|WP|WC

3.3 Query Target Identification

According to the grammar structure mentioned above, if there's only leading word before the question word of the sentence, it's better to use positive search mode; for the others, query targets would be at the sentence tail, it's better to use reverse search mode to search query goals.

Positive matching algorithm:

Scan the user query sentence from the first word; According to the maximal matching principle, put the first matched word into the target array, and delete it and its front string.

Reverse matching algorithm:

Step1 : Delete the leading word before the question word.

Step2 : Scan the user query sentence from the last word and match reversely.

Step3 : If matched, put the string into the query target array and modify the string variable, then truncate the matched word and the behind string; Otherwise judge whether the query target array is empty, and if so, show "no query target!"

3.4 Query Condition Identification Algorithm

Step1: First, delete the useless words and unknown words and record the unknown words for backstage manual monitoring and the building of the learning mechanism, then expand the dictionary of the system and enhance its processing capability at any time;

Step2: For the extraction algorithm of the condition phrase, match the remaining string variable reversely by the maximum matching method, and extract the condition phrase, then delete the matched and the behind string ;

Step3: Judge whether the first word of the modified sentence is the logical operator, comma or comas, and if so, to determine whether it means the relationship of "or" , if it does, mark it;

Step4: Judge whether the remaining string contains R, C, P, I and other highly correlated with the domain knowledge base words, if there is, jump to step1, otherwise the algorithm ends.

4 Conversion Processing Rules

The following are rules used when query is transformed. In the conversion, it will use two list structures to store the query head string and query body string respectively.

4.1 Conversion of the Query Verb

978-1-935068-76-1 © 2011 SciRes.

Omitted directly [9]

4.2 Conversion of the Query Target

1) target phrase=C/WC

target phrase=C, which Corresponding to the concept query in nRQL, that is, to search all instances of a class, add string" (? x)" into the query head list and string" (? xC)"into the query body list.

2) target phrase=P/WP

target phrase=P, which corresponding to the datatype value in Search OWL ontology, add string"(?x (:datatype-fillers(P ?x)))" into the query head list.

3) target phrase=W

target phrase=W, usually the query target conceals in the target phrase, and question word implies the concept of the query obeject like "who", "where", and so on.

Firstly add the string"(?x)" into the query head list, then search the corresponding entry in association dictionary. If the corresponding "C" is found, add "(?x C)" into query body list.

In another case, a question word may correspond to many properties such as "duo da" which may asks age(how old) or house area(how big). At this time all the pairs of class-property which the queries may asked should be stored in the pending query head list. Meanwhile, search C in the query condition, if not found, then search the C corresponded to I and match the query head list which is just modified, finally find P corresponded to the common C.

4) target phrase=WQ

For example, "how old" is asking for age, "how many meters" is asking for length .

4.3 Conversion of the Query Condition

1) No query condition

Scan the current string, and count the "("and ")" as x and y; According to nRQL grammar rules, add n(n=x-y) ")" into the end of the string to complete the back parentheses of the command.

2) condition phrase=/IR/IZR/CIR/RI

The first two modes are on behalf of that there is R relation between the query target and Instance I which turns to"(I ?x R)". In the third mode, C then I is on behalf of the modified relationship between C and I, and then add the string "(IC)". The last mode means the exchange of subject and predicate which turn to "(?x I R)" .

3) condition phrase=RC

According to nRQL grammar, add "(?x ?y R)"and"(?y C)" into query body list.

4) condition phrase =POV/POVZC/PONQ

Corresponding to nRQL, to limit the Datatype property value of the query target, it needs to determine the type of Datatype properties. In general, the common types are character and numeric. For string type, nRQL can only handle the relation of equal and not equal. While for numeric type, it can also handle the relation of more than, less than, more than or equal to and less than or equal to.

Condition phrase=POV means string type, according to nRQL grammar, it turns into the command string"(?x (string= P "V"))"

Condition phrase=POVZC, based on condition phrase=POV, add string "(?x C)" into query body list.

Condition phrase=PONQ means numeric type, according to nRQL grammar, it turns into the command string"(?x(O P N))")

5) condition phrase =RNQBC/ BNQR/RONQ

The first two modes, NQB/BNQ are the number of the relation R between the query target and class C. According to nRQL grammar, add string (?x (B N R C))/ (?x (B N R)) into query body list.

The last mode, ONQ is the number of the relation R formed by modifying the query target. O should use the second formal description stored in the dictionary, then add the string "(?x (O N+1 R))" into query body list.

6) condition phrase =I(Z)

It needs to judge the query head field to find the query target is C or P. If P, add SAME-AS query atoms into query body list, that is, add string "(same-as ?x I)"; If C, it requires further search of related vocabulary to find class C1 corresponded to I and the relation R between C and C1, then add string "(I ?x R)".

7) condition phrase =NQ

There is neither P nor R in condition phrase, so it cannot determine the object that NQ constraints. It needs further search of the related vocabulary of Q to find that whether the most possible P could be found. In addition, this expression doesn't contain comparison operators, so it most possibly express the equal relation which turn to command string "(?x(= P N))"

8) condition phrase =OI

As it doesn't describe the compared contents of the comparison operators, it needs to search the contents in the query target list. Generally it should be C, then search the knowledge base to find the intersection of the comparable properties P of C and O. Depending on the semantic, compare P between query target and instance and turn to command string "(?x I (constraint P P O))".

5 Query Processing Experiment

5.1 Ontology Knowledge Base

For the conversion experiment, we established an experimental body for query analysis and conversion validation. The figure below is an ontology class hierarchy graph drawn by the plug-in OWLviz of Protégé.

5.2 Query Conversion Instances

Example 1: "Shui shi Jia Zheng de er zi?" (Who is Jia Zheng's son?)

Lexical analysis: "Shui(who)" W| "shi(is)" A| "Jia Zheng" I| "de" Z| "er zi(son)" R|

Belongs to who-sentence, query target is W, "shui(who)"

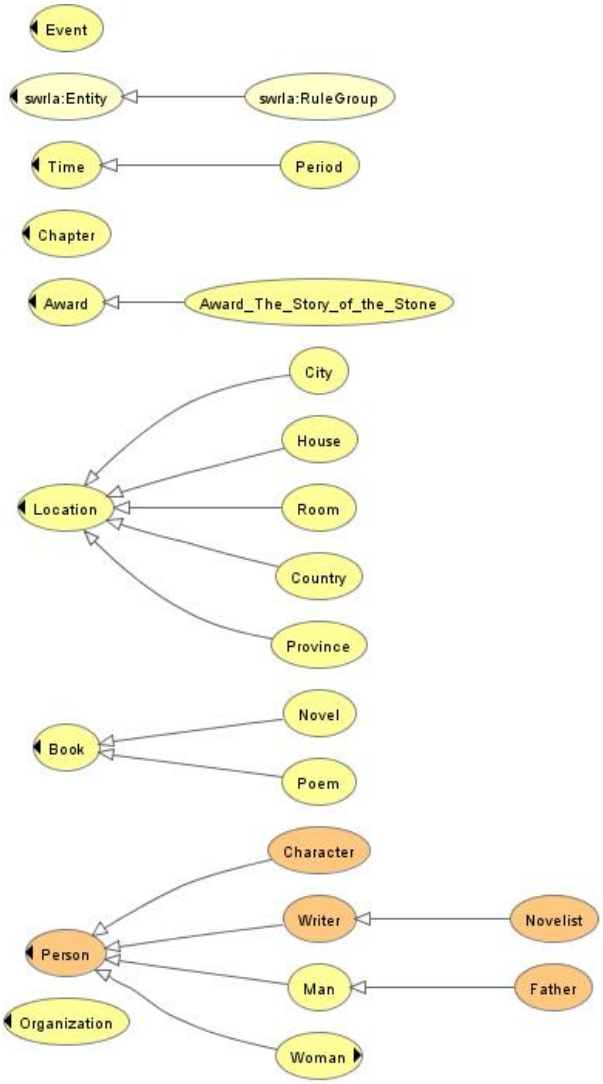


Figure 2. Class structure of ontology used in query sentence conversion experiment

Match query condition reversely, and the maximum match is IZR

In the field dictionary, the formal description of "Jia Zheng" and "er zi(son)" are #!:Jia_Zheng and #!:has_son

In accordance with the transformation algorithm, turn to nRQL command

Search result is as follows:

> (((?x #!:Jia_Baoyu)))

Example 2: "Na xie shu de feng pi shi lv se de?" (The cover of which books is green?)

Lexical analysis: "na xie(which)" W| "shu(book)" C| "de" Z| "feng pi(cover)" P| "shi(is)" O| "lv se(green)" V| "de" Z

As the question word is in the first, search the query target forward.

The query target is the first word "shu(book)" after the first question word, and the part of speech is C.

Delete the string "na xie shu(which books)", then search the query condition reversely, and the maximum match is POV.

In the field dictionary, the formal description of "shu(book)", "feng pi(cover)", "lv se(green)" are #!:Book 、 #!:cover_color and green. While, "feng pi(cover)" is stored in the thesaurus as the synonym of "feng pi yan se(color of the cover)".

In accordance with the transformation algorithm, turn to nRQL command

(retrieve (?x) (and (?x #!:Book) (?x (string= #!:cover_color "green"))))

Example3: "Cao Xueqin xie guo na xie shu?" (What books did Cao Xueqin write?)

Lexical analysis: "Cao Xueqin" I| "xie guo(wrote)" R| "na xie(what)" W| "shu(books)" C

As the question word is in the middle, search the query target forward.

The query target is the first word "shu(book)" after the first question word, and the part of speech is C.

Delete the string "na xie shu(which books)", then search the query condition reversely, and the maximum match is IR.

In the field dictionary, the formal description of "Cao Xueqin", "write(or wrote)", "shu(book)" are #!:Cao_xueqin、#!:write_book and #!:Book.

In accordance with the transformation algorithm, turn to nRQL command

(retrieve (?x) (#!:Cao_xueqin ?x #!:write_book))

Example4: "Cha xun nian ling deng yu 17 sui de nv hai" (Search the girls whose age are 17)

Lexical analysis: "Cha Xun(Search)"| "nian ling(age)" P| "Deng yu(is)" O| 17 N| "sui" Q| "de" Z| "nv hai(girls)" C

For the imperative sentence begin with the query verb, search the query target reversely. The query target is "nv hai(girls)" ,and the part of speech of is C. In the thesaurus, this word points to the synonym "nv ren(woman)".

Search the query condition reversely, and the maximum match is PONQ.

In the field dictionary, the formal description of "nian ling(age)", "deng yu(is)", "nv ren(woman)" are #!:age、=、#!:woman.

nRQL commands : (retrieve (?x) (and (?x #!:woman) (?x (= #!:age 17))))

Example5: "Huo guo liang ci yi shang hong lou meng jiang de zuo pin" (works awarded to "Dream of the Red Chamber" more than twice)

Lexical analysis: "huo guo(awarded)" R| "liang (two)" N| "ci(times)" Q| "yi shang(more than)" B| "Dream of the Red Chamber" C| "de" Z| "zuo ping(works)" C

Search the query target reversely. The query target is "zuo ping(works)", and the part of speech is C. Delete "zuo ping(works)", then the part of speech of the last word of the string is Z, so end the searching.

Search the query condition reversely, and the maximum match is RNQBC.

978-1-935068-76-1 © 2011 SciRes.

In the field dictionary, the formal description of “huo de(awarded)”, “yi shang(more than)”, “Dream of the Red Chamber” are #!:get_award, at-least, #!: Award_The_Story_of_the_Stone, #!:Book. While, “yi shang(more than)” is the synonym of “zhi shao(at least)”.

Turn to nRQL command：

(retrieve (?x) (and (?x #!:Book)(?x (at-least 2 #!:get_award #!: Award_The_Story_of_the_Stone))))

References

[1] RacerPro User’s Guide Version 1.9.2.
[2] Ma Xiaona and Yang Chenglei, Design and Implementation of a Limited Nature Language Query System Based on Object-oriented, Computer Engineering and Applications, 2005, no. 10, pp165-168.
[3] Ma Xiaona and Yang Chenglei, Syntax Analysis and Arithmetic Realization of Natural Language Query Systems, Journal of Shandong Institute of Architecture and Engineering,2005,vol. 1, no. 20, pp76-81.
[4] Xu Haiyun and Qian Bin, Design and Implementation of Natural Language Interface Software, Computer & Information Technology, 2007,Z1, pp74-75,79.
[5] John F. Burger. CONCEPTUAL PROCESSING/NATURAL LANGUAGE INTERFACE.
[6] Xu Haiyun and Qian Bin, Design and Implementation of Natural Language Interface Software, Computer & Information Technology, 2007,Z1, pp74-75,79.
[7] Research on Database Query Technology via Natural Chinese.
[8] Xu Jiuyun, Wang Xinmin and Tong Zhaoqi, Grammar Identification and Structural Analysis of Query Language of Chinese Database, Journal of Petroleum University (Natural Science),Vol. 21,No. 2, 1997.
[9] Xu Longfei, Tang Xiaodi and Tang Shiwei. Research on Natural Language Interface Technology Based on Limited Chinese Database, Journal of Software, 2002, vol. 13, no. 4, pp537-543.

Research on Sentence Level Bibliometrics—Analysis about Abstract of JASIST

Bolin HUA, Yanning ZHENG

Institute of Scientific and Technical Information of China, Beijing, China

Email: huabolin@istic.ac.cn

Abstract: During the past period of time, quantitative analysis of journal articles focused on the author, keyword, subject, body, references and other fields. These fields can't reflect the content of literatures, only title, abstract and body rather can reflect the content of the literatures. Thus title is relatively brief, which include a little information. The body of literature can best reflect the content, but because of large quantities of work to get full-text, more complex analysis, analysis on the summary is a strong feasibility. This research selected 1452 papers abstract published on JASIST 2010- 2001, measured from the sentence-level analysis. This study include quantitative relations among abstracts, sentences and words, analysis on the structure of abstract, analysis on sentence structure. This paper focuses on the data appearing in the summary, on the description of the experiment, on the comparative analysis, and research on methods of quantitative analysis. Through research, we get some descriptive rules as well as some quantitative distribution.

Keywords: content analysis; sentence level; JASIST; knowledge extraction

1 Introduction

1.1 Relative Research

Quantitative analysis of journal articles focused on the author, keyword, subject, body, references and other fields. These fields can't reflect the content of literatures, only title, abstract and body rather can reflect the content of the literatures. Thus title is relatively brief, which include a little information. The body of literature can best reflect the content, but because of large quantities of work to get full-text, more complex analysis, analysis on the summary is a strong feasibility.

Book[1] systematically describes the content analysis of bibliometric methods, and select a certain amount of experimental data, but the experiment is only conducted quantitative analysis of the information on the author, keywords. It lacks deep analysis on the papers content, especially lacks sentence level analysis. Article[2] use a method combining symmetric matrix with asymmetric matrix to detect new sentence in the text. Book[3] build three- dimension and corresponding analysis unit. Words form text, named article of journal; They study relation between words and text, but ignore relations between sentences and text.

1.2 Value of Sentence Level Bibliometric

Sentence of the article is an important component, and also the smallest unit, by which author can show their viewpoint. Generally speaking, word can't give a complete description of knowledge. The smallest unit of describing knowledge completely is sentence. Series of analysis on sentence should be focus of literature processing, in particular knowledge extracting from the literature. Sentence matching and analyzing origins from the machine translation, and has been developing rapidly in the automatic summary as well as questions and answer. Analysis of the sentence has a very good foundation, but research of bibliometrics or content analysis from sentence level is rare. Taking sentence as a unit to analyze paper content has a strong feasibility and superiority.

It is a project supported by National Natural Science Foundation of China (70803048)

1.3 Data Source

This research selected papers abstract published on JASIST 2010- 2001, measured from the sentence-level analysis. Data of this research comes from Web of Science. 1897 papers were retrieved through limiting journal name and publishing year. After filtering out 445 papers without abstracts, we get 1452 effective papers Abstract.

2 Methodology

2.1 Data Preprocess

First, we download retrieved result and convert the data into two-dimensional table format by format conversion. Then filter out the paper information without abstract, cut abstract into sentences for subsequent processing. Cut into the summary sentence. Segmentation of sentence is punctuation mark to prevail, including the period, the question mark, etc. In this process, we also need to deal with the period when then exist names, place names, organization names, such as U. S. A., Inc. V. Tel., R.H. Atkin.

2.2 Method and Procedure

First, we construct rules for extracting knowledge, turn rules into regexp. After extraction system extracting knowledge using regexp, results are stored into excel. Sentences meeting conditions and sentences not meeting

978-1-935068-76-1

conditions are stored different sheets. Then we view the results, see whether the result sheet hit the wrong data, and see whether the filtering results have not hit the missing data. Then modify rules based on results, and then re-run. The whole procedure is as follows Fig. 1

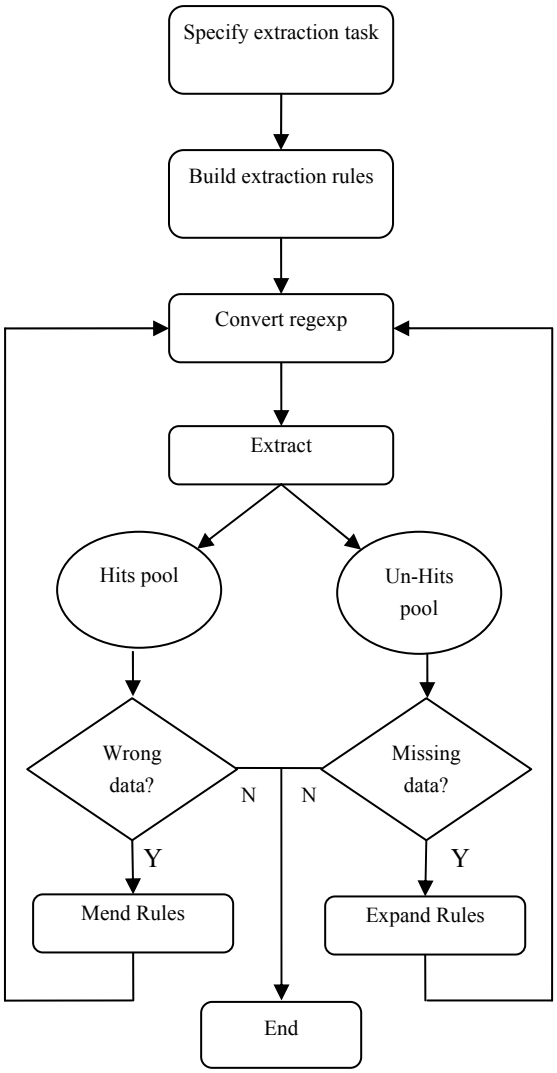


Figure 1. Procedure of Extraction

3 Results

3.1 Form Analysis

3.1.1 Numerical Relationship among Abstract, Sentences and Words

Among 1452 abstracts, the largest amount of sentences in an abstract 21 sentences, and the shortest abstract only has one sentence. There are 10,104 sentences isolated from 1452 abstracts, with an average of 7.0 sentences per article. The average length of each sentence was 24.8 words. An abstract has 164 words on average. The paper which has maximum number of words is "A system for supporting evidence recording in bibliographic records" published in No. 6 in 2009 of JASIST, the word number amount to 445. The paper which has minimum number of words is "On Egghe's construction of Lorenz curves" published in No.10 in 2007 of JASIST, only 18 words. Sentence count among these abstracts is from the paper published on No. 9 in 2006 titled "Documents and queries as random variables: History and implications", identified a total of 21 sentences. Numerical Relationship among Abstract, Sentences and Words is as Tab. 1.

Table 1. Numerical Relationship among Abstract, Sentences and Words

	Sentences	words
Maximum	21	445
Minimum	1	18
Average	7.0	164
Sum	10104	238242

There are 229 articles which abstract include 6 sentences, and 225 articles which abstract include 7 sentences, respectively, 16.1%, 15.7%. Amount of abstract which includes 6 or 7 sentences, 31.9%; Amount of abstract which includes 4-10 sentences amount to 1209, accounting for Summary of the total 84.8%. So, the majority of summary contains 4 to 10 sentences. Relation between sentences and articles is as Fig. 2.

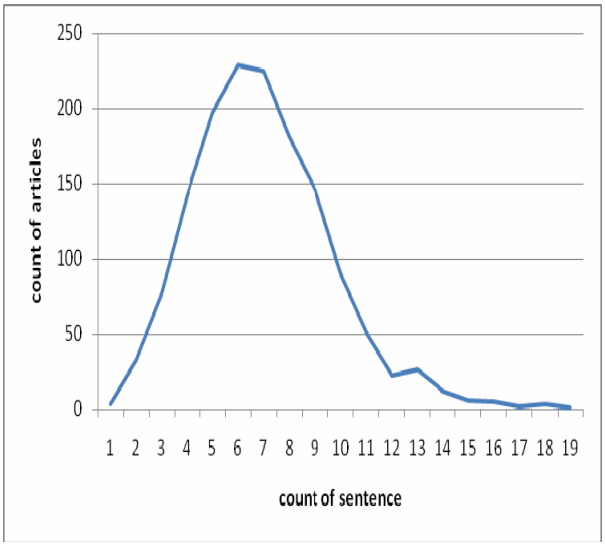


Figure 2. Relations between Sentences and Articles

As can be seen from the above graph, count of abstracts which include different number of sentences meet the normal distribution.

3.1.2 Word Frequency Analysis

Word frequency analysis need get rid of punctuation marks, reserved -, without dealing with singular and plural, as well as ing, ed, etc., which word still retains the

original form. If the complex process is taken to normalize to the single form, and ing, ed treatment normalized the words, then the frequency of content words will be higher.

There are 13217 words in 1452 abstracts, which sum up 238138 times, one word occurs 18 times on average. The most frequently words are "the", "of", "and", "to", "in", "a", and so on. Generally speaking, these words can't reflect subject character of document, which works on conjunctions. Words count more than 1000 times is in Tab.2.

Table 2. Words with high frequency

Word	times
The	15609
of	11141
and	8204
to	5665
in	5333
a	5274
that	2904
is	2857
for	2789
information	2268
this	2173
on	1953
are	1827
as	1666
with	1512
by	1415
we	1411
an	1190
from	1189
research	1038

As can be seen from this table, In the top twenty word lists by frequency only appear two content words which are "information" and "research", and only "information" can reflect professional nature, which appears 2234 times, ranked No. 10.

3.2 Structure Analysis

3.2.1 Analysis on Structure of Abstract

Abstract usually are consisted of introduction, methods, results and conclusions. Structure of abstract is the same to the structure of the body, just more condensed. Introduction in the abstract include several types, as describing purpose, introducing background, presenting problem.

Introduction of describing purpose introduction goals or objectives of this research, common characteristics is as following expression : to…, aim to…, objective…, attempts to solve…, In order to evaluate/testify/ the proposed algorithm.

Introduction of background type (environmental changes) is mainly described as the emergence of some new things, or certain things grow rapid, for example, with the rapid development of web 2.0…, The rapid advance of …, With the emergence of …., The rapid growth of …, … have become popular …, …have recently been adopted by…, …become a new trend in ….

Introduction of presenting problem mainly introduce the problems of current research over a certain period of time, paving the way to conduct this study. For example, during the past…, …Suffer two problems, Lacks …. Analysis, There have been few studies on…, Traditional approaches…, … has a complex structure, which makes it difficult to apply…, Address the problem…

3.2.2 Analysis on Sentence Structure

5518 sentences containing copula are found among 10,104 sentences, there are 950 sentences containing "can be / to be". After a series of processing, 1548 sentences characteristic of "be+adj" are extracted, accounting for 15.3% of the total number of all sentences. So ratio of sentences characteristic of "be+adj" accounting for total sentences in the academic literature is very low. Most sentences use the verb-object structure. Sentences characteristic of "be+adj" are the following types:

… Is essential/ important/ necessary

We are interested in …

… is a difficult problem, despite many efforts from …

3.3 Content Analysis

3.3.1 Analysis on Digital

1389 sentences containing digital are extracted among 10104 sentences, which distributed in 744 articles.

Digital in abstracts can be divided into three categories, including experimental data count, annual data, and experimental results data. Experimental data including amount of experimental objects, amount of questionnaire; annual data includes a point of time or period of time, 395 sentence appeared year are extracted; experimental result data usually present results indicators, such as recall rate, precision, in the form of a percentage. 216 sentences occurred percentage digital are extracted. In addition to these three types of valid digital, there is a class of noisy data to identify characteristics of the data in a term, such as web 2.0, F-1, 2-dimensional plots, TREC-6.

Table 3. Different Digital Count

	Count	Example
Data about experimental objects		Bringing their own 60 search tasks, 31 participants, A randomly selected and representative sample of 246 active research scientists worldwide was surveyed
Data about year	395	The citation image of Conrad Hal Waddington, a developmental

978-1-935068-76-1 © 2011 SciRes.

		biologist and evolutionary theorist, is mapped using both cocitation and tricitation approaches over three successive decades, 1975-1984, 1985-1994, and 1995-2004
Data about experimental results	216	Our methods suggest that our block-weighting ranking method is superior to all baselines across all collections we used and that average gain in precision figures from 5 to 20% are generated.

3.3.2 Analysis on Experiment

Experiments in Abstract are classified with Experimental study and empirical research. Experimental study is characteristic of a strong technique. Empirical studies often used in humanities and social sciences. Experimental study often aims to testify certain problem, select certain data to make an experiment, following the corresponding experimental results. Some experiments aim to verify the validity of method, some experiments aim to show the advantages of method by comparing corresponding experimental results. The latter need common suite sets, such as evaluation suite sets in TREC conference

Empirical researches include case studies, field research, and study by analyzing questionnaire.

184 articles referred to experiments in abstract are extracted, only account for 12.9%, using 215 sentences. So amount of articles mentioning experiment in abstract of the JASIST is not so many. Description about experiments include experiments data, process method, experiment tools and experiment result. Mentioned in the previous bibliometric analysis, 216 sentences involved in experiment results are extracted from 1452 abstracts, it exceeds the amount of explicit "experiment" in abstract because there are a lot of descriptions on the results of survey results and other evidence in abstracts.

An experiment was conducted with…

We examined/observed the…

3.3.3 Analysis on Comparison

1390 sentences involved with comparison from 831 abstract are extracted, accounting for 57.2%. Comparative analysis includes two types, one is comparison between articles, contrast with previous research, aiming to elicit innovative research; another is comparison inside article, such as comparison between different countries. Some discussion style on comparative analysis in abstract are as follows: Many studies…Yet/but/however/ regardless of…, these studies…, In this study/here we/this article ,…. , Early work focus on…

3.3.4 Analysis on Method

Description of the method is usually innovation of the paper, so Extraction method from article is quite significant. Extraction methods include method name, advantages and disadvantages (in particular, as opposed to more than the previous method) of method. Discussion of research methods are classified with the proposed type and improved general categories. 667 abstracts with proposing method are extracted, and 168 abstracts with improving method are extracted. 3058 sentences containing both verbs like "present" and nouns like "method" are extracted. 7046 sentences containing verbs like "present" but nouns like "method" are extracted. 2038 sentences containing nouns like "method" but verbs like "present" are extracted. So, verbs like "present" is preferred using than verbs like "modify" in abstract. The count of these words is as Tab. 4

Table 4. Count of Especial verbs and nouns

	Verbs like Present	Verbs unlike Present
Noun like method	3058	2038
Noun unlike method	7046	—

Verbs in abstract are often used like *Present, produce, employ, explore, describe, offer, set forth, discover, introduce, pose, develop, create, find, probe.* Nouns in abstract are often used like *method, approach, methodology, means, way, measure*.

In addition to verbs like "present", there are also some verbs like "modify", such as modify, revise, mend. There are some neutral descriptors, such as offer, consider. In addition to verbs class, there are some adjectives labeling innovative modifiers, such as the novel / new / different / flexible.

4 Conclusions and Discussion

Viewing from the above analysis, we can extract some useful information and knowledge, which are difficult to obtain through author analysis, theme analysis, and citation analysis. So it is interesting to conduct content analysis from article abstracts. However, due to the complexity of the language, there are also some following difficulties to analyze paper

(1) Tagging attributes of knowledge after parsing for each sentence is difficult.
(2) It is necessary to improve the completeness of the rules. Building a set of common effective rules is a complex project.
(3) Features of explicit description on method is relatively easy to obtain rather than features of implicit description on method, extracting and recognizing explicit method is rather easy. Description of method about method advantages and disadvantages, applicability of the method is difficult to extract from abstracts. These useful information can be found only through the full text analysis. Analysis on body can mine some more novel and valuable knowledge.

5 Acknowledgment

It is a project supported by National Natural Science

Foundation of China (70803048).

References

[1] Junping Qiu, Yuefen Wang. Bibliometrics and Content Analysis Method, National Library Press, 2008.10;

[2] Flora S. Tsai, Wenyin Tang, Kap Luk Chan, Evaluation of novelty metrics for sentence-level novelty mining Information Sciences: an International Journal archive Vol.180 Issue 12, June, 2010

[3] Loet Leydesdorff. The Challenge of Scientometrics. Wu yun et al.translate, Yishan Wu Proofreading, Scientific and Technological Literature Publishing House, 2003.

978-1-935068-76-1

A Study on Testing and Improving Nonlinear Evaluation Methods for Academic Journals

Liping YU[1], Yuntao PAN[2], Yishan WU[2]
[1]*Business school, Ningbo University*
[2]*Institute of Scientific& Technical Information of China,*
Email: chinayangzhou@yahoo.com.cn, panyt@istic.ac.cn, wuyishan@istic.ac.cn

Abstract: In this paper, indicator-based comprehensive evaluation methods for academic journals are divided into two groups: linear and nonlinear evaluation methods. In nonlinear evaluation, sometimes an abnormal phenomenon occurs where the final evaluation score decreases while the value of component indicators increases. Regression adjustment method, a new method for test and improvement, is suggested as a solution for the above abnormality. In regression adjustment method the first step is a multicollinearity test after a regression analysis, where evaluation score is used as dependent variable and evaluation indicators as independent variables; the second step is to judge whether multicollinearity occurs or not. In the case of no multicollinearity, one should give up those indicators with negative regression coefficients, and then renew the previous process if there is multicollinearity, however, one should adopt ridge regression approach for estimation and adjustment. The authors conclude that attention should be paid to hidden defects in nonlinear evaluation methods; prudence is necessary before one decides to give up certain indicators; excessive regression-adjustment circles is not recommended; there may be nonlinear evaluation methods free from the requirements for test and improvement; further study is expected on many issues in journal evaluation methods.

Keywords: academic journals; regression adjustment method; principal component analysis; topsis; grey correlation

1 Introduction

Academic journals are an important window displaying the national scientific and technological development level, and also a wide channel for bridging the supply of scientific and technological achievements with the demand of industry. Evaluation of academic journals is a significant component of the bibliometric study. It involves quantitative analysis of development regularity and growth trend of academic journals, reveals distributional law of publications in journals, and offers an important reference point for optimizing utilization of academic journals. At the same time, it helps to increase intrinsic quality of academic journals and to promote their healthy growth and development.

Studies on performance evaluation often focus on identification of research of the "highest quality", "top research" or "scientific excellence". This focus on top quality has led to the development of a whole series of bibliometric methodologies and indicators (van Leeuwen et al, 2003). Methods of journal evaluation usually include Single Index Evaluation and Multiple Attribute Evaluation (MAE). Representative indicators are the Relative Citation Rate (Schubert et al., 1983), the Relative Subfield Citedness (Vinkler, 1986), the Normalized Mean Citation Rate (Braun & Glänzel, 1990), the Field Citation Score (Moed et al., 1995), the Hirsch Index (Hirsch, 2005), the Article-Count Impact Factor (Markpin et al., 2008), etc.

As research performance is multidimensional it is clear that it cannot be evaluated by a single indicator (Martin, 1996). In MAE indicators are combined into a single index. Multiple attribute evaluation (MAE) could be divided into two types. The first is linear evaluation method, with its basic principle being to seek weighted sum of standardized data after giving weight to evaluation indicators in subjective or objective approaches, thus the relation between the evaluation result and the indicators' value is linear. Some examples are expert panel method, Delphi method, analytical hierarchy process method (AHP), entropy weight method and variation coefficient method, etc. The second type of MAE is nonlinear evaluation method. Here, for evaluation purposes it mainly uses methods often found in fields of operations research, fuzzy mathematics and system engineering etc, so the relation between the evaluation result and the indicators' value is nonlinear. Some systemic evaluation methods need no weighting, such as principal component analysis (PCA), data envelopment analysis (DEA) and TOPSIS etc; others may require weighting for evaluation, either by subjective or objective ways. Most evaluators adopted subjective weighting approaches, such as weighted TOPSIS, ELECTRE, fuzzy comprehensive evaluation and PROMETHEE etc.

Extensive research has been reported in this field. In the field of MAE, Weiping Yue and Concepcion S. Wil-

The National Social Science Foundation of China (Grant No. 10FTQ003) and a grand Sponsored by the National Natural Science Foundation of China (Grant No. 70973118).

son (2004) established an analysis framework of journal influence based on the structure equation model. Qiu Junping and Zhang Rong (2004) used grey correlation method for evaluation. Pang Jingan and Zhang Yuhua (2000) and Li Kaiyang and Jia Yuping (2005）took AHP method in their evaluation. Wang Xiaowei and Yang Bo (2003) used DEA. Cheng Hanzhong (2004) adopted PCA. It seems that various linear and nonlinear evaluation methods are used in academic journal evaluation, and many results have been achieved.

In using principal component analysis (PCA), Yu liping (2008) found a strange phenomenon, which is that the journal evaluation value may decrease even when the value of this or that positive indicator increases. To resolve the problem that sometimes a negative weight appears in the principal component analysis, Su Weihua (2000) proposed that when evaluation indicators are reasonable, one could use unconstrained principal component, or just give up this method. Hou Wen (2006) proposed to sectored PCA in order to make improvement for the case when some indicator coefficients turn out to be negative in the first principal component, which is against common sense in evaluation. In a word, present research is limited to pointing out problems in the principal component analysis, however, no good resolution has been proposed. Even fewer scholars have ever made attempt to consider such problems as generalized problems of nonlinear evaluation methods.

As far as evaluation method is concerned, every method has its own characteristics and application scope. For example, PCA and factor analysis are suitable for evaluation involving highly correlated indicators. DEA is suitable for evaluations with input-output indicators. Generally speaking, however, many evaluation methods (such as TOPSIS, entropy weight and grey correlation method etc.) have wide application scope, thus multiple evaluation methods may be applied to academic journal evaluation, so for the same evaluation object, evaluation results are not the same when different methods are adopted. This leads to two problems: the first is selection of evaluation method, the second is the test and improvement of the method involved. For linear evaluation methods, selection is mainly based on various evaluation principles; but for nonlinear evaluation methods, both the selection based on evaluation principles and the test as well as improvement are involved. This paper focuses on test and improvement of nonlinear evaluation methods.

Based on the medical journals evaluation data provided by the Institute of Scientific and Technical Information of China, a systematic method called Regression Adjustment is presented here to resolve the problems mentioned above. We take PCA, TOPSIS and grey correlation method as examples, first use them in evaluation exercises, then make regression analysis over evaluation results. Finally we identify problems and make corresponding judgments and improvements on evaluation methods.

2 Methods

2.1 Methods of Systematic Evaluation

2.1.1 Principal Components Analysis

The concept of PCA was first proposed by Karl Parson in 1901 and applied only to nonrandom variable. In 1933 Hotelling applied it to random variable. In most practical issues, existence of many indicators and the correlation among indicators will certainly increase complexity of analysis. By Substituting the original indicators with a new set of independent comprehensive indicators which are forms by recombination of the original indicators, the principal component analysis is capable of selecting a few indicators from the comprehensive indicators and these selected ones would reflect information of the original indicators as completely as possible. This method is usually used to seek such comprehensive indicators that could judge certain things or phenomena, rationally explain information included in the comprehensive indicators, thus revealing the inherent law of things more deeply.

The principal component analysis method shows similarities and groups the indicators. As an evaluation method, it is also widely used in the evaluation of Public Research Institutes (D. Krishna,*et al*., 2002), Environmental Degradation (Tahmina Khatun,*et al*., 2009), Redundant Air Quality (J.C.M. Pires,*et al*., 2009), and Banking Performance (Chien-Ta Bruce Ho, *et al*., 2009) etc.

2.1.2 TOPSIS

The acronym TOPSIS stands for technique for preference by similarity to the ideal solution. TOPSIS was initially proposed by Hwang and Yoon (1981), Lai *et al*. (1994), and Yoon and Hwang (1995). It was based on the concept that the selected best alternative should have the shortest distance from the ideal solution and the farthest distance from the negative-ideal solution in a geometrical (Euclidean) sense. That is to say, the ideal alternative has the best level for all attributes considered, whereas the negative ideal is the one with all the worst attributes value. So a TOPSIS solution is defined as the alternative that is simultaneously farthest from the negative-ideal and closest to the ideal alternative.

2.1.3 Grey Correlation Method

The grey theory was founded by Chinese scholar Prof. Julong Deng in 1982 and caused intense interest because of its original idea and wide applicability. Here “grey” means the information is incomplete, so it is neither white nor black. This theory tries to use very limited known information to forecast or tell unknown information. By far, it has already developed a set of technologies including system modeling, analysis, evaluation,

978-1-935068-76-1 © 2011 SciRes.

optimization, forecasting and decision-making, and has been applied in many fields.

2.2 Multiple Regression Analysis

To analyze relation between the nonlinear evaluation result and all evaluation indicators, common multiple regression analysis model is adopted with the following basic form

$$Y = a_1 X_1 + a_2 X_2 + \cdots\cdots + a_n X_n \qquad (1)$$

where Y is the evaluation value as the dependent variable, while $X_1 \ldots\ldots X_n$ are the evaluation indicators of academic journals as the independent variables, coefficients $a_1 \ldots\ldots a_n$ reflect the contribution extent of each indicator and are equivalent to weight in some degree. An indicator with positive coefficient increases the evaluation score as it increases. It is just the normal situation for positive indicators. However, positive indicators with negative coefficient cause problems. It is our purpose to use multiple regression to find such contradictions.

2.3 Multicollinearity Analysis

Multicollinearity is a statistical phenomenon in which two or more predictor variables in a multiple regression model are highly correlated. In such situations the coefficient estimates may change erratically in response to small changes in the model or the data. Multicollinearity does not reduce the predictive power of the model as a whole. However, it affects calculations regarding individual predictors. In other words, a multiple regression model with correlated predictors can indicate how well all the predictors as a whole predicts the outcome variable, but may fail to give valid results about any individual predictor, or about which predictors are redundant with others.

Damodar (2004) thought that multicollinearity causes such problems as larger R^2 and smaller t-test value, very large variance and standard deviation, sign error in regression coefficients, etc. In practice, tolerance and variance inflation factor (VIF) are often used to test multicollinearity.

Tolerance and VIF are two important indicators in multicollinearity test. Tolerance equals to 1 minus R^2 obtained by regressing analysis as shown in Formula (2), that is

$$\text{Tolerance}=1-R_j^2 \qquad (2)$$

Very large tolerance means very small R_j, which implies that X_j includes quite a bit of independent information and may become an important predictor variable; in comparison, very small tolerance plus very large R_j indicate that the more X_j duplicates the information included in other predictor variables, the weaker is its capability to explain dependent variable Y, thus multicollinearity happens easily. Threshold of tolerance is determined according to specific requirement. When tolerance is less than 0.1, it is usually considered that multicollinearity between variable X_j and other predictor variables has gone beyond the tolerance limit.

Variance inflation factor (VIF) is reciprocal of tolerance, that is

$$VIF = \frac{1}{Tolerance} = \frac{1}{1-R_j^2} \qquad (3)$$

VIF equals 10 when tolerance equals 0.1. So when $VIF > 10$ we know that multicollinearity occurs between X_j and other predictor variables.

2.4 Ridge Regression

Ridge regression is in fact a modification of least square method, a kind of biased estimation regression method specially designed for collinear data analysis. Ridge regression analysis, described by Hoerl and Kennard (1970), is a technique for removing the effect of correlations from the regression analysis. The regression coefficients obtained are biased but have smaller sums of squared deviations between the coefficients and their estimates.

The classical regression is to make the following calculation:

$$\hat{\beta} = (X'X)^{-1}X'y \qquad (4)$$

When multicollinearity occurs in data, matrix X'X is almost singular and gives rise to very high variance of final estimation result. This problem can be avoided by transforming matrix eigenvalue as follows:

$$\hat{\beta} = (X'X + kI)^{-1}X'y \qquad (5)$$

Thus the variance of estimation is reduced, but the estimation is biased. If reduced variance is much higher than increased estimation bias, which means that MSE（Mean Square Error）is decreased, then the new method is considered valid.

Ridge regression can be seen as penalty regression, the minimum

$$\sum_i (y_i - b_0 - \sum_j x_{ij} b_j)^2 \qquad (6)$$

can be substituted with

$$\sum_i (y_i - b_0 - \sum_j x_{ij} b_j)^2 + k\sum_{j>0} b_j^2 \qquad (7)$$

Thus big variance can be avoided. With the restraint $\sum_{j>0} b_j^2 \le s$, selection of suitable k or s helps reduce both the variance and the bias of estimation significantly. Heuristics method, graphic method or proof could be adopted for selection of suitable k or s. It is convenient to use SPSS 17.0 for ridge regression estimation.

2.5 Regression Improvement Principle

In using nonlinear evaluation methods, one often find

such puzzling phenomenon that evaluation score decreases while the value of some positive indicator increases. Obviously either the indicator or the evaluation method is irrational here. In the cases where no improperness is found in indicators selection, then the cause leading to the phenomenon mentioned above needs to be analyzed carefully. If the phenomenon still exists after eliminating the influence of multicollinearity with a ridge regression, then the current evaluation methods should be given up. In the cases where there is no doubt about the evaluation method, if the phenomenon still exists after eliminating the influence of multicollinearity with a ridge regression, then one should consider the necessity of deleting some indicators. In comparison with the non-restrained principal component analysis, regression adjustment method is a test method widely used for nonlinear evaluation methods. The test and improvement procedure for systematic evaluation methods used in academic journals evaluation is proposed as follows:

1. Select evaluated object. Standardize all indicators, and transform all negative indicators into positive ones.
2. Adopt a systematic evaluation method for academic journal evaluation.
3. Take the evaluation result as dependent variable, then make regression with all the indicators value as independent variables. Improvement is not needed if no coefficient of the indicators is negative.
4. Make multicollinearity test. If there is no multicollinearity and all regression coefficients are above zero, then the test is over. If some regression coefficients are negative, then delete the relevant indicators and return to the second step.
5. Adopt ridge regression for estimation if there is multicollinearity. If some regression coefficients are negative, then delete the relevant indicators and return to the second step.

If all regression coefficients in ridge regression are positive and all coefficients in normal regression are also positive, the adjustment is over.

If all regression coefficients of ridge regression are positive but there is a negative coefficient in normal regression, showing that the negative coefficient is caused by multicollinearity, then this specific nonlinear evaluation method should not be adopted because it gives rise to the abnormal situation where increasing indicators value causes decrease in evaluation result value. In this case evaluation may be performed by taking ridge regression coefficients as weights or by other methods.

3 Variable and Data

The data used in this paper is about the medical academic journals in CSTPCD 2007, a citation database made by the Institute of Scientific and Technical Information of China. The total medical journals recorded in the database is 537. 13 indicators originally selected are: Total Cites, Impact Factor (standard two year synchronous impact factor), Immediacy Index, Ratio of Other Citations, Diffusion Factor (number of cited periodicals divided by number of citations per 100 times), Disciplinary Impact Indicator (ratio of number of within-discipline journals citing the evaluated journal over the total number of journals within the same discipline as the evaluated journal belongs), Disciplinary Diffusion Factor (ratio of number of all journals citing the evaluated journal to total number of journals within the same discipline as the evaluated journal belongs), Cited Half-life, Average Citations per Article, Average Number of Authors per Paper, Number of Provinces, Ratio of Funded Papers (ratio of papers sponsored by funds to total papers), Citing Half-life. Data description statistics are listed in Table 1.

Table 1. Data Description Statistics

Variable	Meaning	Average	Max	Min	Standard deviation
X1	Total Cites	842.12	5343.00	6.00	838.32
X2	Impact Factor	0.43	1.71	0.01	0.26
X3	Immediacy Index	0.05	0.31	0.00	0.04
X4	Ratio of Other Citations	0.82	1.00	0.32	0.13
X5	Diffusion Factor	31.72	77.94	6.74	14.92
X6	Disciplinary Impact Indicator	0.54	1.00	0.01	0.25
X7	Disciplinary Diffusion Factor	4.97	44.67	0.03	3.71
X8	Cited Half-life	7.63	11.40	0.88	1.13
X9	Average Citations per Article	9.63	37.76	2.87	4.23
X10	Average Number of Authors per Paper	4.02	6.48	2.00	0.83
X11	Number of Provinces	23.38	31.00	2.00	5.91
X12	Ratio of Funded Papers	0.25	0.99	0.01	0.19
X13	Citing Half-life	4.31	7.24	0.29	0.84
n	Journal Number	537			

978-1-935068-76-1 © 2011 SciRes.

Data should be standardized before academic journal evaluation. Maximum of every indicator value is set to be 100 and every indicator is adjusted by its proportion. In addition, cited half-life and citing half-life are negative indicators, the standardization method for negative indicators is:

$$x_j^{'} = (1 - \frac{x_j}{\max(x_j)}) \times 100 \quad (8)$$

4 Empirical Results

4.1 Results of Principal Components Regression Treatment

4.1.1 First round Treatment

In principal components analysis KMO test and Bartlett test should be undertaken at first. KMO, usually above 0.5 as needed, is an indicator testing sufficient degree of the sample. In this paper data is processed with SPSS, KMO value is 0.723 and statistical test is passed; Bartlett value is 2893.908, P<0.000，so the statistical test is also passed. In other words, conditions necessary for principle components analysis in journal evaluation are satisfied. Cumulative contribution rate of the six principle components is 80.20%，so these principle components are adopted for evaluation.

In order to test the evaluation result of principle components analysis, regression is undertaken with the evaluation result score as dependent variable and 13 indicators as independent variables. Results by principle components analysis regression 1 are shown in Table 2. Goodness-of-fit is very fine and nearly equals 1, while the regression coefficients of all indicators pass statistical test at 1% significance level. However, the regression coefficients of diffusion factor, cited half-life (a positive indicator transformed by normalization) and regional distribution number are negative, which means the evaluation value will decrease when these three indicators increase in value. It is contradictory to common sense.

Is there any possibility that negative coefficients of the three indicators are caused by multicollinearity? Multicollinearity test is conduced. Multicollinearity test with SPSS shows that no single indicator's variance inflation factor is above 3, so there is no multicollinearity. It is obvious that these three indicators are not appropriate ones.

4.1.2 Second Round Treatment

Now we decide to delete diffusion factor, cited half-life and regional distribution number, then take the remaining 10 indicators for second round principle components analysis. KMO value is 0.712>0.5, Bartlett value is 1716.810, P<0.000, thus here principle components analysis is allowed for evaluation. Cumulative contribution rate of the first 5 principle components is 79.47%, so the first 5 principle components are adopted.

Regression analysis is undertaken with the evaluation result as dependent variable and 10 indicators as independent variables. Results by principle components analysis regression 2 are shown in Table 2. Goodness- of-fit again is very fine and nearly equals 1, the regression coefficients of all indicators pass statistical test at 1% significance level. However, the regression coefficient of cited half-life (a positive indicator transformed by normalization) is negative, so it is obviously a bad indicator.

Is there any possibility of multicollinearity problem here? Another multicollinearity test shows that variance inflation factors of 10 indicators are all less than 3, so there is no multicollinearity. It is obvious that citing half-life is not an appropriate indicator.

4.1.3 Third Round Treatment

Then we delete citing half-life and continue principle components analysis with the remaining 9 indicators. KMO value is 0.704>0.5，Bartlett value is 1644.173, P<0.000, so the conditions for principle components analysis is satisfied. Cumulative contribution rate of the first 5 principle components is 83.97%，so the first 5 principle components are adopted.

Regression analysis is undertaken with the evaluation result as dependent variable and 9 indicators as independent variables. Results by principle components analysis regression 3 are shown in Table 2. Fitting goodness is very fine and nearly equals 1, and the regression coefficients of all indicators pass statistical test at 1% significance level. However, the regression coefficient of immediacy index is negative, so it is obviously abnormal.

Continue the multicollinearity test. The result shows that variance inflation factors of 9 indicators are all less than 3, so there is no multicollinearity. It is obvious that the immediacy index is a bad indicator.

4.1.4 The Fourth Round Treatment

Now delete the immediacy index. The fourth regression is undertaken with the remaining 8 indicators as variables and the regression result by principle components analysis 4 is shown in Table 2. Fitting goodness is very fine and nearly equals 1, and the regression coefficients of all indicators pass statistical test at 1% significance level. This time the coefficients of all indicators are positive.

Continue the multicollinearity test. The result shows that variance inflation factors of the remaining8 indicators are all less than 3, so there is no multicollinearity. Now the treatment process is closed.

4.1.5 Conclusion

The result of adjusting regression coefficients shows that in the above principle components analysis adopted

for academic journals evaluation, only 8 out of the 13 indicators are appropriate: Total Cites, Impact Factor, Ratio of Other Citations, Disciplinary Impact Indicator, Disciplinary Diffusion Factor, Average Citations per Article, Average Number of Authors per Paper, Ratio of Funded Papers. A principle components analysis with these 8 indicators would guarantee the monotonicity of evaluation results, meaning that the evaluation value should increases when indicator values increase.

However, a thorough analysis is needed over the 5 deleted indicators during our exercise. If some indicators were really important the principle components analysis would be inadequate for evaluation.

A special case should be considered. Sometimes even after many cycles of regression adjustment, the expected convergence is still not reached. That is to say, the regression coefficients are still negative for the last few indicators. In this case, the conditions for the principle components analysis are not met, so the principle components analysis cannot be adopted.

Table 2. Principle component data processing results

Variable	PCA regression 1	PCA regression 2	PCA regression 3	PCA regression 4	TOPSIS regression	Grey Correlation Analysis regression
C	-0.917*** (-1306307)	-1.596*** (-1137270)	-1.971*** (-1769633)	-2.215*** (-2005243)	0.039*** (13.463)	3.134*** (48．049)
X1	0.003*** (535627)	0.003*** (230288)	0.005*** (361691)	0.001*** (67972)	0.0008*** (38.297)	0.009*** (19.912)
X2	0.005*** (900679)	0.003*** (178889)	0.002*** (179320)	0.002*** (179299)	0.0007*** (29.970)	0.004*** (7.047)
X3	0.006 (1276787)	0.001*** (41622)	-0.001*** (-86408)	--	0.0007*** (35.573)	0.005*** (10.907)
X4	0.001*** (160342)	0.008*** (664326)	0.010*** (904981)	0.009*** (834535)	0.0005*** (24.842)	0.008*** (16.835)
X5	-0.001*** (-119540)	--	--	--	0.0008*** (39.934)	0.008*** (17.455)
X6	0.001*** (229330)	0.004*** (502654)	0.005*** (768061)	0.002*** (309340)	0.0007*** (69.212)	0.006*** (24.625)
X7	0.005*** (547553)	0.011*** (441143)	0.016*** (653947)	0.007*** (300954)	0.0005*** (14.640)	0.004*** (4.603)
X8	-0.003*** (-516338)	--	--	--	0.0008*** (31.097)	0.008*** (14.541)
X9	0.009*** (1748183)	0.007*** (500452)	0.004*** (288924)	0.013*** (906988)	0.0006*** (28.082)	0.004*** (9.298)
X10	0.007*** (1636043)	0.007*** (556573)	0.005*** (400043)	0.010*** (787512)	0.0004*** (19.149)	0.004*** (10.128)
X11	-0.002*** (-744457)	--	--	--	0.0007*** (56.998)	0.007*** (22.672)
X12	0.005*** (1688433)	0.006*** (687745)	0.005*** (516077)	0.009*** (1028889)	0.0007*** (49.707)	0.004*** (14.902)
X13	0.002*** (409660)	-0.005*** (-385825)	--	--	0.0007*** (36.605)	0.005*** (12.303)
R^2	1.000	1.000	1.000	1.000	0.992	0.941
VIF>10	NO	NO	NO	NO	NO	NO
KMO	0.723	0.712	0.704	0.661	--	--
Bartlett	2893.908	1716.810	1644.173	1317.259	--	--
Cumulative%	80.20	79.47	83.97	79.06	--	--
n	537					

Note: *** means passing statistical test at level 1%; data source: Chinese S&T Journal Citation Reports 2008

4.2 TOPSIS Evaluation and its Treatment

We also tried TOPSIS for evaluation. Similarly, we undertake regression analysis with the evaluation result as dependent variable and 13 indicators as independent variables. Results by "TOPSIS regression" are shown in Table 2. Fitting goodness R^2 is very high and equals 0.992, and the regression coefficients of all indicators pass statistical test at 1% significance level. What is more, the regression coefficients of all variables are positive, and no multicollinearity is found in the test. All these show that the problem of evaluation value decreasing while indicator value increasing does not exist in TOPSIS and the adjustment is not needed.

4.3 Grey Correlation Analysis Evaluation and its

978-1-935068-76-1 © 2011 SciRes.

Treatment

We conducted grey correlation analysis for evaluation. First we set resolution coefficient ζ as 0.5, then undertake regression analysis with the evaluation result as dependent variable and 13 indicators as variables. Results by "grey correlation analysis regression" are shown in Table 2. Fitting goodness R^2 is very high and equals 0.941; the regression coefficients of all indicators pass statistical test at 1% significance level; the regression coefficients of all variables are positive; no multicollinearity is found in the test. All these indicate that the problem of evaluation value decreasing while indicator value increasing does not exist in grey correlation analysis and the adjustment is not needed.

5 Conclusions and Discussion

5.1 Deficiency of Nonlinear Evaluation Methods

So far, dozens of nonlinear evaluation methods have been created based on different principles. The unusual case that evaluation value decreases when the value of some positive indicator increases was found in the principle component analysis by scholars. Such an abnormality may occur in other nonlinear evaluation methods, however, it has not been found yet or paid attention to. Regression adjustment method, a robust method for test and adjustment, is suitable for all nonlinear evaluation.

5.2 Caution Needed before Indicator Deletion

Regression adjustment method can be used for selecting evaluation indicators. If the value of some positive indicator increases while its regression coefficient turns out to be negative during the regression, implying that this indicator is abnormal in comparison with other indicators, then a thorough analysis is needed to decide whether or not to delete the indicator.

It is our recommendation that if some "abnormal" indicators should not be deleted because of their extraordinary importance, then the already adopted evaluation method has to be abandoned. The regression adjustment method is taken as an objective, in some degree, method for selecting evaluation indicators.

5.3 Cycles in Regression Adjustment

In general case, if the evaluation model is stable after one or two round of regression adjustments as well as deletion of some unsuitable indicators, the nonlinear evaluation method examined could be adopted safely. On the Contrary, if the model becomes robust only after many rounds of regression adjustment as well as deletion of many indicators, the caution is called as to whether adopt this nonlinear evaluation method or not. Of course, the nonlinear evaluation method should be abandoned if convergence cannot be reached even after many rounds of regression adjustment.

5.4 Possible Nonlinear Evaluation Methods Requiring no Test and Improvement

In this paper TOPSIS and grey correlation evaluation do not require further improvement because the abnormal case does not happen. Does it happen just by chance or due to their own intrinsic characteristics? More evaluation exercises are needed before we can answer the above question because there seems no strict mathematical proof is available.

5.5 Many Problems in Journal Evaluation Methods Requesting Further Study

This paper focuses on the test and improvement of nonlinear evaluation methods and the fundamental principles introduced here should be feasible for all nonlinear evaluation methods. However, it does not mean that our stand is to support adopting nonlinear evaluation methods. True, for all linear evaluation methods, the abnormality never happens that the evaluation score decreases while the value of some positive indicator increases. However, such a situation would disappear even for nonlinear evaluation methods after adjustment with the procedure suggested in this paper. So there are still dozens of evaluation methods available for journal evaluations, and they would yield different evaluation results. Obviously it is necessary to decide how to select a specific evaluation method. This is again a huge challenge.

6 Acknowledgement

Thanks for a research fund provided by The National Social Science Foundation of China (Grant No. 10FTQ003) and a grand Sponsored by the National Natural Science Foundation of China (Grant No. 70973118).

References

[1] A.E. Hoerl, R.W. Kennard, Ridge regression: biased estimation for nonorthogonal problems[J]. Technometrics , 1970(12):55-67

[2] Braun T., Glanzel W. World flash on basic research. A topographical approach to world publication output and performance in science[J].1990,Scientometrics,19:159-165

[3] Cheng Hanzhong. Application of the principal component analysis in academic journal evaluation [J]. Chinese journal of scientific and technical periodicals(in Chinese), 2004(6): 658-660

[4] C.L. Hwang and K. Yoon, Multiple attribute decision Making: methods and applications [M]. Springer-Verlag, New York, (1981).

[5] Chien-Ta Bruce Ho;Desheng Dash Wu. Online banking performance evaluation using data envelopment analysis and principal component analysis[J]. Computers & Operations Research,vlo 36,2009(6):1835-1842

[6] D. Krishna;S. Rama Mohan;B. S. N. Murthy; A Ramakrishna Rao. Performance evaluation of public research institutes using principal component analysis[J]. Journal of Scientific & Industrial Research. vol 61,2002(11):940-947

[7] Damodar N. Gujarati. Basic Econometrics [M]. Beijing：China Machine Press(in Chinese), 2004: 204-205

[8] Deng Julong. Theory of grey system[M]. Wuhan: Huazhong university of science and technology press(in Chinese).
[9] Hirsch,J.E. An index to qualify an individual's scientific research output. Proceeding of the national academy of sciences USA[M]. 2005,102:16569-16572
[10] Hou Wen. Discussing to comprehensive evaluation by principal component [J]. Application of statistics and management (in Chinese), 2006(3): 211-214
[11] J.C.M. Pires; M.C. Pereira; M.C.M. Alvim-Ferraz; F.G. Martins. Identification of redundant air quality measurements through the use of principal component analysis [J]. Atmospheric environment, vol 43, 2009(25): 3811-3818
[12] K. Yoon and C.L. Hwang, Multiple Attribute Decision Making: An Introduction [M]. Sage, Thousand Oaks, CA, (1995).
[13] Li Kaiyang, Jia Yuping. Fuzzy comprehensive evaluation for full-text periodical data bases on AHP [J]. Information Science (in Chinese), 2005(11): 1688-1703
[14] Li Xiujie, Cheng Jingwu. Applying discriminant analysis to build index system of periodicals estimate [J]. The Journal of the Library Science in Jiangxi (in Chinese), 2006(3): 48-50
[15] Markpin, T., Boonradsamee, B., Ruksinsut, K., Yochai, W., Premkamolnetr, N., Ratchatahirun, P. and Sombatsompop, N. (2008). Article-count impact factor of materials science journals in SCI database. *Scientometrics,* 75(2), 251–261
[16] Moed. H. F.,R.E. DE Bruin.. New bibliometric tools for the assessment of national research performance[J]. Scientometrics,1995,33:381-422.
[17] Pang Jingan Zhang Yuhua etc. Study on comprehensive evaluation index system of Chinese academic journals [J]. Chinese journal of scientific and technical periodicals (in Chinese), 2000(11): 217-219
[18] Qiu Junping, Zhang Rong etc. Study on academic journal evaluation index system and quantitative method [J]. New Technology of Library and Information Service (in Chinese), 2004(7): 23-26
[19] Schubert A., Glanzel W., Braun T. Relative citation rate: A new indicator for measuring the impact of publications. In: D. Tomov, L. Dimitrova (eds), Proceeding of the first national conference with international participation on scientometrics and linguistics of scientific text, Varna [M]. 1983, PP80-81
[20] Su Weihua. Study on multiple attribute evaluation: theory and methods [D]. Xiamen: Xiamen University Doctoral Dissertation (in Chinese), 2000: 151-166
[21] Tahmina Khatun. Measuring environmental degradation by using principal component analysis [J]. Environment, Development and Sustainability, vol 12,2009(2): 439-457
[22] Vinkler P. Evaluation of some methods for the relative assessment of scientific publications [J]. scientometrics, 1986 (10): 157-177
[23] Wang Xiaowei, Yang Bo et**c.** Secondary relative quality evaluation for sci-tech periodical publications [J]. Acta Editologica(in chinese), 2003(6): 231-232
[24] Weiping Yue, Concepcion S. Wilson. Measuring the citation impact of research journals in clinical neurology: a structural equation modeling analysis [J]. Scientometrics, 2004(3):317-334
[25] Y.J. Lai, T.Y. Liu and C.-L. Hwang, TOPSIS for MODM, European Journal of Operational Research [J]. 76(3):486-500, (1994).

978-1-935068-76-1 © 2011 SciRes.

A Scientometrics Analysis of the Participation in International Science and Technology Cooperation on China's Scientific and Technical Ability

Junpeng YUAN[1], Lan XUE[2], Jie SU[3]

[1]*Institute of Scientific and Technical Information of China, Beijing, 100038*

[2]*School of Public Policy & Management, Tsinghua University, Beijing, 100084*

[3]*School of Management Science and Engineering, Central University of Finance and Economics, Beijing, 100081*

Email: junpengyuan@gmail.com

Abstract: Based on the published paper from 1994-2006 with a Chinese address samples, included in the Web of Science (including SCI, SSCl and A&HCI), this paper tries to adopt scientometrics and data mining approach to analyze patterns of Chinese international academic research collaboration., including disciplinary distribution of papers, institution types, ratio of international collaboration. This analysis revealed some interesting characteristics of China's research cooperation with the world. Policy recommendations are made on how to improve the productivity of basic research and international science and technology cooperation in China.

Keywords: International Science and Technology Cooperation; Scientometrics; Data Mining; Web of Science

国际科技合作对中国的科技能力影响分析*

袁军鹏[1]，薛澜[2]，宿洁[3]

[1]中国科学技术信息研究所，北京，100038

[2]清华大学公共管理学院，北京，100084

[3]中央财经大学管理科学与工程学院，北京，100081

Email：junpengyuan@gmail.com

摘　要：本文利用科学计量学、数据挖掘技术，以 1994-2006 年 Web of Science（含 SCIE，SSCI 和 A&HCI）收录的中国发表论文为样本，从论文的机构类型分布、引用影响分布、国际科技合作与全部中国文献引用影响的对比、中国与国际著名机构的合作分布进行了研究，从科学计量学的视角描绘了 1994 年以来的中国参与国际科技合作对中国科技能力的影响以及中国自主创新能力的演变趋势。进而提出了中国各机构加强国际合作，提高自主创新能力应采取的对策思路。

关键词：国际科技合作；科学计量学；数据挖掘；Web of Science

1　引言

近年来，随着经济全球化的不断发展，一些研究人员开始关心不断发展的科技全球化进程对中国科技能力的影响[1,2]。任何一项科学研究和技术创造，都要以撰写必要的科学文献为其最后阶段，科学文献的数量和质量无疑是对科技能力的一种量度[3]。

学术论文、专著作为传递新学术思想、成果的主要的物质载体，它们之间的关系并不是孤立的，而是相互联系的，突出表现在科学文献之间的相互引用，因此，引文分析就成为是文献计量学、情报计量学、科学计量学的重要组成部分。引文分析（Citation Analysis）就是利用图论、数理统计及其它数学、逻辑思维方法，对科技文献的引用或被引用现象和规律进行分析，以便揭示出它们所蕴含着的研究对象具有的特征或者对象之间的关系[4]。通过深入分析文献间互相引用的关系，引文分析在评价研究团队和个人的科研绩效、揭示学科特点和结构以及反映科学研究的焦点领域和发展态势等方面都发挥着重要的作用[5]。澳大利亚政府工业与科学技术部的工业经济局(BIE)，曾于 1995 年对该国的科技体制和科学研究水平进行了评价。对其科研活动的绩效评价主要根据 1981-1994

*基金项目：国家自然科学基金科研课题（70973118）中国国家博士后基金（20060390049），中信所预研基金

年 SCI 数据库中的指标,通过对科学论文与引文进行国际比较以及对合作论文的考察，分析该国在若干学科领域中的优势与劣势以及国际合作状况(包括与不同地域、在不同研究领域的合作),进而对国家基础研究的整体水平进行评估[6]。近年来,国外的研究者对中国科学异军突起表现出极大的关注[7,8]，David A. King 用引文分析的方法对 31 个国家和地区的科学影响进行了分析，其中中国在归一化的篇均引文指标中位于倒数第四[8]。

本文在上述研究基础上，利用数据挖掘、科学计量学技术，以 1994-2006 年 Web of Science（含 SCIE,SSCI 和 A&HCI）收录的中国发表论文为样本，对中国机构与国际各类型机构合著 WoS 论文的机构类型分布、引用影响分布、国际科技合作与全部中国文献引用影响的对比、中国与国际著名机构的合作分布进行研究，试图描绘 1994 年以来的中国参与国际科技合作对中国科技能力的影响以及中国自主创新能力的演变趋势。该研究将为我国更好地整合和利用全球优势资源，在更大范围、更广领域和更高层次上参与国际科技合作与竞争，提高自主创新能力和建设创新型国家提供科学决策支持。

2 数据来源与数据库的建立

科学引文索引(SCI)数据库历来被公认为世界范围最权威的科学技术文献的索引工具和科学计量、科学评价的重要工具，是因为它收录期刊的标准高，代表性强以及独特的编排方法所带来的特殊作用。为了获得更全面更权威的数据，我们将 SCI 数据库和 Thomson Scientific 公司的 ISI 开发的另外两个同等重要的数据库——社会科学引文索引数据库(SSCI)和艺术与人文科学引文索引数据库(A&HCI)作为研究的数据来源。

2006 年 10 月 25 日，在清华大学购买的 WoS 中，以“CU=Peoples R China”为检索对象[1]，检索出 WoS 中所有中国发表的文献，得到结果为 441769 篇。同日在 Thomson Scientific 公司的 WoS 以同样的条件得到结果为 442697 篇，其查全率（Recall）为 99.79%，所以该数据是可靠的。利用信息抽取和自动监测技术将检索到文献信息建立基于 SQL Server 的 ChinaSCI 数据库，得到 441769 条文献记录。我们定义中国国际合著关系为一篇文章的合著作者中至少有一位来自中国并且至少有一位来自其他国家。以此为依据，设计国际科技合作自动识别算法，得到所有的中国的国际科技合作的 WoS 文献（SCI、SSCI 和 A&HCI）为 103585 篇。我们从著者地址出发，将涉及的 44149 个机构分为大学、研究机构、企业、医院、政府与国际组织、其它几种类型[2]，然后分析各类型机构的文献情况。

3 结果分析

3.1 各机构类型总体数量分布分析

1994.1-2006.10 期间，中国共有 26627 个机构发表过 WoS 文献[3]，总体情况如图 1 所示。

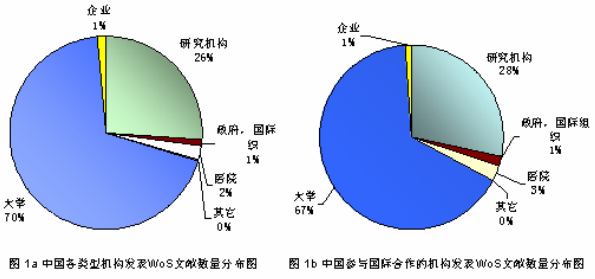


图 1a 中国各类型机构发表WoS文献数量分布图　　图 1b 中国参与国际合作的机构发表WoS文献数量分布图

Figure 1. WoS literature quantity distribution of each type institution of China

图 1. 中国的各类型机构 WoS 文献数量分布图

数据来源: 本文中未作说明的图表均来自于Web of Science（清华大学镜像），2006年10月25日.

在中国，学校和研究机构是 WoS 文章发表的主要类型，占到中国所有文章的 96%，而且在参与国际合作的机构中，学校和研究机构的文章数量也占到了 95%。对比图 1.a 和 1.b 可以看出，在国际合作的文献中，研究机构的比例上升为 28%，医院由 2%上升为 3%，这二者的上升恰恰是由于学校的比例下降导致，其它类型的机构发表的文献所占的比例没有变化。从中可以看出，学校是发表 WoS 的绝对主要类型，其次是研究机构，医院排在第三位，企业等在其中分别贡献 1%的文献，这也符合 Jaffe & Trajtenberg 在 2002 年的判断[9]：企业并不是科技文章的主要发表者，其创新的成果多在专利中体现。

[1] 清华大学的 web of Science 回溯到 1994 年，所以 Time span =1994-2006，另外设定 Doc Type=All document types；Language=All languages; Databases=SCI-EXPANDED, SSCI, A&HCI。

[2] 由于机构名称书写不规范，加上很多大学改名，所以，机构数量存在较大的误差，分类后则可以避免这个问题。

[3] 不包括港、澳、台的 505 个机构，同样，机构数量有一定误差。

978-1-935068-76-1 © 2011 SciRes.

3.2 引用影响因子分析

这里有一个公理性的潜在理念是：引用是质量的关系函数，即一般情况下，被引用频次反映的是某种学术思想的影响程度，一项研究成果被引用的次数越多，表明这项成果的影响力越大[10]，所以，我们定义引用影响因子 CI (Citation Impact)为

$$CI = CT / N \quad (1)$$

其中，CT ＝总被引用次数，N＝总文献数量。

（1）各类型机构的引用影响因子分析

利用公式（1），计算各类型机构的历年的引用影响因子，得到图 2。

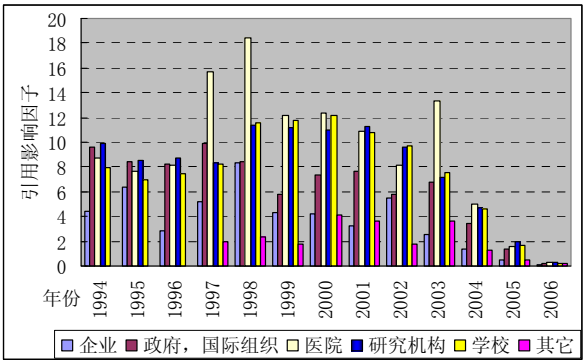


Figure 2. Time distribution of all the impact factor of WoS reference of each type institution of China

图 2. 中国的各类型机构发表的全部 WoS 的引用影响因子按年份分布

从图 2 可以看出，就引用影响因子来看，由于医院基本只发表特定学科的文章，而这些学科的引用影响因子较高，所以导致医院的一直较高，而且在 1997、1998、2003 年明显高于其它类型的机构。学校和研究机构不但文献数量较多，而且二者的引用影响因子也较高。从年代的变化趋势来看，医院的引用影响因子自 1994 年的 8.70 开始下降到 1995 年的 7.63 然后又上升到 1996 年 8.13，随后有个快速上升期到 1998 年的 18.41，然后开始呈现下降趋势至 2002 年的 8.15，2003 年上升到 13.37，其后一直呈下降趋势。学校和研究机构的引用影响因子历年基本保持持平状态，特别是在 1997 年以后，二者的趋势基本完全一致。政府，企业的引用影响因子历年也是呈现波动的趋势，具体趋势请见图 2。

（2）各类型机构的国际科技合作的引用影响比较分析

为了比较各类型机构参与国际科技合作与全部的 WoS 文献的引用影响因子的区别，我们首先以历年的国际合作文献的 CI 除以历年全部文献的 CI，然后按照不同区间统计其分布频次，得到图 3。

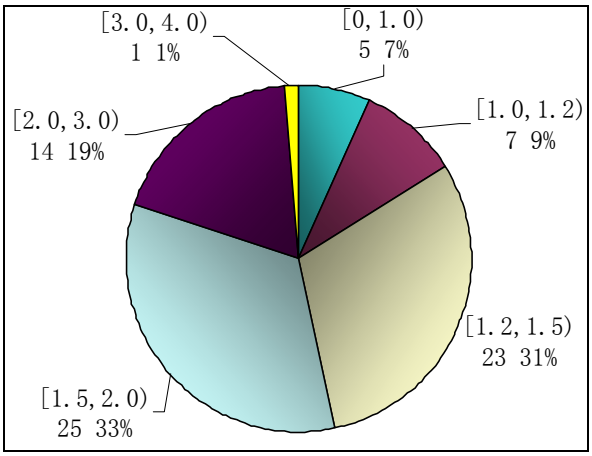


Figure 3. Frequency distribution of the ratio of all the cooperation papers of each type institution of China and all the impact factor of WoS reference

图 3. 中国的各类型机构发表的合作文献与全部 WoS 的引用影响因子比值的频次分布

注：图中，$[0,1.0)$ 表示比值分布在大于等于0且小于1.0的区间中，5表示频次为5个，7%表示这5个频次占所有频次的比例为7%.

从图 3 来看，小于 1.0 的比值只有 5 个，占所有比值的 7%，说明只有 5 次（7%）是国际合作文献的 CI 小于全部文献的 CI，其余的 93%全部是国际合作文献的 CI 高于全部文献的 CI，而且，有 53%的比例分布在 $[1.5,4.0)$ 区间，说明有 40 次的国际合作文献的 CI 是全部文献的 CI 的 1.5 倍以上，这充分说明参与国际科技合作有助于提高我国论文的引用影响力，有助于提高我国的科技能力。有 40%的比例分布在 $[1.0,1.5)$ 区间，说明有些年，有些机构的中国的全部文献与合作文献的引用影响因子差距并不是很大，中国在充分利用国际科技合作，加强自主创新能力后，完全有可能提高我国论文的引用影响。由于统计源的学科结构差别，以及各个学科自身发展的特点和特有引文行为的不同，如科学家研究行为的社会性、学科间交叉渗透的程度、学科发展所处的阶段等，引用影响因子在各个学科之间具有较大的差异性，由此产生了不同学科论文之间影响因子的不可比性，而且，引文中还有转引、崇引、不恰当的自引、伪引等自身的局限性[11]，所以，仅仅使用引文分析的方法不能充分说明问题，下面我们进一步进行分析我国机构与国外

知名机构的合作情况。与国际知名机构合作的文章数量变化趋势可以从一个方面表明中国的研究质量和科技能力的变化。

3.3 与国际知名机构的合作情况

（1） 国际著名大学

依据美国新闻周刊全球前 100 著名大学最新排名选取其中的前 30 名：

Table 1. World TOP 30 famous universities
表 1. 全球前 30 著名大学

1.哈佛大学	11.密歇根大学	21.苏黎世瑞士联邦理工学院
2. 斯坦福大学	12.加州大学洛杉矶分校	22. 西雅图华盛顿大学
3.耶鲁大学	13. 宾夕法尼亚大学	23. 加州大学圣地亚哥分校
4.加州理工学院	14. 杜克大学	24. 约翰·霍普斯金大学
5.加州大学伯克利分校	15. 普林斯顿大学	25. 伦敦大学学院
6.剑桥大学	16. 东京大学	26. 洛桑瑞士联邦理工学院
7.麻省理工学院	17. 帝国理工学院	27.德州大学奥斯丁分校
8.牛津大学	18. 多伦多大学	28. 威斯康辛大学麦迪逊分校
9.加州大学旧金山分校	19. 康乃尔大学	29. 京都大学
10.哥伦比亚大学	20. 芝加哥大学	30. 明尼苏达大学

计算中国与这些大学的历年合作趋势，得到图 4：

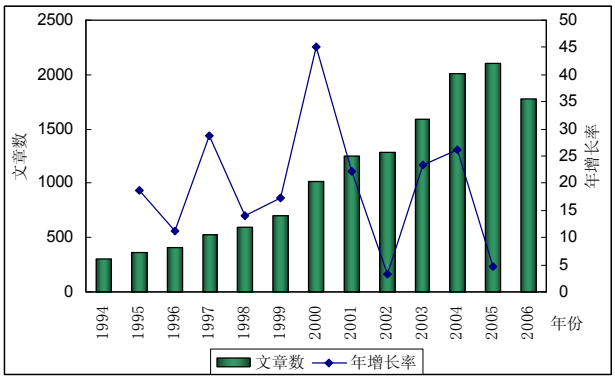


Figure 4. Trend of cooperated paper number between China and international famous universities (TOP30) by year
图 4. 中国与国际著名高校（前 30 位）合作历年文章数变化趋势

如图 4 所示，中国与这 30 所著名高校合作的文章数量逐年增长，由 1994 年的 309 篇一直增加到 2005 年的 2101 篇，年增长率一直大于 0[4]。但是，从增长速度来看是不均衡的：1996-1997、1999-20002003-2004 期间的增长率（IR，Increasing Rate）较高，尤其是 2000 的 IR 达到 45.16%，呈现三个波峰；1996，1998，2002 的 IR 处于波谷。进一步，我们分析中国与其中前 10 名的特别优秀高校的合作情况，得到图 5：

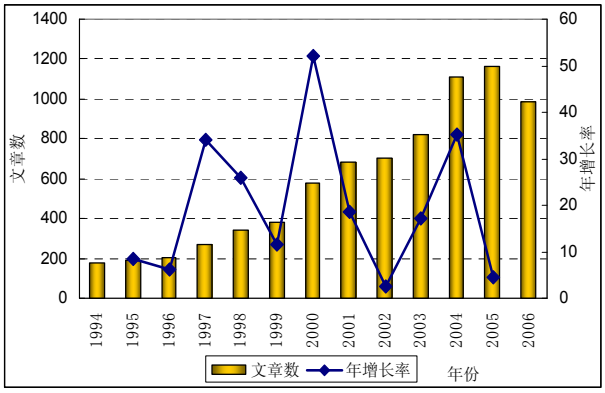


Figure 5. Trend of cooperated paper number between China and international famous universities (TOP10) by year
图 5. 中国与国际著名高校（前 10 位）合作历年文章数变化趋势

如图 5 所示，中国与这些著名高校合作的文章数量逐年增长，由 1994 年的 175 篇一直增加到 2005 年的 1163 篇，年增长率一直大于 0。但是，从增长速度来看是不均衡的：1996-1997、1999-2000、2003-2004 期间的增长率（IR）较高，都超过到 34%，呈现三个波峰；1997-1998 增长率达到 25.83%；1998-1999、2000-2001、2002-2003 期间的增长率为 11.4%-18.5%；1994-1995、1995-1996、2001-2002、2004-2005 的增长率只有 2.6%-8.6%左右。

（2） 国际优秀机构

对于优秀机构的认定目前没有一个权威的、统一的标准，我们利用 ESI 数据，按照发表论文数、被引用次数、篇均被引用次数、文章数大于 500 篇的篇均被引用次数、文章数大于 1000 篇的篇均被引用次数、文章数大于 2000 篇的篇均被引用次数等标准取排名前 20 位的机构，对这 120 个机构去除重复的机构的得到 83 个机构，我们分析中国与这 83 个机构合作发表

[4] 2006 年数据只有 10 个月，所以增长呈负数。

978-1-935068-76-1 © 2011 SciRes.

论文的情况得到图 6：

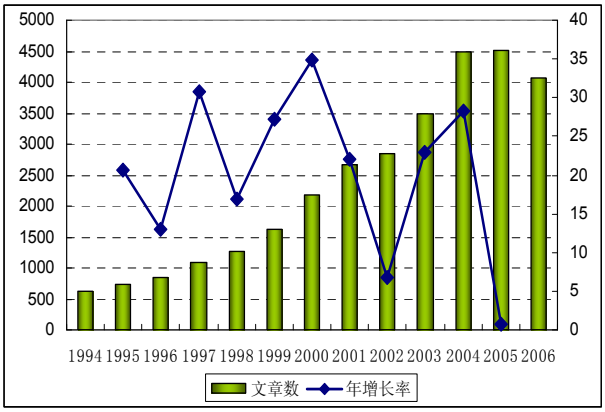


Figure 6. Trend of cooperated paper number between China and world's excellent institutions
图 6. 中国与国际优秀机构合作历年文章数变化趋势

从图 6 可以看出，我国与这些优秀机构的合作文献数量呈现逐年增长的趋势，其文章数量由 1994 年的 612 篇增加到 2005 年的 4515 篇。但其增长速度也是不均衡的：1997，2000 年的 IR 分别达到 30.82%和 34.77%；2004 年的 IR 达到 28.18%，这三年形成 3 个波峰，对比与国际著名大学的趋势来看，这三个波峰是同一年达到的；1996，1998，2002，2005 年的 IR 形成 4 个波谷，与国际著名大学的趋势对比可以发现，形成波谷的年份也很类似，仅在 1998 年时与前 10 名高校的有区别。

考虑到以上 83 个机构过多，尤其是按照篇均被引用次数引入了发表篇数仅为 1-2 的机构，这些机构是否为优秀机构不是很稳定，所以我们取这 83 个机构中在上述方法中出现 2 次以上的为比较优秀机构，得到 30 个机构：

Table 2. World's good institutions
表 2. 国际比较优秀机构

AMGEN	GENENTECH INC	UNIV TEXAS	STANFORD UNIV	UNIV TOKYO
BURNHAM INST	UNIV MICHIGAN	NHGRI	NIAID	UNIV TORONTO
CANCER RES UK	HARVARD UNIV	NHLBI	UNIV PENN	UNIV WASHINGTON
NETHERLANDS CANC	HOWARD HUGHES	UNIV CALIF	GLAXO WELLCOM	UNIV WISCONSIN
INST	MED INST	BERKELEY	E INC	
DANA FARBER CANC INST	JOHNS HOPKINS UNIV	ROCKEFELLER UNIV	UNIV CALIF LOS ANGELES	WELLCOME TRUST SANGER INST
EUROPEAN MOLEC BIOL LAB	MAX PLANCK SOCIETY	SCRIPPS RES INST	COLD SPRING HARBOR LAB	SALK INST BIOL STUDIES

分析中国与这 30 个机构合作的情况得到图 7。

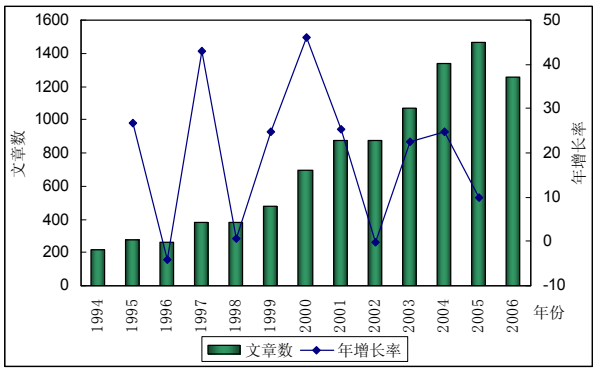


Figure 7. Trend of cooperated paper number between China and world's good institutions
图 7. 中国与国际比较优秀机构合作历年文章数变化趋势

从图 7 可以看出，我国与这 30 个比较优秀机构的合作文献数量与前 30 名高校一样，也呈现逐年增长的趋势，而且其增长速度也基本与前 30 名高校的一致。最终 2005 年的文章数量达到 1467 篇，较前 30 名高校的 2101 篇少了 634 篇，总体来说，中国与前 30 名高校的合作文章数量达到 13934 篇，而和这 30 个机构的合作总文章数只有 9568 篇，差距为 4366 篇。所以，对比这 30 个机构与 30 个高校可以看出，中国已经与这些优秀的机构和高校开展了合作，而且其增长趋势基本一致，但是与机构的合作的文章数量较纯粹与高校合作的数量少很多。

为了与前 10 名高校对比，我们取这 83 个机构中在上述方法中出现 3 次以上的为特别优秀机构，得到 7 个机构：BURNHAM INST、CANCER RES UK、EUROPEAN MOLEC BIOL LAB、GENENTECH INC、HOWARD HUGHES MED INST、ROCKEFELLER UNIV、SALK INST BIOL STUDIES，分析中国与这 7 个机构合作的情况得到图 8。

从图 8 可以看出，中国与这些机构虽然有合作的文献，但文献数量较少，总量只有 131 篇文献，数量基本保持一定的增长，但是 2000-2001、2002-2003 年的增长率为负值，其增长率也是不均衡的。对比前 10 名高校的数据可以看出：虽然二者都有三个波峰，但是三个波峰的时间点基本不一致，除了 1997 年外。在数量上来看，与前 10 名的高校合作的文章数量达到 7614 篇，较与 7 名机构的 131 篇高出 57 倍多，更加证明了 30 个机构的结论。

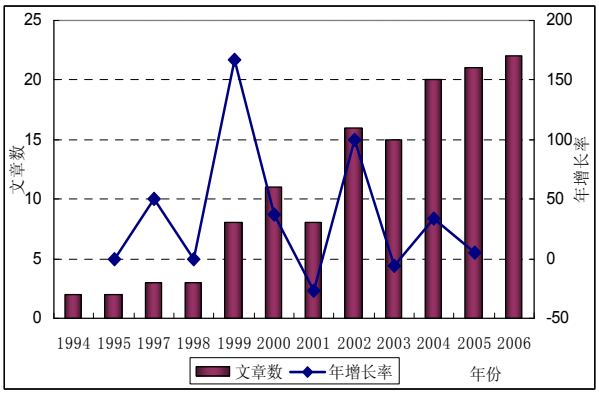


Figure 8. Trend of cooperated paper number between China and world's outstanding institutions

图 8. 中国与国际特别优秀机构合作历年文章数变化趋势

4 结论与政策建议

对中国机构与国际各类型机构合著 WoS 论文的机构类型分布、引用影响分布、国际科技合作与全部中国文献引用影响的对比、中国与国际著名机构的合作分布的考察，使我们从科学计量学和数据挖掘的视角了解了 1994 年以来的中国参与国际科技合作对中国科技能力的影响以及中国自主创新能力的演变趋势。从引用影响上来看，医院最高，学校和研究机构这两个发表文章最多的机构保持比较稳定的态势，而且国际科技合作有助于加强我国论文的引用影响，并且，中国已经开始与国际著名机构开展合作，而且基本呈现逐年上升的态势，但是其增长率却是不均衡的，从数量上来看，中国与国际著名高校合作的论文数量远远大于与优秀研究机构的。

这些发现对于进一步加强我国参与国际科技合作有着重要的意义。首先，我们应该进一步鼓励中国学者积极参与国际科技合作，这有助于提高我国学者的论文的引用影响力。同时，国家在鼓励国际科技合作的总体政策框架下，也应当注意到不同机构类型和不同机构层次的差异，通过具体的合作政策，鼓励中国学者选择国际上优秀的机构和著名的学者进行合作，加强与世界一流机构的交流与合作，国家有关政策应该鼓励中国学者在这方面发挥更大作用。

本研究基于 WoS 论文，我们假定文献能够反映科研的实际水平和学者的主要思想，但其信息传递的可靠性和有效性问题需要进一步研究。另外，引文分析自身的局限性也使本文的结论受到限制。

References（参考文献）

[1] Research group of Annual Report of Science and Technology Development of China(2000). Annual Report of Science and Technology Development of China—Globalization of Science and Technology and its challenges in China. Beijing: Social Sciences Academic Press; 2000.
中国科技发展研究报告（2000）研究组. 中国科技发展研究报告(2000)——科技全球化及中国面临的挑战. 北京: 社会科学文献出版社; 2000.

[2] Xiaojuan Jiang. [J]. Comprehension of Technical Globalization —Resources Reorganization, Advantages Integration and the Ascension of Independent Innovation Ability [J] *Management World*. 2004, 6: 4-13.
江小涓. 理解科技全球化——资源重组、优势集成和自主创新能力的提升[J]. 管理世界. 2004, 6: 4-13.

[3] Liju Mao, Qingkui Xi, Bin Zhang. Statistic Analysis of the Relation between Papers Collected by SCI and Sci-Tech Input and Research Incentive Mechanism in the Nanjing Agricultural University[J]. *Journal Of Library And Information Sciences In Agriculture*. 2006, 04.
毛莉菊, 席庆奎, 张彬等. 南农大 SCI 论文量与科研经费投入及科研激励机制的关系分析[J]. 农业图书情报学刊. 2006, 04.

[4] Garfield E. Citation indexes for science. A new dimension in documentation through association of ideas (Reprinted from Science, vol 122, pg 108-11, 1955)[J]. International Journal of Epidemiology. 2006, 35(5): 1123-1127.

[5] Nan Ma, Jianchen Guan. The History And Perspectives on Research Fronts Identification Based On Citation Analysis[J] *Forum On Science And Technology In China*. 2006, (4): 110-114.
马楠, 官建成. 利用引文分析方法识别研究前沿的进展与展望[J]. 中国科技论坛. 2006, (4): 110-114.

[6] Xu Gong. SCI, Appraisement of scientific Research and Allocation of Resource[J]. *Optimization Science & Technology Review*. 2002, (2): 36-39(Ch).
龚旭. SCI、科研评价与资源优化配置[J]. 科技导报. 2002, (2): 36-39.

[7] Kostoff R. N. The (Scientific) wealth of nations[J]. Scientist. 2004, 18(18): 10-10.

[8] King D. A. The scientific impact of nations[J]. Nature. 2004, 430(6997): 311-316.

[9] Jaffe A. B., Trajtenberg M. Patents. Citations. and Innovations: A Window on the Knowledge Economy. Cambridge MA/London: MIT Press; 2002.

[10] Harsanyi. M. A. Multiple authors,multiple problems bibliometrics and the study of scholarly collaboration - a literature review[J]. Library & Information Science Research. 15(4): 325-354.

[11] Long Pang, Peifu Zhang, Liying Yang. Comparison of Citation Method between China and Foreign Countries[J]. *Journal Of Shanxi University(philosophy And Social Science Edition)* 2006, 29(3): 134-137.
庞龙, 张培富, 杨立英. 引文分析方法在国内外应用的比较研究[J]. 山西大学学报(哲学社会科学版)2006, 29(3): 134-137.

978-1-935068-76-1 © 2011 SciRes.

Current Research Status of the 4G Technology around the World

Zhenglu YU[1], Yong WANG[2]
Institute of Scientific and Technical Information of China (ISTIC), Beijing, China
Email: luluyu@istic.ac.cn

Abstract: Information technology plays an important role in social and economic development. 3G technology has been widely applied and people begin to pay more attention on 4G technology. This article analyses the 4G technologies by the perspective of papers and patents, comparing the 4G papers and patents between countries (regions), institutes and companies. It also analyses the key technologies of 4G. China, the United States and South Korea are the top paper output countries. The growth of Chinese paper is significant. China and the United States are very attractive markets of 4G who attract the most 4G patents. South Korea and Japan attach great importance to the output of 4G patents. Samsung, Qualcomm and Matsushita have more advantages in this field. MIMO and radio link for data networks are getting more and more attention.

Keywords: 4G (fourth-generation); Patent analysis; Paper analysis

1 Introduction

ITU (International Telecommunication Union) Secretary-General Hamadoun Touré said, "ICTs and broadband networks have become vital national infrastructure — similar to transport, energy and water networks — but with an impact that promises to be even more powerful and far-reaching." With large-scale commercial deployment of 3G networks, 4G communication technology has developed as the technology focus of global communication companies and institutes. In October 2010, ITU completed the assessment of six candidate submissions for the global 4G mobile wireless broadband technology, otherwise known as IMT-Advanced. LTE-Advanced and Wireless MAN-Advanced technologies were each determined to have successfully met all of the criteria established by ITU for the first release of IMT-Advanced.

4G technology provides a global platform on which to build the next-generations of interactive mobile services that will provide faster data access, enhanced roaming capabilities, unified messaging and broadband multimedia. These key enhancements in wireless broadband can drive social and economic development [1]. Pre-4G technologies such as first-release 3G Long term evolution (LTE) has been on the market since 2009. According to the Global mobile Suppliers Association (GSA), as of January 2011, there have been 180 operators in 70 countries which are deploying, trialling or evaluating LTE [2]. LTE proves to be the fastest developing mobile communications system technology ever. Enterprises and research institutes have devote many resources to 4G technology. Outputs in this field are growing significantly these years.

Papers and patents are the main output of research. They can reflect the research level of one country (region) or organization to a great extent [3-5]. Papers can reveal scientific development, and patents can reveal technical development [6-7]. From this perspective, the whole status of 4G technology will show up. This article is a macroscopical study, so the specific technology is not discussed here.

2 Methodology

The paper data was retrieved from Science Citation Index Expanded (SCIE) database which is one of the Web of Science databases and downloaded on Feb 10th, 2011. Science Citation Index Expanded is a multidisciplinary index to the journal literature of the sciences. It fully indexes over 6,650 major journals across 150 scientific disciplines. SCIE is the well known database which can be use for analysis of scientific and technical papers. There are great differences in different types of papers. This study limits papers' type: Article, Review, Letter and Editorial materials. These articles are more complete description and have full bibliographic entry. 4G and related keywords are indexed in the field of topic which includes abstract, title and keywords. The confirmation of keywords is based on the suggestions of related experts.

The patent data were retrieved from Derwent Innovation Index (DII) database which is also one of the Web of Science databases and downloaded on Feb 10th, 2011. DII database is a professional patent database, containing patent data of all major powers. The data related to 4G technology covers from 1996 to 2010. Patent classification is another retrieval means besides keyword. Some Derwent classification codes related to 4G technologies are combined with keywords during this study to make the retrieval result more integrated and veracious.

This study limits basic patents as basic data. The basic patent is the first member of a Derwent patent family. In other words, a patent family includes a basic patent, which ordinarily is the first patent issued on an invention, and subsequent or equivalent patents, which are issued by

other authorities for the same invention. So the basic patent can best reflect the development of technology.

3 Results and Discussion

3.1 Overview of the Output of Papers and Patents

The number of SCIE papers in the field of 4G technology in 2010 was 1772 and the number of patents was 2320. As shown in Fig 1, the number of papers and patens was low and stable from 1996 to 2001. It can be seen from the chart that a very noticeable trend from 2002 was the dramatic increase. After 2001, 3G technology began to be used in the commercial field in a large scale, so more and more researchers began to pay attention to Next Generation Network technology. ITU launched the 4G initiative with its strategic future vision in 2002. That meant that 4G technology industry was put on the agenda. After 2003, the number of patents began to increase faster than that of papers.4G technology began to enter the period of rapid growth. The growth rate of basic patents in 2010 was 15.5%, the growth rate of papers was 11.4%.

It can be seen from Fig 2, the number of patent assignees was increasing steadily from 2002. By the end of 2007, the number of patent assignees reached the peak of 1127. The assignees of patents began to be centralized from 2008. As the same time, the number of patents was rising year by year. It indicted 4G technology was changing from growing period to mature period. The control force of some organizations with technology strength was growing. From the current trend, more and more patents will be applied by international telecommunications giants, whereas small companies would withdraw from this field gradually.

2010 was an important year for 4G technology, with ever-growing operator commitments to deploy LTE networks, more commercial launches, and the future roadmap of LTE-Advanced gaining official approval from the ITU as an IMT-Advanced (4G) technology. The pace of development for 4G technology and network deployments are expected to speed up in 2011.

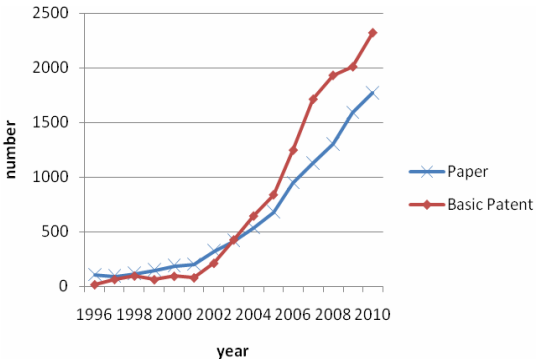


Figure 1. Papers and Patents of 4G Technology by Year in the World

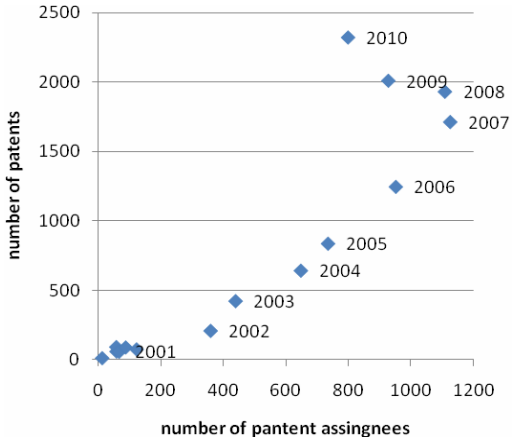


Figure 2. Life Cycle of 4G Technology

3.2 Outputs of Different Countries (Regions)

The United States as a technological power is still being the leading position in terms of papers and patents output. The number of papers of the United States and China are far more than the others in Tab 1. The paper's growth rate of the United States and China was also higher than the others. As shown in Fig 3, China's 4G papers were rising quickly, especially after 2008. In 2010, the number of papers of China had surpassed the United States. The researchers of South Korea, Japan and Taiwan are also very prolific.

In order to secure the future market better, Patent Assignees from all over the world would like to choose attractive countries (regions) to apply their patents. So from the Tab 1, we can conclude that the United States was the most attractive market of 4G which had a large number of patents. Its patent number was more than double, compared with No 2 –China. Besides, Japan and South Korea are also very attractive.

3.3 Outputs of Different Organizations

Table 1. Top Countries (regions) in the Number of Papers and Basic Patents

Countries(regions)	Papers	Basic patents
USA	1763	3651
PEOPLES R CHINA	1390	1456
South Korea	946	782
JAPAN	696	1633
TAIWAN	544	48
U.K.	537	176
CANADA	529	11
Germany	335	97
FRANCE	280	88
Australia	279	3

978-1-935068-76-1 © 2011 SciRes.

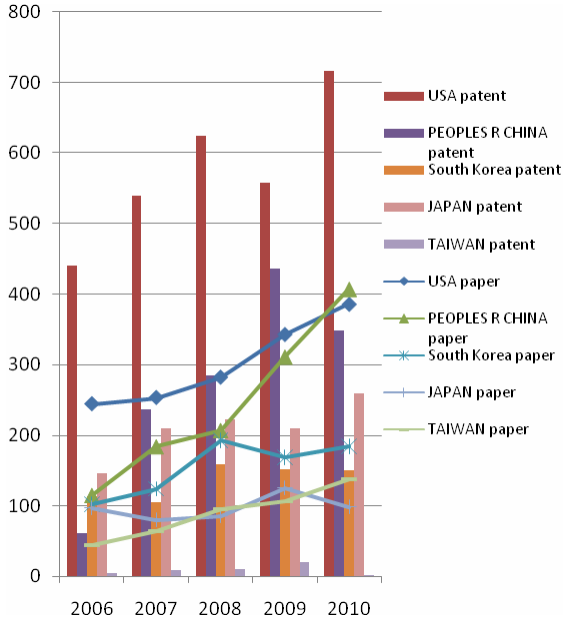


Figure 3. Top Countries (regions) in the Number of Papers and patents by Year

Usually, the first author of a paper is the main contributor for one research. So we choose the first author's affiliation as an indicator. The affiliations are scattered relatively. From the statistics given in Tab 2, it can be seen that among the top 10 institutes, 5 are from Peoples R China and 3 are from South Korea. Tsing Hua University is the No.1. The total paper number of Tsing Hua University was 162.

Parts of researches depended on the support of governments or other organizations. Tab 3 is the top 10 funding organizations. The number of papers from funding of National Natural Science Foundation of China is dramatically higher than the others

Cited number of patents can represent one company's influence in some ways. Ranked by the number of cited patents, we got the Tab 4. Top 3 are Samsung from South Korea, Matsushita from Japan and Qualcomm from the United States.

Japan values the output of Patents. Japanese companies apply a large number of patents in order to fight for the future market. As is shown by Tab 5, there are 5 Japanese companies in the Top 10 assignees counted by the number of basic patents. Samsung has the obvious advantage in the number of basic patents. But most of them were applied in homeland- South Korea. Qualcomm, Matsushita and LG had more patents in World Intellectual Property Organization (WPIO) that indicates they place more importance on the international markets.

The rapid growth of the number of patents didn't let Samsung avoid infringement action. After 2008, the number of patents of Samsung began to decrease. Samsung adjusted its 4G patent strategy and emphasized to buy patent rights from other organizations. Last year, Samsung agreed to pay Qualcomm $1.3 billion for the use of

Table 2. Top 10 Institutes in the Number of Papers (by the first author)

Rank	Organization of the first author	Number of papers
1	Tsing Hua Univ	162
2	Yonsei Univ	124
3	Sejong Univ	110
4	Shanghai Jiao Tong Univ	101
5	Xidian Univ	96
6	Beijing Univ Posts & Telecommun	93
7	Hong Kong Univ Sci & Technol	92
8	Natl Chiao Tung Univ	90
9	Korea Adv Inst Sci & Technol	82
10	Univ Southampton	78

Table 3. Top 10 Funding Organizaitons in the Number of Papers

Rank	Funding Organization	Number of papers
1	National Natural Science Foundation of China	397
2	National Science Foundation	232
3	National Science Council, Taiwan	160
4	National High Technology Research and Development Program of China (863)	159
5	MKE (Ministry of Knowledge Economy), Korea	132
6	European Community	125
7	National Basic Research Program of China	113
8	Natural Sciences and Engineering Research Council (NSERC) of Canada	110
9	Hong Kong Research Grant Council	54
10	Office of Naval Research, USA	54

Table 4. Top 10 Patent Assignees in the Number of Cited Patents

Rank	Patent Assignee	Cited Number
1	SAMSUNG ELECTRONICS CO LTD	2651
2	MATSUSHITA ELECTRIC IND CO LTD	2555
3	QUALCOMM INC	1864
4	SONY CORP	1799
5	TELEFONAKTIEBOLAGET ERICSSON L M	1716
6	MOTOROLA INC	1763
7	NIPPON TELEGRAPH & TELEPHONE CORP	1471
8	TOSHIBA KK	1665
9	NOKIA CORP	1685
10	NEC CORP	1264

Table 5. Top 10 Patent Assignees in the Number of Basic Patents

Rank	Patent Assignee	Number of basic patents
1	SAMSUNG ELECTRONICS CO LTD	1246
2	QUALCOMM INC	563
3	MATSUSHITA ELECTRIC IND CO LTD	518
4	LG ELECTRONICS INC	493
5	ELECTRONICS&TELECOM RES INST	367
6	ZTE CORP	349
7	NIPPON TELEGRAPH & TELEPHONE CORP	330
8	SONY CORP	306
9	TOSHIBA KK	286
10	HUAWEI TECHNOLOGIES CO LTD	261

certain 3G and 4G patents over the next 15 years [8]. Samsung will pay not only an up-front license fee, but also additional royalties. Samsung is also reported to have paid some of the lowest rates due to massive volume. According to Telecoms Korea, Samsung and LG represent a big profit source for Qualcomm.

3.4 Key Technology Analysis

As shown in Tab 6, OFDM technology apparently held the leading position among technology focuses in patent applications, with the number of 8925 basic patents.

It is essential to most of 4G candidate systems. Leading companies, such as Samsung, Qualcomm, Panasonic, etc., have devoted enormous resources to R&D of OFDM. As a result, the number of patents in this field increased dramatically from 2001 to 2007. However, the number of patent application had stopped rising since 2008. This may means that this technology becomes mature. Samsung takes the lead in the number of OFDM patents, while Qualcomm has the most OFDMA patents.

Table 6. Technology Foucus of 4G by Number of Basic Patents

Technology focus	number of basic patents
OFDM	8925
MIMO	2628
Broadband and modulated carrier systems using multi-frequency codes	2136
Radio link for data networks	1731
FDM	1454
Cellular telephone system	1217
Claimed software products	1075
Error detection and prevention by diversity, repeating or returning	1032
Portable mobile radio telephone	763
Transformation function of data processing systems	744

MIMO technology is also attached great importance. The number of patents in this field grew rapidly in recent years. MIMO technology has attracted attention in wireless communications, because it offers significant increases in data throughput and link range without additional bandwidth or transmit power. It achieves this by higher spectral efficiency (more bits per second per hertz of bandwidth) and link reliability or diversity (reduced fading). Because of these properties, MIMO is an important part of modern wireless communication standards. It can be seen that MIMO technology will be a growing technology focus in the next few years.

Besides OFDM and MIMO, radio link for data networks is another growing technology focus. The number of patents in other technology hardly changed in recent years. From the view of number of patents in different technical fields, Samsung's technology focuses include OFDM, MIMO and broadband and modulated carrier systems using multi-frequency codes. Qualcomm's technology focuses include OFDM, MIMO and claimed software products.

4 Conclusion

4.1 4G Technology is in the Growth Phase

After 2002, the number of papers and patents in 4G field was growing rapidly. The rate of patent's growth was faster than the paper's. The United States is the most attractive market who has the highest number of patents. The markets of China, South Korea and Japan are also very attractive. 4G papers of China were rising greatly especially after 2008. As for papers, among Top 10 funding organizations, 4 are from China. The number of papers funded by the National Natural Science Foundation of China was 397. The researchers of Taiwan, U.K. and Canada are also very prolific.

4.2 High Production Organization is Relatively Centralized

The companies of South Korea and Japan emphasize the output of patens. The companies of these countries account for 7/10 in the Top 10 patent organizations. Samsung has the obvious advantage on the terms of number of patents. Compared with Samsung, Qualcomm, Matsushita and LG pay more attention on the layout of international markets. On the billboard, we can see 2 Chinese companies – ZTE and Huawei. Facing competitors with power, Samsung has to abandon the strategy of seeking patents quantity, turning to get advantageous position by cross license and etc. In the field of information technology, the quality of patens is more important than the quantity.

4.3 OFDM and MIMO are the Most Important Technology Focuses

978-1-935068-76-1 © 2011 SciRes.

Among all 4G technology focuses, OFDM has the most patents, with a great quantitative dominance over others. It is essential to most of 4G candidate systems. However, after years of rising, the number of patent application becomes stable since 2008. This may means that this technology becomes mature. MIMO technology has attracted attention in wireless communications. The number of patents in this field has a rapid growth in recent years. It can be seen that MIMO technology will be a growing technology focus in the next few years. Besides OFDM and MIMO, radio link for data networks is another growing technology focus. The number of patents in other technology is steady in recent years.

References

[1] ITU. ITU paves way for next-generation 4G mobile technologies http://www.itu.int/net/pressoffice/press_releases/2010/40.aspx, Oct 21, 2010, Feb 14, 2011.
[2] GSA. GSM/3G and LTE Market Update http://www.gsacom.com/gsm_3g/market_update.php4, March 3, 2011, March 14, 2011.
[3] D. R. Chen, H. W. Chang, M. H. Huang and F. C. Fu, "Core technologies and key industries in Taiwan from 1978-2002: A perspective from patent analysis," Scientometrics, vol. 64, pp. 31-53, 2005.
[4] J. C. Guan and H. Ying, "Patent-bibliometric analysis on the Chinese science-technology linkages," Scientometrics, vol. 72, pp. 403-425, 2007.
[5] I. D. Lacasa, H. Grupp and U. Schnoch, "Tracing technological change over long periods in Germany in chemicals using patent statistics," Scientometrics, vol. 57, pp. 175-195, 2003.
[6] Narin, F., & Olivastro, D. (1998) Linkage between patents and papers: and interim EPO/US comparison. Scientometrics, 41(1-2), pp. 51-59.
[7] Huang, M. H. & Chi, P. S. "Analysis of academic papers in the field of material science among universities and colleges in Chinese world". International Conference on Engineering Education (ICEE) 2007. Portugal: International Network for Engineering Education & Research.
[8] Chris W. Samsung takes the IPR battle to Nokia. http://www.rethink-wireless.com/2010/04/21/samsung-takes-ipr-battle-nokia.htm. April 21, 2010.

Scientific Research

Evaluation of Scientific Researcher's Academic Performance Based on Contribution Ratio

Jianyu TANG[1], Keping WANG[2], Jing GUO[3]

[1]*Archives Center of Shanghai Aircraft Design and Research Institute, Commercial Aircraft, Corporation of China, Ltd., Shanghai, China*

[2]*Institute of Scientific & Technical Information, Shandong University of Technology, Zibo, China*

[3]*Shanghai Jiao Tong University Library, Shanghai, China*

Email: jytang84@163.com, wangkeping007@163.com, jguo@lib.sjtu.edu.cn

Abstract: The contribution of authors are not differentiated in the tradition evaluation methods of scientific researcher's academic performance, which not only leading to unfair and unscientific results, but also resulting in bad effects. Firstly, the shortcomings of tradition methods, such as counterpart evaluation, published articles and cited frequency, impact index, H index and so on, are reviewed because of undifferentiating contribution ratio. A new evaluation method of scientific researcher's academic performance which is based on contribution ratio is brought forward in this paper. After that, the method is applied into the evaluation process of researchers in the civil aero-engine subject of China. The analysis results which are accurate and logic demonstrate the new method is applicable, logic and fair.

Keywords: scientific researchers; academic performance evaluation; published articles; cited frequency; contribution ratio

1 Introduction

The evaluation of academic performance for scientific researchers makes up the bulk of science and technology evaluation. Evaluating scientific researchers' academic performance objectively, fairly and equitably is the base and way to optimize efficiency of allotting science and technology resources and promote R&D activities and science developing. The premise of fully putting functions of academic performance evaluation into effect is setting up the right evaluation methods. Basic principles of appraisement method are scientific, logistic, objective, comparative and operational [1]. In recent years, especially since the *h* index having been put forward by Hirsch of America physical scientist in 2005 [2], the evaluation of academic performance for scientific researchers has become an hot spot in the area of scientometrics at home and abroad.

2 Literature Review

There is a large quantity of published research literatures concentrated on methods of academic performance evaluation for researchers in domestic and overseas. These methods of peer review, published articles, cited frequency (total cited frequency and average cited frequency per paper), influence factor, representative work, *h* index and so on are adopted frequently in practice. The application of these methods is based on one hypothesis actually, that is differentiating authors' order in papers, so the contribution ratio of every author is regard as the same and every author share the same and all of credit. The assumption is completely held only for papers of sole author, otherwise, it's invalid. The reason is contribution ration of each author in a paper is likely different. Normally, contribution ratio can be reflected according to the signature order, and more anterior of a author's rank, the greater contribution of the author. As a result, authors of a paper can't share the same amount of credit and honor. Furthermore, with the production of cooperation, any authors of a paper, especially the author who is nominal or signed posteriorly, couldn't share all credit of the paper. Otherwise, if such traditional methods are adopted to evaluate researchers' academic performance, it's impossible to get objective and fair appraisement results even with perfect process controlling.

For example, *A*, *B* and *C* are three researchers and published *20* papers by working together. *A* is the first, *B* and *C* are the second and the third authors in *15* papers, which are totally cited for *200* times. For the other 5 papers, which are cited for *100* times, *B* is the first, A and *C* are the second and third authors respectively. Under this condition, based on above traditional measures, the evaluation results of academic performance of *A*, *B* and *C* are exactly the same, i.e. the number of these three authors' published papers is separately *20*, and the cited frequency is identical *300*. Obviously, the results are unscientific. Because it cannot obtain equitable effects of three researchers' academic performance (they may even complain about the unfair results), and the credits of these *20* papers (the number of published papers and cited frequency) are repeatedly calculated by these authors. It's only a simple instance. In practical work, scientific researcher's cooperating in writing papers is much more complicated than this example. According to the statisti-

This paper is sponsored by open fund of ISTIC-THOMSON scientometics joint laboratory

978-1-935068-76-1

cal data of SCI issued in 1992 by Glanzel, more than 90% papers are finished in cooperative way, the average author number is *4.5* per paper, and the largest number of the paper's writers is *102* [3].

Overall, these appraisement methods of academic performance which disregard contribution ration could result in three negative effects at least: (1) unfair evaluation; (2)honor of a paper being double counted;(3)leading to a culture of reaping without sowing and cooperation bubble. In addition, with the growing size of coauthor papers [4,5], these bad effects become more serious. Therefore, it's very necessary to add the factor of considering contribution ratio in the practice of researchers' academic performance.

Up to now, there are very few literatures concerns the issue and no evaluation methods based on contribution ratio. In 2006, Professor Su Xinning in China firstly proposed a method of calculating researchers' published papers based on their signature sequence in papers when he analyzed the academic influence in library, information and documentation science [6]. In the method, each paper is assigned one score and an author could get one score if he/she finished the paper by own him/her. For the cooperated ones, every author gets his/her score according to his/her order in a paper. This idea should be more scientific and reasonable in computing published papers of researchers. Based on the *h* index, Academician Zhang Chunting in Academy of Sciences of China bought forward the *w* index in 2009 which is the weighted *h* index and is equal to the product of a paper's cited frequency and an author's weighted coefficient[7,8]. This is a method of considering weighted cited frequency based on an author's signature sequence in a paper. As a result, as long as the researchers' order of varied, the w index will be different, even if the same of researchers' total cited number and h index, thus helping to solving the problem of cited number bubble. However, the above two methods don't reflect researchers' academic performance because of the shortcoming of singularity indicator.

3 New Model Based on Contribution Ratio

The number of published papers, especially papers released in core journals, is a key indicator to measure researchers' achievements and activity. The cited frequency of papers is an important index to appraise researchers' Academic influence. The new model designed in the paper and showed below is based upon two indicators of published papers in core journals and cited frequency:

$$AP=PS*80\%+CS*20\% \quad (1)$$

Where:

AP: The score of researchers' academic performance.

PS: The score of researchers' published papers in core journals.

CS: The score of cited frequency of papers in core journals.

Besides:

$$PS=\sum_{j=1}^{p} pc_j \quad (2)$$

$$CS=\sum_{i=1}^{k} cc_i * c_i \quad (3)$$

Where:

P: The quantity of published papers in core journals of researchers.

pc_j: The contribution of a researcher in a paper of No.*j*.

k: The number of cited papers of a researcher.

cc_i: The contribution of a researcher in a cited paper of No.*i*.

c_i: The cited frequency of the paper of No.*i*.

It's very crucial to assign the values of pc_j and cc_i. The best and most accurate way is based upon the real contribution devoted by researchers which is non-operational in practice. In most cases, researchers' order in a paper is the order of their dedication. Therefore, the paper quoted the contribution assignment scheme according to researchers' order in a paper [6]. The following Table 1 shows that every paper's score is *1* and each researcher's dedication ratio is fixed on the basis of his/her sequence and the number of researchers in a paper. For example, if researcher A published three papers totally in key journals. He is sole author in one paper. For the other two papers, A is the second author and the number of authors is *3* and *4* respectively. As a result, researcher A's dedication in these three papers is *1*, *0.25* and *0.2* respectively. According to this method, two rules could be got: firstly, the larger number of researchers in a paper, the smaller contribution ratio of each researcher; secondly, more anterior of a researcher in a paper, more contribution devoted to the paper by him/her.

Table 1. Contribution Assignment Way[a]

		Order of an Researcher in Paper					
		first	*second*	*Third*	*Fourth*	*Fifth*	*sixth*
The Number of Researchers in a Paper	*One*	1					
	Two	0.6	0.4				
	Three	0.6	0.25	0.15			
	four	0.6	0.2	0.1	0.1		
	Five	0.6	0.1	0.1	0.1	0.1	
	six	0.6	0.1	0.1	0.1	0.05	0.05

a. Authors ranking behind the seventh are neglected because of too low score

On the basis of formula of (1), (2), (3) and Table 1, researchers' evaluation results of academic performance could be got easily. The main advantages of this new method are shown below.

1) Rationality: Because of adding the factor of contribution ratio makes the result more realistic.

2) Comparability: Not only the last indicator can be

compared, but also the sub-index of score of published papers and cited frequency. What the most important is that no same evaluation results are got even if the same quantity of published papers, cited frequency and *h* index.

3) Synthesis: The model combines the indicators of published papers and cited frequency and assigns corresponding weight according to importance so could reflect researchers' academic performance comprehensively.

4) Objectivity: Because the model is completely based on objective data which is not be effected by subjective activity in practice.

5) Easy implementation: The model applies simply computing formula and can be easily conducted. Besides, the evaluation results could be quickly got.

4 An Application Case

The aero-engine, which is regarded as the heart of a civil aircraft, is the machine frequently being operated long-term in the complicated and atrocious environment of high temperature, high pressure, high rotate speed and high burthen. It's a crucial factor to determine the airspeed, altitude, range, reliability, economic, comfort and environmental conservation of an aircraft. It also influences the developing and advancement of the aircraft in much degree [9]. Until now, there have only been four countries which are America, Britain, France and Russia respectively having complete capabilities and technologies to develop advanced civil turbofan engines independently in the world. The whole global civil aero-engine market is shared by three manufacturers of General Electric Company, Pratt & Whitney and Rolls-Royce PLC. From the beginning of repair, strypped-down, remodel, China masters the technologies to develop military engines independently now, but still falls behind greatly in the area of civil turbofan engines especially the turbofan engines of big bypass ratio. Researchers are one of crucial factors to accelerate the developing of civil aero-engine. Therefore, this paper tries to adapt the new model into the academic performance appraisement of researchers in the civil aero-engine subject.

CNKI is the largest full text database which updates constantly and collects Chinese journals in China. The paper selects research articles about civil aero-engine from 2000 to 2009 published on ten domestic key and professional journals as the data source. These journals which embodies civil aero-engine papers mainly and are Journal of Aerospace Power, Journal of Propulsion Technology, Acta Aeronautica ET Astronautica Sinica, Journal of Beijing University of Aeronautics and Astronautics, Journal of Nanjing University of Aeronautics and Astronautics, Journal of Northwestern Polytechnical University, Acta Aerodynamica Sinica,, Journal of Aeronautical Materials, Journal of Mechanical Engineering and Nondestructive Testing. The papers could be downloaded easily from academic literatures online publishing database of China in CNKI. The literatures retrieval lasts one month from July to August in 2010. After the process of a series of relative data compiling, the researchers' academic performance are analyzed easily by using the above model. Appraisement results are showed in Table 2.

On the basis of the results showed in Table 2, the following conclusions are listed below.

1) The new method gets very exact results and so could be adapted to compare academic performance among researchers.

2) Total score of a researcher is impacted greatly by his/her score of cited frequency, indicating the character of emphasizing the academic influence evaluation.

3) Two points enhance the rationality of the method further: From the aspect of title, 90% of those researchers have the senior tile and the others have the assistant senior title; Secondly, all of researchers come from professional and outstanding aero-engine institutes in China.

4) The model integrates merits of dynamic developing and cumulating because of activity of published papers' score and cumulation of cited score.

Table 2. Important Researchers of Civil Aero-engine in China

No.	Researchers	Published Papers Score	Cited Frequency Score	Total Score	Title	No.	Researchers	Published Papers Score	Cited Frequency Score	Total Score	Title
1	Chen Guo	5.35	312.75	66.83	Professor	26	Zhou Sheng	5.65	64.4	17.4	Professor
2	Fu Huimin	3.4	270.05	56.73	Professor	27	Zhou Zhengping	5.9	63.05	17.33	Professor
3	Sun Jianguo	15.9	146.35	41.99	Professor	28	Liu Bo	11.75	38.65	17.13	Professor
4	Zhang Jinzhou	13.4	135.15	37.75	Professor	29	Han Wanjin	5.55	59.7	16.38	Professor
5	Zhu Huiren	17.7	101.8	34.52	Professor	30	Li Ling	7.85	48.05	15.89	Professor
6	Yue Zhufeng	3.7	153.9	33.74	Professor	31	Zhao Jianxing	7.15	49.9	15.7	Professor
7	Qiao Weiyang	15.2	86.65	29.49	Professor	32	Li Huacong	4.15	61.75	15.67	Assistant Professor
8	Wen Weidong	4.1	127.05	28.69	Professor	33	Guo Yinqing	4.45	59.15	15.39	Professor
9	Wang	7.4	107.3	27.38	Professor	34	Wu Hu	11.35	30.75	15.23	Professor

978-1-935068-76-1 © 2011 SciRes.

	Zhongqi										
10	Huang Jin-quan	10.9	88.7	26.46	Professor	35	Ji Honghu	8.75	37.85	14.57	Professor
11	Wang Song-tao	6.5	101.5	25.5	Professor	36	Wang Zhanxue	8	39.15	14.23	Professor
12	Gao Deping	5.45	104.7	25.3	Professor	37	Xiang And-ing	6.7	43.4	14.04	Senior engineer
13	Feng Guotai	5.85	98.45	24.37	Professor	38	Liu Qianzhi	4.45	51.65	13.89	Assistant Professor
14	Liu Songling	10.2	75.6	23.28	Professor	39	Chu Wuli	6.4	38.35	12.79	Professor
15	Tao Zhi	15.25	46.8	21.56	Professor	40	Zhu Xingen	4.7	44.25	12.61	Professor
16	Zhou Zheng-gui	8.95	71.15	21.39	Professor	41	Xu Duchun	5.35	40.8	12.44	Senior engineer
17	Liao Minfu	4.95	86.1	21.18	Professor	42	Wang Xi	6.2	36.35	12.23	Professor
18	Ji Lucheng	5.05	83	20.64	Principle Investigator	43	Chen Fou	4.3	42.4	11.92	Professor
19	Xu Guoqiang	15.35	38.8	20.04	Professor	44	Wang Yan-rong	4.7	39.65	11.69	Professor
20	Song Yin-dong	5.7	74.85	19.53	Professor	45	Zhu Chang-sheng	4	41.55	11.51	Professor
21	Huang Xianghua	4.3	78	19.04	Professor	46	Lin Yuzhen	6.3	32.1	11.46	Professor
22	Ding Shuit-ing	14.8	34	18.64	Professor	47	Wang Qiang	5	37.15	11.43	Professor
23	Liu Baojie	5.7	69.75	18.51	Professor	48	Zhang Dalin	3.4	43.35	11.39	Professor
24	Li Qihan	4.2	75.1	18.38	Professor	49	Xu Quan-hong	5.4	35.1	11.34	Assistant Professor
25	Hu Jun	11.05	47.6	18.36	Professor	50	He Xiaomin	4.8	37.1	11.26	Professor

5 Conclusions

On the whole, the evaluation method of a scientific researcher's academic performance proposed in this article is rational, comparable, objective and easy to conduct. Based on this measure, equitable, objective and accurate evaluation results can be quickly obtained. Moreover, as each researcher's contribution is distinguished, it is effective to contain paper cooperation bubble (Since the nominal researcher is in the backward position, his/her score is lower correspondingly, and the effects on his/her totally academic performance are weaker.), and to avoid repeated calculation of scientific achievements. Meanwhile, this measure features dynamic expansibility and the nature of cumulation, so it cannot only evaluate experienced scientific researchers, but also appropriately appraise the developing potential of young ones, thus, it is useful to lead young researchers to do academic research.

One more point is needed to be mentioned. The same as other evaluation methods, the model introduced in this paper also has some defects. For example, the distribution of one's contribution perhaps is not exactly correspondent with the real condition (In China, in some cases, the second author's contribution is bigger than the first one's). Therefore, the authors of this paper expect that experts and scholars could provide better solution to contribution distribution scheme, and related educational organization and scientific research administration could take researchers' contribution into consideration in the process of evaluating his/her academic performance.

References

[1] Z.F. WANG and G. X. YANG, "Research on the academic evaluation function of sci-tech periodicals, " Science and Technology Management Research. Guangzhou, vol. 26, pp. 170-172, 2006.

[2] J. E. Hirsch, "An index to quantify an individual's scientific research output," PNAS. Washington, vol.102, pp. 16569- 16572, 2005.

[3] W. Glanzel and H. J. Czerwon, "A new methodological approach to bibliography coupling and its application to national, regional and institutional level," Scientometrics. Budapest, vol. 37, pp.1995-1996.

[4] Y. LIU and Q. CHANG, "Scientometrical measurement and evaluation on international collaboration of bas ic research in China," Journal of Management Sciences. Tianjing, vol. 4, pp. 64-70, 2001.

[5] H. Z. Zuckerman, "Patterns of name ordering among authors of scientific paper: A study of social symbolism and its ambiguity," American Journal of Sociology, Chicago, vol. 74, pp. 276-291, 1968.

[6] X. N. SU, "Report on Academic influence in Library, Information and Documentation Science (2000--2004), "Journal of the China Society for Scientific and Technical Information. Beijing, vol. 25, pp.131- 154, 2006.

[7] C. T. ZHANG, "A proposal for calculating weighted citations based on author rank," EMBO reports. Heidelberg, vol. 10, pp. 416-417, 2009.

[8] C. T. ZHANG, "How to Evaluate a Researcher's Academic Performance?—On the Citation- bubble Problem and Its Solution," Science & Technology Review. Beijing, p. 1, 2009.

[9] X. W. HAN, L. Y.CHEN and H. WU, "Countermeasure of Speeding the Aero-engine Development in China," Journal of Beijing University of Aeronautics and Astronautics (Social Sciences Edition). Beijing, vol. 16, pp. 37- 41, 2003.

Scientific Research

The Integrated Evaluation about the Academic Impact of Chinese Researchers in Clinical Medical Science Based on Citation Analysis

Bin ZHANG, Min WANG, Jian DU, Peiyang XU, Yeding CAO, Xiaoli TANG, Yanwu ZHANG, Xiaoting LIU, Yang LI

Institute of Medical information, Chinese Academy of Medical Sciences & Peking Union Medical College, Beijing, China

Email: xupeiyang@vip.163.com

Abstract: Objective: To provide reference for researcher evaluation, comprehensive assessment for academic impact of researchers in clinical medical science has been done by multiple citation indicators. **Methods**: A factor analysis model for composite academic impact scores has been constructed for 10 Clinical PhD supervisors from Chinese Academy of Medical Sciences Fu Wai Hospital with 7 indicators. And a correlation analysis for the performance of domestic and international academic impact has been done. **Results**: Two main factors F1 and F2 are extracted from the 7 indicators. The number of publications, total citations, h index and g index are larger load in the F1 main factor, cumulative impact factor is also the larger factor loading for the international academic influence, the average citations number and A index are the larger load on the main factors F2. Comprehensive evaluation scores of domestic academic impact and international academic impact showed moderate correlation. **Conclusion**: Citation analysis indicators can be classified into two types of factors, one is "total" and the other is "average", with the establishment of integrated academic impact assessment model. The clinical researchers with higher "average" indicator tend to have higher academic impact scores. The domestic academic impact and international academic influence should be assessed separately.

Keywords: Academic Impact; Chinese Researchers; factor analysis; Integrated Evaluation; Citation analysis

1 Introduction

Under the circumstance of health system reform, the importance of researcher evaluation in medical field becomes more prominent. Researcher evaluation is an effective method to promote health development, accelerate talent team construction [1]. The important task for the department of research management is how to evaluate researchers output objectively and impartially. Generally, there are two important methods, one is peer-review method, and the other is citation analysis [2]. Peer review is a generic term for a process of self-regulation by a profession or a process of evaluation involving qualified individuals within the relevant field, which plays an important role in research management, but there is some disadvantages, such as time consuming, much of the decision-making power rests in the hands of the experts and so on[3]. As an objective tool to measure the influence of scientific research output, citation indicators are widely applied in performance evaluation[4].For traditional indicators, such as total number of papers, total number of citations, average number of citations, total number of high impact papers and so on[5], which have many disadvantages.

In 2005, h index was suggested by Jorge E. Hirsch[6], a physicist in USA, which is an index that attempts to measure both the productivity and impact of the published work of a scientist or scholar, and many researches have done in its application [7,8], limitation, and modification [9,10].g index was proposed in 2006 by Leo Egghe, which is an index for quantifying scientific productivity based on publication record [11], especially for the high cited papers. In 2007, Chinese scientist Jin Bihui put forward A index, R index [12], which can overcome the shortcomings of the high sensitivity of h index for evaluating the literature less output but high cited. Since then, f index [13], e index[14], h-share index derivatives and other indicators were proposed. In spite of many indicators, none of which is perfect in researcher evaluation [15], therefore an integratd method to evaluate the academic scholar was given. In this study, a factor analysis model for composite academic impact scores has been constructed for 10 Clinical PhD supervisors with 7 indicators, which can give a reference to evaluate the individual's scientific research output.

2 Methods

2.1 Research Objects and Indicators

According to the principal compared researches in the same subject, 10 Clinical PhD supervisors from Chinese Academy of Medical Sciences Fu Wai Hospital were included. We use the ISI Web of Science database, and seven indicators were analyzed, which involve total number of papers, total number of citations, average number of citations, h index, g index, A index and cumulative impact factors.

978-1-935068-76-1

2.2 Data Collection and Indicators Calculation

The published international papers of 10 Clinical PhD supervisors were retrieved by ISI Web of Science database, and then download the data to the software of NoteExpress, which is reference management software. NoteExpress can be used to remove duplicate data, delete papers published by other authors with same name. After that, relevant data were stored in the EXCEL, which used to calculate the value of indicators.

2.2.1 Total Number of Papers

Using EXCEL, sum of the published papers of each Clinical PhD supervisors, named international published papers and abbreviated sN.

2.2.2 Total Number of Citations

We use the ISI Web of Science database, retrieve the number of citations for each paper, named $sCit_j$, then sum of $sCit_j$ of each clinical PhD supervisor by Excel, and the formal is $sCit = \sum_{j=1}^{aN} sCit_j$.

2.2.3 Average Number of Citations

Calculate the average number of citations of each clinical PhD supervisor, and the formal is $sCit_j / sN$.

2.2.4 The h Index

The h index of a researcher is the number of papers coauthored by the researcher with at least h citations each，and then calculated h index of each clinical PhD supervisor by Excel, named sh.

2.2.5 The g Index

Ranked a set of papers in decreasing order of the number of citations, the g-index is the (unique) largest number such that the top g articles received (together) at least g^2 citations. The g index is calculated based on the distribution of citations received by a given researcher's publications, named sg.

2.2.6 The A Index

The A-index provides an average number of citations for the papers included by calculating h index, the formal is $sA=\frac{1}{sh}\sum_{j=1}^{sh} sCit_j$.

2.2.7 Cumulative Impact Factors

For the researcher's papers were published in recent years, the total number of citations can't reflect the quality of individual scientific achievement. However, The Journal Impact Factor can reflect the average level of papers published in it, so add cumulative impact factors in researcher evaluation. The data of impact factors are from Journal Citation Report published by Thomson Reuters in 2010. Calculate the cumulative impact factors of each clinical PhD supervisor by Excel, and the formal is $CIF=\sum_{j=1}^{aN} CIF_j$.

2.3 Statistic Method

Factor analysis is a statistical method used to describe variability among observed variables in terms of a potentially lower number of unobserved variables called factors [16]. Factor analysis is a multivariate analysis to deal with a statistical method for dimensionality reduction. The original variables were classified into some groups based on the correlation between variables, the same group with higher correlation and different groups with low correlation [17].We used SPSS software to analyze the values of indicators.

3 Results and Discussion

3.1 Raw Data

Table 1 shows the raw data of scientific research output indicators of 10 clinical PhD supervisors.

Table 1. Scientific research output indicators of 10 clinical PhD supervisors

NO.	sN	sCit	sCit/ sN	sh	sA	sg	CIF
1	185	479	2.589	11	14.636	13	677.405
2	112	337	3.009	9	18.222	14	284.056
3	187	505	2.701	11	18.364	15	545.433
4	19	124	6.526	8	12.500	10	31.055
5	21	31	1.476	3	7.667	5	49.056
6	248	2455	9.899	25	63.360	44	1800.361
7	131	1362	10.397	18	60.389	35	703.204
8	84	179	2.131	7	12.571	10	201.216
9	38	213	5.605	5	32.800	14	103.046
10	127	458	3.606	11	22.273	17	547.505

updated：2010-10-9

3.2 Data Standardization and Suitability for Factor Analysis

A standard score indicates how many standard deviations an observation or datum is above or below the mean. It is a dimensionless quantity derived by subtracting the population mean from an individual raw score and then dividing the difference by the population standard deviation. This conversion process is called standardizing or normalizing.

The number of standard deviations from the mean is called the z-score and has mean 0 and standard deviation 1, the formula as follows,

$$x_{ij}^{*} = \frac{x_{ij} - \overline{x_1}}{\sigma_j} \quad (i=1,2,3,\ldots,10，\ j=1,2,3,\ldots,n)$$

$\overline{x_1} = \frac{1}{n}\sum_{j=1}^{n} x_{ij}$, $\sigma_j = \sqrt{\frac{1}{n-1}\sum_{j=1}^{n} x_{ij}}$, n mean the total number of indicators. Then get the normalized data, named x_1, x_2, x_3, ... x_7.

The test of KMO and Bartlett shows that the KMO value is 0.692, Bartlett's test of sphericity of the X2 value is 346.831 ($p < 0.05$), which means the data is suitable for factor analysis.

3.3 Determine the Main Factors

We use SPSS to analyze the raw data by factor analysis in table 2 and table 3. Table 3 shows the first two Eigen values of the cumulative contribution rate has reached 95%, which meet the dependent variable to explain the whole issue their own requirements. Maximum variation of variance (Varimax) loading matrix was performed for the implementation of orthogonal rotation. Factor rotation, the cumulative variance ratio did not change, only a redistribution of the various factors to explain the variance of the original variables, making it easier to explain the factors. Table 2 shows factor loading matrix after the shaft.

Table 2. Component Score Coefficient after Rotated

Indicators	Main factors	
	F1	*F2*
total number of papers	0.133	0.976
Cumulative impact factors	0.472	0.865
Total number of citations	0.706	0.692
The h Index	0.654	0.739
The g index	0.805	0.586
Average number of citations	0.975	0.141
The A Index	0.896	0.389

Table 2 shows, among the two extracted main factors, Average number of citations and the A Index in the load on the F1 large amount, these are the "Average amount "relevant indicators. And the number of papers, total number of citations, h index and g index 4 main factors on the main factors in the F2 a larger load capacity, and both with the "Total amount" relevant indicators, in fact, included in A index reflects the average index of papers Cited. Thus, bibliometric indicators from 7 to extract the two composite indicators obtained F1 and F2, the variance rate and the factor scores of each index coefficient matrix in Table 3.

Table 3. Extraction of the total variance decomposition of the main factors and factor scores of the system matrix

The factors	*Average amount of factor F1*	*Total amount factor F2*
Total Initial Eigenvalues	**5.878**	**0.931**
% of Initial Eigenvalues	**83.968**	**13.293**
Cumulative % of Initial Eigenvalues	**83.968**	**97.261**
The number of papers(X1)	-0.361	0.576
The total number of times cited(X2)	0.106	0.133
The Average number of the cited(X3)	0.514	-0.348
h Index(X4)	0.054	0.187
A Index(X5)	0.355	-0.151
g index(X6)	0.212	0.019
Cumulative impact factor(X7)	-0.111	0.351

3.4 Comprehensive Evaluation Model

Coefficients by the factor scores can be obtained scores of academic influence model:

$$F_1 = -0.361X_1 + 0.106X_2 + 0.514X_3 + 0.054X_4 + 0.355X_5 + 0.212X_6 - 0.111X_7$$

$$F_2 = 0.576X_1 + 0.133X_2 - 0.348X_3 + 0.187X_4 - 0.151X_5 + 0.019X_6 + 0.351X_7$$

Factor score model based on standardized data and the indicators you can calculate the factor scores Ph.D., and then the variance factor was weighted to calculate the 10 international academic influence doctoral comprehensive evaluation model is:

$$Z = 0.8397F_1 + 0.1329F_2$$

Calculated on this basis, 10 doctoral composite score of academic influence, and sort the results shown in Table 4 (this result is for reference only, since only the citation analysis based on the results of periodical literature, academic and practical impact of multiple factors Force will determine the gap.) Table 4 shows the influence of academic evaluation scores more by the "total" factor expression, the number and total published papers were cited more the number of doctoral comprehensive evaluation of the higher the score.

Table 4. 10 doctoral academic influence of the domestic influence and international academic evaluation results

No.	*Score*	*RANK*	*F1*	*F2*
1	38.214	3	-81.326	801.387
2	26.288	4	-24.771	354.308
3	-20.320	10	-62.839	244.133
4	19.193	6	13.182	61.133
5	23.330	5	-5.038	207.378
6	69.040	2	9.121	461.862
7	78.457	1	54.200	247.896
8	-12.078	8	-25.629	71.050
9	18.667	7	15.185	44.515
10	-19.522	9	-44.114	131.828

4 Conclusion

Compared with the peer-reviewed, bibliometrics method is more quantitative, objective, cost saving and less interference from man-made. Academic literature citation analysis as a way of measuring influential and effective,

978-1-935068-76-1 © 2011 SciRes.

has been widely used in scientific research and performance evaluation [4]. This measure of academic influence in the early indicators, based on research and preferred method of using factor analysis of the indicators of comprehensive academic impact of researchers conducted a study of the total citations, h index, g index, A index, impact factors and other indicators used in the Chinese academic influence of clinical research scientists. The results showed that the performance of these indicators can influence the performance of academic researchers. In 2009,Wei SP using h index as the main index combined with w index, total citations, the total number of the cited papers the number of derived 43 most influential scholars in the Chinese Society from 1998 to 2007[18], citation index that can assist the literature to carry out the actual peer-reviewed evaluation of talent.

For researcher's academic activities are complex, many indicators should be used for the academic impact assessment. Single indicators or evaluation of some indicators, although convenient, but they will lose a number of other useful information. This requires an integrated evaluation method. Statistical aspects of the comprehensive evaluation method is comprehensive, comparable and advantages of objectivity, for the evaluation of the relevance of a large object classification, reflecting the dependence of various types of evaluation object[19]. COMPREHENSIVE STUDIES applications such as in 2008 the principal component analysis of 10 of Shandong Province Outstanding academic leaders of academic health systems, comprehensive evaluation of the influence[20]. In this paper, factor analysis of the indicators cited by many as the "total amount" and "average amount" two factors, the evaluation model of the 10 doctoral academic impact evaluation showed that the number of published papers and total number of times cited More clinical researchers tend to have higher academic influence, reflecting the wide range of clinical research personnel to participate in academic exchanges, can expand their academic horizons and research, improve their academic influence.

It should be noted that the academic influence is a broader knowledge system under evaluation, which include academic influence the level of the commitment issue, participate in academic assessment of the situation, published work, leading academic status, students develop Quality, access to awards and patents circumstances. In this study, citation analysis is only based on the perspective of the evaluation indicators involved primarily with academic papers related to one aspect of this citation, is relative and limited, because of its statistical analysis of data from different sources, evaluation vary. The individual's scientific research evaluation should be based on objective and scientific principles. Other methods, such as researchers' cooperative network, the accuracy of citation analysis in academic impact of the issue needs further study.

References

[1] Cao Wei. On health personnel evaluation and Reform of [J]. Medicine and Society 2010; 23 (8): 44-46.
[2] Best of Zheng, Jie. A personal evaluation of academic influence of [J]. Journal of Science and Technology of China, 2007,18 (6): 957-960.
[3] Li Xiongwen. Talk about "peer review" the work of the advantages and disadvantages of science and technology awards [J]. China Science and Technology Award, 2000 (4) :22-23.
[4] Zhu Daming. Researchers based on citation impact evaluation of academic discussion [J]. Science and Technology Management Research, 2008, 28 (11): 86-87.
[5] Wang Lian. Scientometrics performance evaluation applied to the challenges of scientific research personnel [J]. Science of Science and Technology Management, 2007 (4) :165-168.
[6] Hirsch J E. An index to quantify an individual's scientific research output [J]. Proceedings of the National Academy of Sciences of the United States of America, 2005, 102 (46): 16569-16572.
[7] Qiu Ping, Miu Wenting. Bibliometrics talent evaluation in the new exploration application to _h _ _ for the method of index [J]. Evaluation and Management, 2007,5 (2) :1-5.
[8] Bornmann L, Mutz R, Daniel H. Are there better indices for evaluation purposes than the h index? A comparison of nine different variants of the h index using data from biomedicine [J]. Journal of the American Society for Information Science and Technology, 2008,59 (5): 30-837.
[9] Schreiber M. The influence of self-citation corrections and the fractionalised counting of multi-authored manuscripts on the Hirsch index [J]. ANNALEN DER PHYSIK, 2009,18 (9): 607-621.
[10] Jeang K T. H-index, mentoring-index, highly-cited and highly-accessed: how to evaluate scientists? [J]. Retrovirology, 2008,5:106.
[11] Egghe L. Theory and practise of the g-index [J]. Scientometrics, 2006,69 (1) :131-152.
[12] SCIENTIFIC AND, Rousseau Ronald. R index, AR Index: h index extensions of additional indicators [J]. Scientific observation, 2007,2 (3): 1-8.
[13] Ye Ying. F index and evaluation of the number of significance [J]. Information Science .2009, 27 (7): 965-967.
[14] Zhang CT. The e-index, complementing the h-index for excess citations [J]. PLoS One, 2009, 4 (5): e5429.
[15] Thompson DF, Callen EC, Nahata M C. New indices in scholarship assessment [J]. Am J Pharm Educ, 2009, 73 (6): 111.
[16] Yuanjun Peng. Scientometrics Advanced Course [M]. Beijing: Science and Technology Literature Press, 2010.
[17] Xiao-group. And application of modern statistical analysis [M]. Beijing: China Renmin University Press, 2003.
[18] Weishun Ping, Lu Qiu Li, INDUSTRIAL ENGINEERING. China Education Research Performance Evaluation research in the field of [J]. Open Education Research, 2009,15 (3) :73-79.
[19] Qiu Ping, Xiao-Wen Ting. Evaluation theory • theory • Practice [M]. Beijing: Science Press, 2010.
[20] COMPREHENSIVE STUDIES, Liu Yan, cold Rong, et al. Using principal component analysis of the health impact of scientific and technological personnel of academic evaluation of [J]. Chinese Medical Science Research Management, 2008,21 (5) :282-284, 294.

Scientific Research

An Overview of the Knowledge Discovery Approaches in Ancient China

Qi YANG

The School of Public Management, Tianjin University of Commerce, Tianjin, China, 300134

Email: yq228@163.com

Abstract: Knowledge discovery is not only the method of computer technology in knowledge management but also the critical function of human knowledge activities. Computer technology is essentially just one type of tool in knowledge discovery activities. In ancient China, the work of knowledge discovery was mainly charged by government. Special officials of the Imperial College and departments did the work of knowledge discovery including commenting the classics, collecting folk poems, and writing history books and reference books etc. The third part presents the knowledge discovery model in ancient China. The fourth part of this paper explores six types of knowledge discovery approach in ancient China including notes and comments; textology; analogy; the tabulation method; bibliography and editing reference book. The functions of six Ancient Chinese Knowledge Discovery Approaches are presented in the fifth part.

Keywords: Knowledge Discovery; Note & Comment; Textology; Analogy; Tabulation method; Bibliography; Editing Reference Books

1 Introduction

Knowledge discovery is a concept of the field of computer science that describes the process of automatically searching large volumes of data for patterns that can be considered knowledge about the data [1]. There are two main bodies of work in knowledge discovery. One is concentrated around applying machine learning and statistical analysis techniques towards automatic discovery of patterns in knowledge bases, and the other is concentrated around providing a user guided environment for exploration of data [2]. It is considered as nontrivial extraction of implicit, previously unknown, the branch of data mining: Knowledge Discovery in Databases (KDD). From another point of view, knowledge discovery is not only the method of computer technology in knowledge management but also the critical function of human knowledge activities. In essentially, computer technology is only one type of tool in knowledge discovery activities. This paper advocates that the activities of knowledge discovery have a long history, where there is knowledge, there is knowledge discovery, and knowledge discovery in database (KDD) is only one of the modern approaches of knowledge discovery.

Knowledge discovery is in essential the activities of creating new knowledge: upgrading the information to new knowledge, searching the existing knowledge and achieving clustering knowledge by modern computer technology, obtaining new knowledge by analogy, and other activities. Computer data analysis is nothing but the new method of knowledge discovery in new era. It is undoubted that human development is always be accompanied by knowledge discovery activities.

Humanities and Social Sciences project of Ministry of Education (*09YJC870019*)

The famous sinologist, John King Fairbank pointed that the China of AD 1200 was on the whole more advanced than Europe at the same period, and China had a vast amount of documents which was about half of all over the world before 1750, the year of the Industrial Revolution that began in England [3]. This paper will explore the approaches of knowledge discovery in ancient China, and find the value of Chinese traditional knowledge management methods.

2 The Knowledge Discovery Activities in Ancient China

In ancient China, the work of knowledge discovery was charged by government. There are special officials to do the work of knowledge recording, knowledge organizing, knowledge converting etc., including historian official, education official and the officials of the Imperial College (it was named by Taixue in HAN Dynasty and Guozijian from Tang Dynasty to Qing Dynasty). The officials had two kinds of identities: official and scholar. They were responsible for not only knowledge recording, organizing and innovation, but also public administration. Their knowledge discovery activities included commenting the classics, collecting folk poems, and writing history books and reference books etc.

2.1 Commenting the Classics

There are six ancient classics in the *Spring and Autumn Periods*. Confucius was the first person to use Six Classics as teaching materials in private education. These are the *Yi* or *Book of Changes*, the *Shi* or *Book of Odes* (or *Poetry*), Shu or Book of History, the Li or Rituals or Rites,

978-1-935068-76-1

the *Yue* or *Music* (no longer preserved as a separated work), and the *Chunqiu* or *Spring and Autumn Annals*, a chronicle history of Confucius' state of Lu extending from 722 to 479 B.C.. Because the Six Classics had existed before the time of Confucius, they were difficult to study and need to be commented. Confucius said that his mission was not innovating but explaining the Classics. However, as a matter of fact, Confucius and his students had established the Confucian knowledge system during the process of commenting and editing. With the growing status of the Confucian, the Six Classics were gradually expanded into Thirteen Classics in later centuries. Commenting the classics was the main approach to formulate new knowledge in ancient China. Hence, there large number of classics comment books in Chinese literatures.

2.2 Collecting Folk Poems

Collecting folk poems is the specialized work of the officials who were responsible for public opinion and knowledge. In ancient China, the official who charged in collecting folk poems was called messenger from central government or the official from far away or messenger by carriage. On the surface, their mission was collecting folk poems, just like their official title, but actually, their work is to get information and knowledge from the bottom of the society, and gather it into empire knowledge system, and deliver it to government. Because the ancient people usually criticized or praised government by poems in China, all the poems can be sung by local people, and the government appointed officials of collecting folk poems was to ensure the government to get the accurate information and governance knowledge of local people. In the *Book of Poetry*, the poetry of 2500 years ago, we can find large number of poems that focus on the governance knowledge and ethic knowledge.

From the modern perspective, poem is a form of literary art. But The *Book of Poetry,* as the oldest known purely classic in Chinese history, is different from modern poetry. It provided not only the first extant examples of narrative and emotion-expressing verse and rhyme in Han Chinese history, but also the principles in composition and rhyme. Most importantly, it recorded the knowledge of people's daily life and people's opinion of governance. The Odes deeply influenced political dealings in ancient China. When kingdoms or feudal leaders wished to express delicate or difficult positions, they would sometimes couch the knowledge within a poem. This practice became common among educated Chinese in their personal correspondences.

Modern scholarship on the *Book of Poetry* often focuses on doing linguistic reconstruction and research in Old Chinese by analyzing the rhyme schemes in the Odes, which show vast differences when read in modern Mandarin Chinese. The scholars seldom research the activities of collecting folk poems as a typical function of knowledge discovery.

2.3 Writing History Books and Editing Reference Books

Writing history books is the great tradition of China. There were two kinds of historian officials, the left historian record the king's speech, and the right historian record the king's action. All the records should be saved in the central archives. The historian officials also wrote history books according large amount of history records of the central archives. These activities can save valuable information and knowledge. Zhang Xuecheng, the famous thinker of Qing Dynasty, said that all of the Six Classics were history books. Why did history books play such a big role in ancient China? We should find reasons from knowledge management angle. The process of writing history book included several special activities of knowledge management, such as knowledge recording, knowledge organizing, knowledge analysis, and knowledge commenting, etc.

With the development of Chinese ancient culture, large numbers of books were published, how to get the knowledge rapidly in the large amount knowledge products has become the critical problem to the ancient Chinese people. In order to solve this problem, Chinese ancient scholar edited the reference books according to some special classifications.

Currently, people seldom research ancient Chinese knowledge management activities, and there still exists the suspicions on these activities. Hence, next part of this paper will discuss the model of Chinese ancient knowledge discovery.

3 The Model of Knowledge Discovery in Ancient China

In modern knowledge discovery activities, people can discover the potential and significant knowledge through computer analysis tools from database. In ancient society, there was no computer, the ancient people cannot get new knowledge through modern tools, but they could not stop the activities of knowledge discovery. China has been a huge country for thousands of years, and the governance of huge country inevitably needs the support of knowledge management. That's why ancient China has vast amount documents and firstly invented the art of paper-making and printing thousand years ago. In particularly, people seldom research the ancient China's knowledge discovery systems and many people don't know this field. According to ancient Chinese documents and knowledge management systems, we can describe the model of Chinese ancient knowledge discovery (Fig.1). This model includes four parts as follows:

- **Collecting data**: collect the documents and establish archives system
- **Data standardization**: establish knowledge standard system and texts system by knowledge management tools such as note and comment, textol-

ogy, analogy, bibliography etc.

- **Establish huge database**: edit huge reference books according some topics
- **Data mining**: users obtain or discover new knowledge by the huge reference books

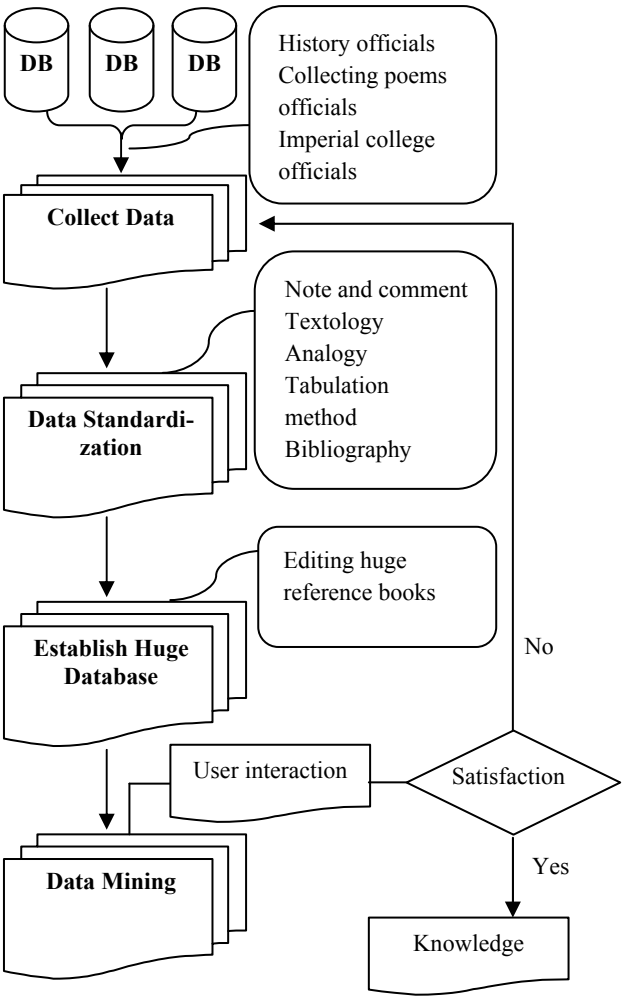


Figure 1. The model of knowledge discovery in ancient China

In additionally, what we need to notice is the step of the data mining. The concept of data mining in ancient China is different from modern knowledge discovery, and it is inefficient compared with modern knowledge discovery. But in that period of time, it was the advanced technology all over the world. The Chinese huge reference books can be utilized in data mining of Chinese ancient knowledge discovery, and Chinese ancient government attached importance to this tool just like modern computer database.

4 The Approaches of Knowledge Discovery in Ancient China

Before considering the approaches of knowledge discovery in ancient China, we should notice the fact that China is a vast country and has large amount of literatures and massive total population. How to achieve good governance? Qin Shihuang, the first emperor of Qin Dynasty, decided to fire Confucian books and want to achieve governance by the theory of Legalist School. Chinese ancient scholars considered that it was the main reason that led to the empire perished. Zhang Jie, the poet of Tang Dynasty, wrote a poem to criticize Qin Shihuang. He said, the empire was weakened in the smoke of firing books; the defense facilities only protected the emperor's house; the ashes was still warming, while the uprising happened in east side of Xiao Shan Mountain; yet the leaders of uprising, Liu Bang and Xiang Yu, did not read the books. This poem indicated that knowledge can make the country stability and prosperity, and if the government destroyed the documents abandoned knowledge, the empire would be faced with social unrest and war. After Qin Dynasty, almost all the Chinese government attached great importance to documents and knowledge discovery. Chinese ancient intellectuals explored many efficiency methods to discover new knowledge and establish Chinese knowledge system. These approaches include note and comment, textology, analogy, tabulation method, bibliography, editing reference books, etc.

4.1 Note and Comment

Recording the important speeches and events was the tradition of ancient China, and China has vast amount of history documents. Confucius had edited *Six Classics* and used them to teach students. But several hundred years later, people could not understand these texts. The successors of Confucius began to note and comment these classics. Han Dynasty was the first important period of commenting classics. The most famous scholar was Zheng Xuan, and it was said that he had commented all the classics. During the process of note and comment work, the scholars should note the pronunciation, the structure and meaning of characters, and they also should point the allusions of some special phrases, analyze the implication of the whole article according the social background of the documents.

The second period of comment classics was from Tang Dynasty to Song Dynasty. During this period, there were many famous scholars who commented the classics, such as Kong Yingda, Yan Shigu, Jia Gongyan, Xing Bing, Zhu Xi, etc. Ruan Yuan, the most famous scholar of late Qing Dynasty, had chaired the enormous project of collating the Collection of Notes and Comments of Thirteen Classics and published these books. It was one of the most important achievements of Chinese traditional knowledge management. According to statistics of Qian Taiji, the scholar of Qing Dynasty, this collection books had 416 volumes and the main body has 647,500 characters (does not include the characters of notes and com-

978-1-935068-76-1 © 2011 SciRes.

ments) [4]. In 2004, the full-text searchable database of *the Notes and Comments of Thirteen Classics* was provided by Beijing Guoxue Times Culture Co. Ltd., and it has approximately 10,000,000 characters (the sum of text, notes and comments).

The objective of note and comment approach is to give the critical interpretation of ancient texts. There are several methods as follows:

- to explain the ancient character by its synonym (Huxun)
- to explain the ancient character by the characters that having the same phone, same meaning and different strokes (Shengxun)
- to explain the ancient character by the character strokes and font (Xingxun)
- to explain the ancient character by the meaning of contemporary characters (Yixun)
- to explain the ancient character by its antonym (Fanxun, some characters' ancient meaning is opposite to their modern meaning.)
- to explain the ancient character by using several related characters (Dixun)

This approach could ensure how to understand the character's right meaning, and promote the knowledge qualities during the process of ancient knowledge diffusion.

4.2 Textology

The approach of textology was formed in the late of Ming Dynasty and was popular in Qianlong and Jiaqing period of Qing Dynasty. It was also named textual criticism or Pu Xue (simple and plain learning). It was used in almost all the Chinese tradition academic area, including classics subject, philology, phonology, note and comment subject, history, geography, astronomy, bibliography, etc.

Seeking the truth from facts is the principle of Chinese ancient textology. There are several methods as follows:

- to collate the ancient text for all the different edition books
- to seek the truth by using the materials of archaeological discoveries
- to compare the texts and find the illogical records
- to collect the materials of the ancient book that had been lost, and to collect these materials from the quotations in the other ancient books)
- to prove the issues by induction
- to prove the issues by the method of comparison
- to integrate all the knowledge from different fields

The approach of textology was the most important tool to seek the truth. It had a profound effect on Chinese academic researches.

4.3 Analogy

Analogy is a cognitive process of transferring information or meaning from a particular subject (the analogue or source) to another particular subject (the target), and a linguistic expression corresponding to such a process. In a narrower sense, analogy is an inference or an argument from one particular to another particular, as opposed to deduction, induction, and abduction, where at least one of the premises or the conclusion is general [5]. Analogy is the most commonly used method in ancient China. Confucius said that drawing the analogy from the simple idea and knowledge was the critical method to help the individuals accept benevolence value. Analogy was applied in almost all of the fields of Chinese academic researches. There are three kinds of analogies in ancient China [6].

The first is having analogy between the similar things. The second is having comprehensive analogies by using several similar relationships. The third is having analogies of attributes between different classifications.

There were lots of analogy examples in ancient Chinese documents. In the Book of Odes, the ancient poets used the method of Bi Xing (Bi: to compare this thing with that thing as an analogy; Xing: to associate with something in thinking by facing this thing), and formulated new social knowledge from the phenomenon of nature. Kong Anguo said that Bi Xing was the method to get knowledge of category through analogy (Yin Pi Da Lei).

4.4 Tabulation Method

The approach of tabulation method was invented by Sima Qian, the famous historian of Han Dynasty. He gathered the historical data by the Ten Forms in generation order or chronological order or monthly order in the Book of Historical Records. In ancient China, the historian's mission was to record the historical facts and help the readers more easily to discovery new knowledge from his history books. Sima Qian's *Historical Records* was the first huge historical book that recorded the history of 3000 years in ancient China. The vast amount history documents were classified suitably in all the biographies, but some kinds of knowledge could not be found in biographies. Hence, Sima Qian collected all the facts and put them into the forms in some kind of order according the materials. According to the actual situation of history materials, he created the generation table of three Dynasties, the chronological table of twelve vassal states, the menology table of period of Qin and Chu, and other chronologies. As the result of this approach, the new knowledge of social events is formulated by the comparison of events in one table. It is not only the outline of history knowledge but also the new knowledge of social value. Until now, Chinese scholars have always used the tabulation method to get new knowledge from the complex materials.

4.5 Bibliography

Bibliography is the academic study of books as physical, cultural objects, it is not concerned with the literary content of books, but rather how the books were designed,

edited, printed, circulated, reprinted and collected [7]. It is a systematic list of books and other works such as journal articles. In ancient China, with the increasing number of books, the scholars of Han Dynasty met the problems that how to manage the large amount of books. Liu Xin and Liu Xiang, the son and father who lived in AD 1 century, were the first bibliographers in China. They wrote the books of *Bie Lu* and *Qi Lue*, which were the earliest bibliography works. There are lots of bibliography works in ancient China. The largest and most completed bibliography work is *the Synopsis of Si Ku Quan Shu*. Because the bibliographic works can provide the categories of books, pictures and the other documents, the ancient Chinese scholars were much emphasized in this method and established Chinese special bibliography.

The approach of bibliography is the important tool to help the scholars to check the literatures and discovery special knowledge.

4.6 Editing Reference Books

The reference book is the Chinese special research tool, and it is different from modern encyclopedia. It was named Lei Shu (the books' summary of some categories), and it is reference book with material taken from various sources and arranged according to subjects. The first reference book named *Huang Lan* was edited by Liu Shao and Wang Xiang in Three Kingdoms period of Chinese history. This reference book has 8,000,000 characters according to the book of *Wei Lue*. It was lost in later centuries and was only indexed by the book of Qi Lue. After that period of time, the ancient Chinese government had appointed the scholars to edit the large reference books, such as the book of *Yi Wen Lei Ju* (the Collection of the Category of Art and Literature, by Ouyang Xun, Tang Dynasty), the book of *Yun Fu Qun Yu* (the Reference Book Rhyme Characters, by Yin Shifu, Yuan Dynasty), *the Large Book of Yongle* (the encyclopedia of Yongle, Ming Dynasty), *the Book of Pei Wen Yun Fu* (the Reference Book Rhyme and Characters, Qing Dynasty), and *the Books Collection From Past to Now* (Qing Dynasty), etc. Among all of these large books, *the Large Book of Yongle* is the largest reference book, and it has 0.37 billion characters and the index of this book has 60 volumes.

The reference book was the huge literature database that edited according classifications. It was used to research a kind of knowledge and the scholar can obtain the knowledge in short time. In a sense, the ancient Chinese reference book is the big knowledge database, and it is provided to the scholars discovering the knowledge.

5 The Functions of Ancient Chinese Knowledge Discovery Approaches

In ancient China, the different approach has different function, and the scholar can select the method to research the documents. They are described in Table 1. The function of notes and comments is to help the readers to understand ancient characters; the method of textology is to ensure delivering the accurate meaning of text; the approach of analogy is to convert the information to knowledge and get new knowledge by analogies; the tabulation method is to compare the complex documents and outline the events in the time order and help the readers to discovery new knowledge in history records; the method bibliography is to help the scholar search the books rapidly; the method of editing reference book is to help scholar search for the text rapidly. All of these approaches are invented by ancient Chinese scholars and had been formulated Chinese characteristics academic approach system.

Table 1. The function of the knowledge discovery approaches in ancient China

The approaches	*The functions*
notes and comments	to help the readers to understand ancient characters
textology	to ensure delivering the accurate meaning of text
analogy	to convert the information to knowledge and get new knowledge
the tabulation method	is to compare the complex documents and outline the events in the time order and help the readers to discovery new knowledge in history records
bibliography	to help the scholar search the books rapidly
editing reference book	to help scholar search for the text rapidly

6 Conclusion

China is an ancient country, it has vast amount documents. How to manage the knowledge effectively? How to discovery new knowledge? Almost all the Dynasties should face these problems and seek the solutions. As the main knowledge management tools, the six approaches were explored by ancient Chinese people and played important role in the development of China. The ancient Chinese government utilized the country's financial, material and human resources to establish the huge national archives, special reference books, intellectual development systems, and ordered the scholar officials to research the documents and practices, ensured the formulation and implementation of the national knowledge innovation system. There is no doubt that the ancient Chinese people who making the knowledge strategies, knowledge management activities and knowledge discovery approaches were the forerunners of government knowledge management.

References

[1] Frawley William. F. et al. (1992), "Knowledge Discovery in Databases: An Overview", AI Magazine (Vol 13, No 3), 57-70.

[2] Ronen Feldman and Ido Dagan, Knowledge Discovery in

978-1-935068-76-1 © 2011 SciRes.

Textual Databases (KDT). KDD-95 Proceedings. 1995, pp.112–117. http://www.aaai.org/Papers/KDD/1995/KDD95-012.pdf
[3] John King Fairbank and Merle Goldman, "China, A New History", second enlarged edition, the belknap press of Havard University Press, Cambridge, Massachusetts, London, England, 2006, pp. 2.
[4] Enbao WANG. "The number of volumes and words of the Collection of Notes and Comments of Thirteen Classics". Literatures, 1982, vol.2. pp.82 (Chinese).
[5] http://en.wikipedia.org/wiki/Analogy
[6] Qi YANG, "To Be Able to Get the Figures from the Nigh Things: The Basic Moral Education Method of Confucius", Journal of Socialist Theory Guide. 2008,Vol.2, pp.113-115. (Chinese).
[7] http://en.wikipedia.org/wiki/Bibliography

Scientific Research

Exploring Associations between Words in English and Chinese Sentences

Junsheng ZHANG, Wei YU, Huilin WANG

IT Support Center Institute of Scientific and Technical Information of China, Beijing, China

Email: zhangjs@istic.ac.cn, yuwei@istic.ac.cn, wanghl@istic.ac.cn

Abstract: Study associations among words can help distinguish the semantic roles of words and find the important words for information organization and retrieval. This paper studies co-occurrence and sequence associations between words based on a parallel sentence corpus in English and Chinese. The experimental association networks are constructed to study the characteristics from the complex network view. Experiment results show the difference between association networks in English and those in Chinese.

Keywords: association; co-occurrence; sequence; complex network; link analysis

1 Introduction

Short text based applications have become popular in recent years. Micro-blog is widely used to capture the feelings of people, and describe events happened in the daily life. An article of micro-blog may contain several sentences, a single sentence or even only several words. In scientific information service, abstract of literature only contains several sentences to represent the meaning of the whole article. To understand short texts, it is necessary to study the associations among words.

Semantic associations exist among words in the natural language. Statistical characteristics of words such as *sequence* and *co-occurrence* reflect the semantic associations between words. The statistical associations among words are studied in a *context*. Words and separators in the same text window together formulate a *context*. If text windows chosen for context have different size, the context is also different.

A sentence could be regarded as a minimum unit to represent a relatively complete meaning. Words and separators together formulate a sentence of natural language. Study the associations between words in the same sentence is important for the study on the semantic meanings of words and distinguishing semantic roles of words. Associations between words can be used to calculate the similarity between words, which is the basis of clustering sentences and short texts.

In this paper, we study two common statistical associations among words (i.e., *co-occurrence* and *sequence*) in sentences, that is, *the scope of context is limited to be in a sentence*. Study association networks of words can help distinguish the semantic roles of words in sentence level, and find the distributional rules of words. Associations between words can be used to cluster sentences. Words can be chosen from the association network for information organization. Besides, bilingual association networks of words can be used in machine translation.

This work was supported by ISTIC project Grant No. YY-2010018.

2 Related Work

Associations between words in a context can be represented by an association network, in which vertices represent words/separators in the sentence, while edges between vertices are associations between words in the sentence.

Words/separators and their co-occurrence associations together formulate *co-occurrence network*, while words/separators and their sequence associations together formulate *sequence network*.

To study the characteristics of association networks, many network measures should be defined first. The definitions of various network measures are based on graph $G=(V, E)$, where V and E are the vertex set and the edge set, respectively; $|V|=n$ and $|E|=m$ represent the number of vertices and edges, respectively. The commonly used network measures are defined as follows.

Degree: In an undirected graph, a vertex's degree is the number of edges incident to the vertex.

Self-loop: If the start vertex and end vertex of an edge are the same, the edge and the vertex formulate a *self-loop*. Self-loops are counted in degree metrics.

In-degree: In a directed graph, a vertex's in-degree is the number of incoming edges incident to the vertex.

Out-degree: In a directed graph, a vertex's out-degree is the number of outgoing edges incident to the vertex.

In an undirected graph, in-degree and out-degree are undefined and not computed. In a directed graph, degree is undefined and is not computed.

Links between vertices can be used to evaluate the vertices. The *authority* and *hub* reflect the in-degree and out-degree characteristics, respectively, of vertices in network. The idea of Hyperlink-Induced Topic Search (HITS) is that a good hub links many authorities while a good authority is linked by many good hubs [1]. Vertices with the highest authority/hub are centers. The authority and hub of a vertex are calculated by the following formula:

978-1-935068-76-1

$$\begin{cases} A(v_i) = \sum_{(v_j,v_i)\in E} H(v_j) \\ H(v_j) = \sum_{(v_i,v_j)\in E} A(v_i) \end{cases}$$

where $A(v_i)$ and $H(v_j)$ are the authority and hub of vertex v_i and v_j, respectively.

The *degree centrality* describes the degree information of each vertex [2,3] according to the idea that more important vertices are more active and therefore should have more connections. Degree centrality can be used to find the core vertices, but it only considers the hub characteristic and ignores the authority characteristic. The degree centrality for a vertex v is calculated as follows: $C_D(v)$=degree(v)/(n−1). Calculating the degree centrality for all vertices in a dense graph has a time complexity $O(n^2)$, which becomes $O(m)$ in a sparse graph.

The *betweenness centrality* describes the frequencies of vertices in the shortest paths between indirectly connected vertices [2-4]. It is based on the idea that if more vertices are connected via a vertex, then the vertex is more important. It can be used to find the edges between two communities in a complex network. The betweenness centrality for vertex v can be calculated by the following formula,

$$C_B(v) = \sum_{\substack{s\neq v\neq t\in v \\ s\neq t}} \frac{\sigma_{st}(v)/\sigma_{st}}{(n-1)(n-2)},$$

where σ_{st} is the number of the shortest geodesic paths from s to t, and σ_{st} (v) is the number of the shortest geodesic paths from s to t that pass through vertex v.

The shortest paths between each pair of vertices in a graph can be found by Floyd–Warshall algorithm with time complexity $O(n^3)$ [5], so the time complexity of the betweenness centrality also is $O(n^3)$. The betweenness centrality has been used to study the community structure of social and biological networks [6].

The *closeness centrality* describes the efficiency of information propagation from one vertex to the others [3, 7, 8]. Its idea is that a vertex is central if it can quickly reach others. The closeness centrality can be regarded as a measure of the time to spread information from a vertex to other reachable vertices in the network. The closeness centrality is defined as the mean geodesic distance (i.e., the shortest path) between a vertex v and all of the vertices reachable from v as follows, where $n\geqslant 2$ is the size of the connected component reachable from v. Calculating the closeness centrality for each vertex in the graph has time complexity $O(n^3)$.

$$C_C(v) = \frac{n-1}{\sum_{t\in V\setminus v} d_G(v,t)}.$$

The *eigenvector centrality* measures the importance of vertices according to the adjacent matrix of a connected graph [9,10]. It assigns relative scores to all vertices in the network based on the principle that connecting high-scored vertices contributes more to the score of a vertex than do connecting low-scored vertices. PageRank is a variant of the eigenvector centrality measure [11].

The information centrality describes vertices' influence on the network efficiency of information propagation [12].

$$E_G = \frac{\sum_{v_i\neq v_j\in G}\varepsilon_{ij}}{n(n-1)} = \frac{1}{n(n-1)}\sum_{v_i\neq v_j\in G}\frac{1}{d(v_i,v_j)},$$

where the efficiency ε_{ij} in the communication between vertices v_i and v_j is equal to the inverse of the length of the shortest path d(v_i, v_j).

The *information centrality* of a vertex v is the relative drop in the network efficiency caused by the removal of the edges incident with v from G defined by the following formula:

$$C_I(v) = \frac{\Delta E}{E} = \frac{E[G] - E[G_v']}{E},$$

where G_v' indicates the network resulting from removing the edges incident with vertex v from G. The information centrality has been used to study the structures of communities in complex networks [13].

Clustering coefficient: the ratio between the three times of the number of triangular in a network and the number of edge pairs sharing a common vertex [15].

$$CC = \frac{3 * number\ \ of\ \ triangles}{number\ \ of\ \ connected\ \ triples\ \ of\ \ vertices}.$$

Geodesic distance: the distance between two vertices in a graph is the number of edges in a shortest path connecting them.

Diameter: the diameter of a graph is the maximum eccentricity of any vertex in the graph. That is, it is the greatest distance between any pair of vertices.

Graph density: The density of a graph is the ratio of the number of edges and the number of edges in a complete graph.

3 Association Network of Words

In English and Chinese, a sentence often contains many words and separators. The number of words and separators in a context C is defined as the *length* of C, denoted by $|C|$. A Chinese sentence has to be segmented first. After segmentation, the chunks split by "*blank*" is called *words*. For example, given a sentence "This is a sentence." in English, and "这是一个句子。" in Chinese.

- "This is a sentence." becomes "This is a sentence ." after preprocessing. Chunks *This*, *is*, *a*, and *sentence* are words, while . is a separator.
- "这是一个句子。" becomes "这 是 一 个 句子 。" after segmentation. Chunks 这，是，一，个, 句子 are words, while 。 is a separator.

Co-occurrence and sequence are two common associations between words. Each pair of words/separators has two associations, that is, *co-occurrence* and *sequence*, if they occur in the same context.

Sequence has two types: *before* and *after*. Given two words x and y in the same context C, if the position of x is less than the position of y, then x is *before* than y, while y is *after* than x.

Sequence associations between words are *transitive*. For example, if x is *before* than y, y is *before* than z, then x is *before* than z. In this case, x and y has *direct sequence* association, while x and z has *indirect sequence* association. Accordingly, the sequence network is classified into two types: *direct sequence network* and *sequence network*.

Suppose $s = t_1, t_2, \ldots, t_n$ ($n \geq 1$), and t_i and t_j ($1 \leq i, j \leq n$, $i \neq j$) are two tokens, then the distance between t_i and t_j is defined by $d(t_i, t_j) = |i-j|$.

- Two words in a sentence have sequential association. If $i > j$ then t_i is *after* t_j or t_j is *before* t_i. All the associations sequence between t_i and t_j formulate the *sequence network* of s.
- Two words in a sentence s have co-occurrence association. Each pair t_i and t_j ($1 \leq i, j \leq n$, $i \neq j$) has co-occurrence association. All the co-occurrence associations formulate co-occurrence network of s.

Suppose a sentence s contains n words and m separators, then sentence s contains $m+n$ elements. The number of sequential relations between $m+n$ elements is

$$1 + 2 + \cdots + (\mathrm{m} + \mathrm{n} - 1) = \frac{(m+n-1)(m+n)}{2}.$$

So the concurrence network of a sentence contains $(m+n-1)(m+n)/2$ concurrence associations.

Association *co-occurrence* is position-independent, while association *sequence* is position-dependent. If two words/separators have *sequence* association, they have *co-occurrence* association; if two words/separators have *co-occurrence* association, they also have *sequence* association.

A word may occur two or more times in a sentence. *During the construction of association networks, same words with different positions in a sentence share the same identifier*. So cycles may occur in the directed sequence network.

All the co-occurrence/sequence associations can be combined into a united co-occurrence/sequence network. The united co-occurrence/sequence network can be used to study the rules hidden in the co-occurrence/sequence network globally.

Given a corpus of sentences, association networks of words could be constructed and analyzed as follows:

1) *Preprocessing*. Read a sentence from the corpus and segment the sentence into many chunks separated by "*blank*". The preprocessing processes for English sentence and Chinese sentence are different. English words are separated by *blanks* and *separators*, so the segment is easier. However, Chinese words are continuously written, so it is necessary to split the sentence into many chunks by Chinese segmentation programs such as ICTCLAS [16].
2) *Establishing associations*. Build sequence and co-occurrence associations between words/separators and words/separators in a sentence according to their positions in the sentence. Each word/separator in a sentence has a position: the first word/separator has position 0, and the succeeded word/separator is 1; if a word/separator has position n, the succeed word/separator has position $n + 1$.
3) *Combining associations*. Count the frequency of the associations (edges), and merge all the associations (edges) into the united co-occurrence network and sequence network, respectively.
4) *Visualization and analysis*. The association networks of words can be visualized into graphs with different layout. The characteristics of association networks can be analyzed by using analysis tools for complex network.

4 Experiments

To study the characteristic of association networks of English and Chinese words, we do experiment based on the data set of NIST MT 2005 English and Chinese bilingual parallel corpus. It contains 1082 pairs of English and Chinese parallel sentences. The English sentences and Chinese sentences in data set are used for building the association networks of English words and Chinese words respectively.

After the association networks of words in English and Chinese are constructed respectively, network analysis tool *NodeXL* is used to analyze the association networks, which is a free and open add-in for Microsoft Excel that supports network overview, discovery and exploration [14].

In association network analysis, duplicate edges are grouped by counting their frequencies. In the visualization graphs, duplicate edges are represented by the same edge.

The numbers of edges in co-occurrence and sequence networks are far more than those in direct sequence network, so it needs more time and memory to visualize them. The overview graphs of association networks with more than 10 thousands of edges look very dense, and the visualization graph is hard to distinguish the vertices and edges clearly. Fortunately, *NodeXL* has provided the zoom-in, filtering and selection functions on vertices and edges.

Figure 1 and Figure 2 show the visualization graphs of direct sequence networks of English words and Chinese words, respectively.

The visualization graphs of association networks have so many vertices and edges, and it is hard to find characteristics based on the visualization graph. So we have to study the association networks from the statistical measures of complex network.

978-1-935068-76-1 © 2011 SciRes.

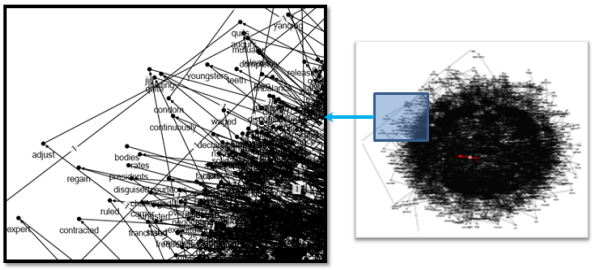


Figure 1. Direct sequence network of English words with 5052 vertices and 20346 edges: the right is the whole network, and the left is the zoom-in of a part in the whole network; the arrow of an edge means *before*

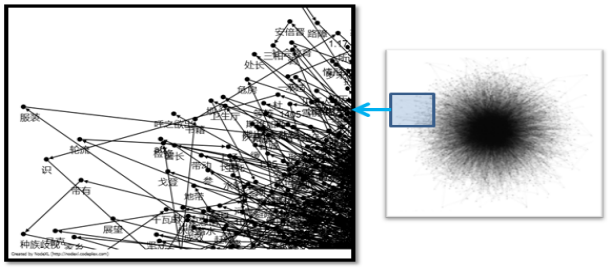


Figure 2. Direct sequence network of Chinese words with 5665 vertices and 20700 edges: the right is the whole network, and the left is the zoom-in of a part in the whole network; the arrow of an edge means *before*

Table 1 shows the statistical characteristics of direct sequence, co-occurrence and sequence networks of English words, while Table 2 shows the statistical characteristics of direct sequence, co-occurrence and sequence networks of Chinese words.

The characteristics of sequence network and co-occurrence network are very similar. The main differences are as follows: (1) sequence network is a directed graph, while co-occurrence network is an undirected graph; (2) sequence network has in-degree and out-degree characteristics, while sequence network only has degree characteristics. (3) The betweenness centralities are different.

English direct sequence network has two connected components, while Chinese sequence network has one connected components. Co-occurrence network and sequence network have more loops. It means that there are some duplicate words in the same sentence. The association networks of words obey the *small-world* characteristics of complex network.

The strange results are as follows: (1) for the co-occurrence network and sequence network of English words, the clustering coefficients are different; however, the clustering coefficients of co-occurrence network and sequence network of Chinese words are equal. (2) The closeness centrality is 0 in co-occurrence network and sequence network.

Table 1. Characteristics of association networks of English words

Characteristics	direct sequence network	co-occurrence network	sequence network
Graph Type	Directed	Undirected	Directed
Vertices	5052	5052	5052
Total Edges	20346	309654	309654
Self-Loops	2	624	624
Connected Components	2	1	1
Max Geodesic Distance	7	3	3
Avg. Geodesic Distance	2.878501	1.985562	1.985562
Graph Density	0.0008	0.0208	0.0208
Min Degree	*	4	*
Max Degree	*	4980	*
Avg. Degree	*	102.929	*
Min In-Degree	0	*	0
Max In-Degree	740	*	4976
Avg. In-Degree	3.997	*	60.851
Min Out-Degree	0	*	0
Max Out-Degree	930	*	4225
Avg. Out-Degree	3.972	*	60.844
Min C_B	0	0	0
Max C_B	4917054.7	1676699.7	3353399.3
Avg.C_B	9360.018	2425.419	4850.839
Min C_C	0	0	0
Max C_C	1	0	0
Avg. C_C	0	0	0
Min C_E	0	0	0
Max C_E	0.008	0.003	0.003
Avg. C_E	0	0	0
Min PageRank	0.239	0.184	0.184
Max PageRank	146.758	49.602	49.602
Avg. PageRank	0.99	0.994	0.994
Min *CC*	0	0.021	0.012
Max *CC*	1	1	1
Avg. *CC*	0.169	0.782	0.631

Table 2. Characteristics of association networks of Chinese words

Characteristics	direct sequence network	co-occurrence network	sequence network
Graph Type	Directed	Undirected	Directed
Vertices	5665	5665	5665
Total Edges	20700	252466	252466
Self-Loops	11	705	705
Connected Components	1	1	1
Max Geodesic Distance	7	4	3
Avg. Geodesic Distance	3.11758	1.999507	1.985562
Graph Density	0.0006	0.0157	0.0208
Min Degree	*	3	*
Max Degree	*	5591	*
Avg. Degree	*	89.132	*
Min In-Degree	0	*	0
Max In-Degree	847	*	5589
Avg. In-Degree	3.653	*	44.566
Min Out-Degree	0	*	0
Max Out-Degree	742	*	2413
Avg. Out-Degree	3.653	*	44.566
Min C_B	0	0	0
Max C_B	10463296.3	3595776.8	7191553.6
Avg. C_B	11999.208	2831.604	5663.207
Min C_C	0	0	0
Max C_C	0	0	0
Avg. C_C	0	0	0
Min C_E	0	0	0
Max C_E	0.011	0.003	0.003
Avg. C_E	0	0	0
Min PageRank	0.244	0.176	0.176
Max PageRank	194.391	64.458	64.458
Avg. PageRank	1	1.000	1.000
Min *CC*	0	0.016	0.016
Max *CC*	0.5	1	1
Avg. *CC*	0.098	0.768	0.768

5 Conclusion and Future Work

This paper studies co-occurrence and sequence between words in the same sentence, which can reflect the semantic associations among words. Co-occurrence and sequence associations between words in sentences formulate the association networks. We study the characteristics of the association network from the complex network viewpoint by using network analysis tool *NodeXL*. Experiment results show the difference between association networks in English and Chinese.

This work is preliminary in exploring the associations between words in English and Chinese. In the future network, we will study: (1) choose the important words from association network for information organization. (2) the alignment between words in English association network and those in Chinese association network. After the alignment between English words and Chinese words in parallel sentences, the rules in association networks of words in English and Chinese could be used to improve the machine translation between English and Chinese.

References

[1] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," Journal of the ACM, 46(5), 1999, pp. 604–632.

[2] L. Freeman, "A set of measures of centrality based on betweenness," Sociometry, 40(1), 1977, pp. 35–41.

[3] L. Freeman, "Centrality in social networks: Conceptual clarification," Social Networks, 1(3), 1979, pp. 215–239.

[4] J. Anthonisse, "The rush in a graph," Amsterdam: University of Amsterdam Mathematical Centre, 1971.

[5] S. Warshall, "A theorem on boolean matrices," Journal of the ACM, 9(1), 1962, pp. 11–12.

[6] M. Girvan , M. Newman, "Community structure in social and biological networks," Proceedings of the National Academy of Sciences, USA, 99(12), 7821, 2002.

[7] G. Sabidussi, "The centrality index of a graph," Psychometrika, 31(4), 1966, pp. 581–603.

[8] S. Wasserman, K. Faust, "Social network analysis: Methods and applications," Cambridge, United Kingdom: Cambridge University Press. 1972.

[9] P. Bonacich, "Factoring andweighting approaches to status scores and clique identification ," Journal of Mathematical Sociology, 2(1), 1972, pp. 113–120.

[10] N. Perra, S. Fortunato, "Spectral centrality measures in complex networks," Physical Review E, 78(3), 036107, 2008.

[11] L.Page, S. Brin, R. Motwani, T. Winograd, "PageRank citation ranking: Bringing order to the Web," (Tech. Rep). Stanford, CA. ,1998.

[12] V. Latora , M. Marchiori, "A measure of centrality based on the network efficiency ," 2004. Retrieved March 30, 2010, from http://arxiv.org/PS_cache/cond-mat/pdf/0402/0402050v1.pdf

[13] S. Fortunato, V. Latora, M. Marchiori, "Method tofind community structures based on information centrality," Physical Review E, 70(5), 56104, 2004.

[14] M. Smith, N. Milic-Frayling, B. Shneiderman, E. Mendes Rodrigues, J., Leskovec, C. Dunne, "NodeXL: a free and open network overview, discovery and exploration add-in for Excel 2007/2010"，2010.

[15] T. Opsahl, P. Panzarasa, "Clustering in Weighted Networks," Social Networks, 31 (2), 2009, pp. 155–163.

[16] ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System), [Website], http://ictclas.org/index.htm

978-1-935068-76-1 © 2011 SciRes.

Research on Knowledge Discovery Framework in Digital Library: An NSTL Case Analysis

Li WANG, Bin LIANG, Xiaodong QIAO
IT Support Center, Institute of Scientific and Technical Information of China, Beijing, China
Email: wangli@istic.ac.cn, liangb@istic.ac.cn, qiaox@istic.ac.cn

Abstract: With the advance of information technology, problems of "information explosion but knowledge poverty" are already apparent. How to extract useful content from such abundant information has produced increasing demands for, as well as increased activity to provide tools and techniques for knowledge discovery. Knowledge discovery is a concept of the field of computer science that describes the process of automatically searching large volumes of data for patterns that can be considered knowledge about the data [1]. Ascending research on utilizing knowledge discovery techniques in digital library includes Web data mining, personalization services, information intelligent push, resources development, knowledge sharing, etc. In this paper, we use Chinese National Science and Technology Library (NSTL) as a case, analysis the current literature resource discovery mechanisms, then propose a knowledge discovery framework based on STKOS, and further analysis knowledge services under the framework.

Keywords: Knowledge discovery framework; Knowledge base; NSTL; Digital library; Knowledge service; Literature resource discovery

1 Introduction[2]

Chinese National Science and Technology Library (NSTL), a public service organization providing S&T information resources in a network environment, was established in 2000 with the approval of the State Council PRC.

Over the past decade, NSTL continues to strengthen the guarantee of national S&T resource access, improve information services and sharing, and so become the support center for national S&T literature information service development. Currently, the S&T literature collected by NTSL includes academic journals, conference proceedings, reports, dissertations, patents, standards, and measurement specifications and so on. The total number of foreign periodicals holdings is about 2.6 million. NSTL embodies a whole set of standardized and scientific operation workflows and mechanisms, from literature-collection through resource processing, data storage, networking services, and data analysis of S&T literature. The total document record count is up to 110 million. While stabilizing the collection and availability of printed resources, NSTL is also actively purchasing on-line and full-text editions of journals for users nationwide.

2 NSTL Digital Service Platform and the Literature Resource Discovery Mechanisms

NSTL provides scientific researchers and educators with nationwide information availability via the Digital Service Platform. This service platform provides a central web service based in Beijing, and also mirror image websites in eight other cities, to ease access from across the nation. In addition, for the user's convenience, NSTL has developed multi-service modes, such as service stations, embedded services and user management platforms. Currently a nationwide service network, covering 25 provinces and cities around the country has been formed.

One especially critical aspect of the online service platform is "all roads lead to Resource".

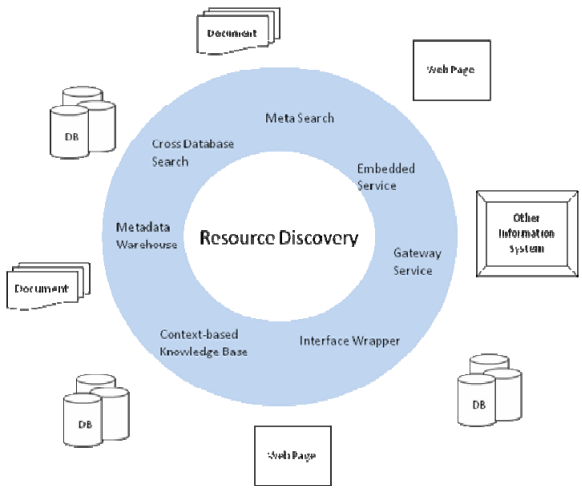


Figure 1. Literature Resource discovery mechanisms of NSTL

2.1 Metadata Warehouse

The resource discovery mechanism based on metadata warehouse is a physical integrated method. All kinds of metadata are stored in the same physical location (data warehouse) in accordance with specific modeling methods (usually by subjects, or other multi-dimensional ap-

proach). The unified target data model (schema) provides a one-station access to a variety of digital resources.

NSTL proposed different XML schemas adapt to the different document types. These metadata are stored in the same metadata warehouse, and released indexing databases by a dissemination system.

NSTL's search function, based on metadata warehouse, is a multi-search engine. Users can select one or more databases (organized by document type, such as foreign journals, patents). The multi-search engine sends user's search words to the indexing databases, then extracts metadata from the metadata warehouse and shows the organized results.

2.2 Meta Search

In order to compensate for the great gap of foreign S&T literature in Chinese national history, NSTL places great emphasis on the purchase of major international retrospective literature databases. Currently, NSTL has purchased 6 retrospective literature databases (metadata and full-text, not just access privilege), including 1500 kinds of periodicals and more than 400 million articles.

NSTL provides a general metadata model, which contains only the essential elements, to integrate these different resources. Different data sources can be accurately mapped to this model, and also the general model must be able to reverse the original records to obtain the target source. NSTL's retrospective system, as a part of the digital service platform, extracts metadata from the raw resource and builds a unified set of metadata, to support one-stop searching and full-text download directly.

2.3 Context-based Knowledge Base

Context-based knowledge base is designed to provide combined information discovery/disclosure capabilities for agencies. The knowledge base can be deployed in NSTL center sites, mirror sites, or service sites. The institutions providing services jointly with NSTL may register their own resources into the base, enabling them to be found through the NSTL retrieval module. During information retrieval, users will be shown both local resources and centralized resources from NSTL, enabling the maximum use of local sources.

The role NSTL plays is not only that of an information provider, but also an infrastructure supplier for resource discovery, retrieval, and services. It is possible for users to choose whatever mode of service as they like, with NSTL serving as the channel of last resort for literature resources, when others are not available. NSTL integrates these roles, providing literature discovery services and guaranteeing the infrastructure for related services, supporting the availability of domestic S&T literature services.

2.4 Cross-database Search

NSTL is developing its own cross-database search, to offer one-stop searching of specific web sites, which provide online and full-text foreign current periodicals ordered by NSTL. Clicking on a result link will go to the original web site. Anyone can try it out, but those not part of authorized agencies won't see results for licensed databases.

The key challenge for this cross-database search is how to package up the search and process the results. This service uses a somewhat primitive technique called "screen scraping". Screen scraping [3] is the process of collecting needed information by clues such as the location of the information on the screen. The problem is that the slightest change in screen displays can break your process. In general, the more databases a search service covers the more challenges it will face.

2.5 Interface Wrapper

NSTL offers many types of web service interfaces (including retrieval, document delivery, and so on) to support the use of the third-party information service providers. So, NSTL online service has moved from the simple mode of providing services directly to users, to the multi-service mode, efficiently supporting of technology and resource usage to the nationwide science and technology service institutions.

2.6 Embedded Service

This embedded service offers a set of resources and tools that help users to interact with information between NSTL spaces and the personal information spaces. It is designed to support diversity among platforms, information and users intensifies, using a distributed architecture.

2.7 Gateway Service

This gateway service operates in 31 databases from 7 information organizations, including NSTL, China Academic Library & Information System (CALIS), NSLC, National Library of China (NLC), Zhejiang Institute of Scientific and Technological Information, MIS, the Lanzhou Branch of the National Science Library. The key components are a registration system and a retrieval system. The registration system was established using UDDI, to enable cross-database searches on the resources registered by information organizations. The retrieval system offers a simple smart interface.

3 Knowledge Discovery Framework based on STKOS [4-11]

In the digital information environment, user needs are changing from data, information to knowledge. Information profession services are required to identify, analyze and coordinate the various needs of various user groups.

Knowledge discovery is a concept of the field of computer science that describes the process of automatically searching large volumes of data for patterns that can

978-1-935068-76-1 © 2011 SciRes.

be considered knowledge about the data [1]. Knowledge discovery technology can help discovery hidden knowledge in huge resources collection of digital library, and improve resource utilization, information service quality and user satisfaction. How to use knowledge discovery technology in the digital library is causing more and more attention. In foreign countries, representative applications include KDWebS[1][9], the Bio-energy Knowledge Discovery Framework[2], GeoLinkedData[3], applications of link data in BBC[4] etc. Although related researches include web data mining, personalization services, information intelligent push, resources development, knowledge sharing, most applications focus on link data and visual mashup. CNKI[5] (China Knowledge Resource Integrated Database), a specific example in China, is an integrated system of knowledge resources and knowledge service. The core of CNKI is knowledge element database, which emphasis knowledge element node. The link between knowledge elements only remain in the lexical level, and lack of semantic analysis.

NSTL is funding some projects and programs in the fields of machine translation, data link, user behavior analysis, intelligent retrieval, personalized recommendation, to explore key technologies involved in knowledge service for long-term goals. The literature resource discovery network has some characteristics of knowledge discovery such as facet browsing in the retrospective system, multi-link in the search engine, although there is a wide gap between the current online services and knowledge services.

In the 12th Five-Year plan, NSTL expands investment to research on construction and application of Science and Technology Knowledge Organization System (STKOS), focuses on key issues such as content construction of the domain-specific KOS, deep indexing of document content, S&T terminologies service. A complete knowledge space will be formed, combining KOS with literature resource, to support efficient and convenient knowledge service.

3.1 Plug-n-Play Analysis Engines

Analysis Engines process entity/relation detection. A different engine is used to identify different entities, for example, person entity, organization entity and project entity. An analysis engine can be a simple one which

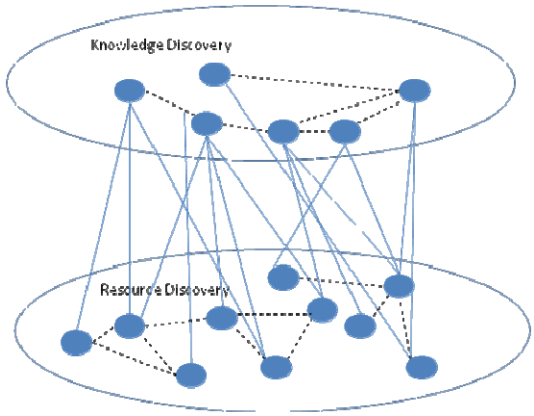


Figure 2. From "resource discovery" to "knowledge discovery"

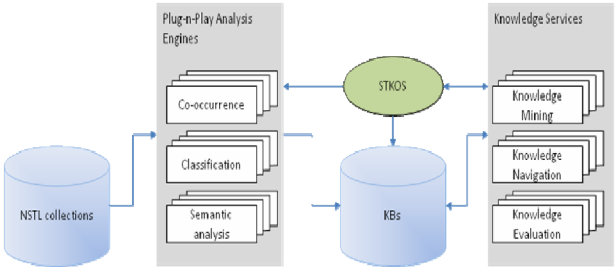


Figure 3. Knowledge Discovery Framework based on STKOS

parses out the original information to facilitate further analysis. Analysis engines may be organized and composed together to form reusable components that encapsulate rich workflows of cooperating engines. Those recursive partitioning engines combination by the other analysis engines are named "Aggregate Analysis Engine". Parsing out any results produced by previous analysis engines, they discover more entities and relations.

Specific steps of an analysis engine can be described as:

- Parsing out raw data obtained from documents or the previous analysis engine;
- Finds Mentions of entities in documents (each engine has its specific extraction ontology);
- Performs in document co-reference;
- Creates data in terms of common data representation.

Common data representation following standard XML syntax conventions is offered to support pluggable. The logical interface for an analysis engine is simple -- read access to the raw data being analyzed and write access to the analysis results. This simplicity facilitates interoperability and composability of independently developed engines.

3.2 KBs

Knowledge elements, extracted from literature resources through analysis engines, expresses knowledge

[1] KDWebS: Knowledge Discovery System based on Web Services, a scalable web services based tool to facilitate automated knowledge discovery in NDLTD. The key of KDWebS is emphasis of link labels, not just node labels, using information that is based on Natural Language Processing (NLP) and an ontology of computing science.
[2] The Bio-energy Knowledge Discovery Framework: BIOMASS program supported by U.S. Department of Energy. (https://bioenergykdf.net/)
[3] GeoLinkedData: an open initiative whose aim is to enrich the Web of Data with Spanish geospatial data. (http://geo.linkeddata.es/browser/#dashboard)
[4] http:// www.bbc.co.uk
[5] http://www.cnki.net/

contents clearly. The logical relationship between knowledge elements is called "knowledge chain". Knowledge elements, knowledge chain, and instances of literature resources constitute a knowledge base providing knowledge analysis and mining.

The main aspects of KBs are:

- Knowledge entities and link
- Identifies instances of the entities at the collection level
- Includes evidence chain back to original documents and mentions
- Builds view of the data in terms of triples

KBs offer multiple views and multi-modal, such as:

- *Document network*: each document as a node, around which organized many related information nodes, such as references, similar documents, related media, scholars in the same research field, and so on.
- *Knowledge element network*: constructed by coupling and link of concepts.
- *Classification network*: to support multi-category and multi-navigation, rather than a simple classification tree.
- *Scholar network*: constructed by links among collaborators, tutors, published literatures, literature evaluation feedbacks and other academic activities.
- *Citation network*: consist of nodes such as references, cited literature, secondary references.

3.3 STKOS

STKOS can be divided into domain-specific KOSs and research KOSs. "Domain-specific" means four major areas (science, technology, agriculture, and medicine) This KOS, based on the super S&T vocabulary, is a lightweight ontology including normative concept, semantic standardization and semantic relationships, which can support knowledge organization and link discovery of literature resources and scientific research objects. Research KOS focuses on academic activities and the links between research objects (such as researchers, institutions, equipments, achievements).

Knowledge Discovery Framework based on STKOS, with the help of ontological hierarchy and logical reasoning, can effectively support knowledge representation, knowledge discovery and semantic resolution in different areas.

4 Scenarios of Knowledge Services

The Knowledge Discovery Framework based on STKOS is designed to support effective exploration of knowledge.

1) Multilingual Information Discovery: STKOS, bilingual designed in English and Chinese, will provide automated support for multilingual information discovery. This multilingual information discovery based on concept mapping beyond query-translation, can effectively improve accuracy and professional level of translation. [12]

2) Multidisciplinary Information Discovery: is based on knowledge relationship among science, technology, agriculture, and medicine.

3) Intelligent Retrieval and Visual Mashup: Information retrieval based on STKOS, with the technological improvement of knowledge mining, co-occurrence analysis, relation computing, is developed into a semantic-based retrieval, to provide knowledge navigation, automatic clustering, natural language retrieval, visual knowledge map and so on. This intelligent retrieval engine can expand what can be searched with knowledge reasoning, and visually combines search results from diverse data source in a mashup [13] visualization. One of the most common mashup is integrating geographic data of organizations with literatures produced by these organizations.

4) Scalable Multi-facet: Knowledge network is actually a collection of different facet spaces. An individual facet is a hierarchy tree and these basic tree structures are extended by Related-To links and poly hierarchies. This multi-facet schema, describing literature resource from multiple dimensions, allows the user to restrict the scope of the search and to post hits against facets, when used in conjunction with search.[14]

5) Personalized Recommendation: to provide related literature based on data mining of user behavior information (focus on search and browsing behaviors), and to detect potential social network and enhance academic activities.

6) Monitoring and Trend Analysis on Science and Technology: to offer multi-angle, multi-level analysis of development in subject areas based on the knowledge network, and detect developing trends and hot spots all of the world.

7) More details to Improve the user experience: [15]

a) Emphasizes the simple and productive interface ("Simple search but supported by smart results and rich browse.");

b) Increases user participation, such as tag, review;

c) Completely or partly embedded into the user environment (Toolbar, Widgets—Portal);

d) Provides a personalized overview of a subset of the whole knowledge repository of interest to a user, based on his or her profile.

5 Summary

The goal of digital libraries is to enable ubiquitous access to knowledge cooperating with other knowledge infrastructure (e-learning, e-research, e-science, e-government, e-business), and to promote knowledge's understanding, sharing and utilization. The knowledge discovery framework based on STKOS, integrating information at knowledge element level and establishing relations

978-1-935068-76-1 © 2011 SciRes.

between knowledge elements, is essential to achieve the goal. Based on the knowledge network, the new literature can be found through the existing one, and the potential knowledge can be found through the explicit knowledge. [15,16]

References

[1] Frawley William. F. et al. (1992), Knowledge Discovery in Databases: An Overview, AI Magazine, 13(3), pp. 57-70.
[2] Qiao Xiaodong, Liang Bing, and Yao Changqing, China National Science and Technology Digital Library (NSTL), D-Lib Magazion, vol. 16, no. 5/6. [Online: http://www.dlib.org/dlib/may10/xiaodong/05xiaodong.html.]
[3] Roy Tennant, Digital Libraries- Cross-Database Search: One-Stop Shopping, Library Journal.[Online: http://www.libraryjournal.com/article/CA170458.html.]
[4] NSTL 12th Five-Year plan, unpublished.
[5] UIMA Architecture Highlights, [Online: http://domino.research.ibm.com/comm/research_projects.nsf/pages/uima.architectureHighlights.html.]
[6] LI Yan, JI Xin-sheng, and XIANG Jun, Research and implementation of UIMA-based knowledge discovery framework, Computer engineering, Vol. 36, No. 21, 2010, pp. 277-279.
[7] SHAO Long-bin, Concept of semantic search and UIMA, Journal of Ningbo Ploytechnic, Vol.13, No.2, 2009, pp. 80-83.
[8] Werner Klieber, Vedran Sabol, and Markus Muhr, Knowledge discovery using the knowminer framework, [Online: http:// know-center.tugraz.at/download_extern/papers/IADIS_Knowminer.pdf.]
[9] W. Ryan Richardson, Venkat Srinivasan, and Edward A. Fox, Knowledge discovery in digital libraries of electronic theses and dissertations: an NDLTD case study, International Journal on Digital Libraries, Vol. 9, 2008, pp.163-171.
[10] Martin Labský, Vojtech Svátek, and Marek Nekvasil, Information extraction based on extraction ontologies: design, deployment and evaluation, [Online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.142.6837&rep=rep1&type=pdf.]
[11] Winyan Chung, Hsinchun Chen, and Jay F. Nunamaker Jr, A Visual Framework for Knowledge Discovery on the Web: An Empirical Study of Business Intelligence Exploration, Journal of management information systems, Vol. 21 No. 4, Spring 2005 pp. 57-84.
[12] Douglas W. Oard. Global Access to Multilingual Information. [Online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.45.9508&rep=rep1&type=pdf.]
[13] Anselm Spoerri, Visual Mashup of Text and Media Search Results, 11th Int. Conf. on Information Visualization (IV 2007), Zurich, Switzerland, July 3-6, 2007, [Online: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.157.4453&rep=rep1&type=pdf.]
[14] R.B.Allen. Retrieval from facet spaces. [Online: http://www.ischool.drexel.edu/faculty/ballen/PAPERS/FACETS/facets.html.]
[15] Knowledge lost in information, Report of the NSF Workshop on Research Directions for Digital Libraries, June 15-17, 2003. [Online: http://www.sis.pitt.edu/~dlwkshop/report.pdf.]
[16] Zhang Wenxiu, Futures of Digital Library--The Post Digital Library Age, New technology of library and information, 2006,no.5, pp. 1-5,39.

Scientific Research

The Service Platform of National Science and Technology Library

Bing LIANG, Li WANG, Xiaodong QIAO

Information Technology Support Center, Institute of Scientific and Technical Information of China, Beijing, China

Email: Liangb@isitc.ac.cn; wangle@istic.ac.cn; qiaox@istic.ac.cn

Abstract: NSTL (National Science and Technology Library) was jointly founded in 2000 by five Ministries, including Ministry of Science and Technology and Ministry of Finance, with approval of China's State Council. NSTL consisted of nine members. NSTL provided the nation-wide services in ST literature information mainly focused on science, technology, agriculture and medical fields. This article not only presented the background of ST literature service system of NSTL, but also introduces the NSTL's internet-based service system construction and its future development orientation of service construction.

Keywords: NSTL; Digital Library

1 Introduction

1.1 Background

Chinese National Science and Technology Library (NSTL), an organization of public service with ST information resources based on network environment, was established on June 12, 2000, with the approval of the State Council PRC. NSTLC is made up of National Science Library of Chinese Academy of Sciences (NSLC), Institute of Science and Technology Information of China(ISTIC), Chinese Machinery Industry Information Institute, China National Institute of Standardization, China Metallurgical Information and Standardization Institute(MIS) and China National Chemical Information Center (CNCIC), Agricultural Information Institute of CAAS, Institute of Medical Information of Chinese Academy of Medical Sciences, Standards Library of China National Institute of Standardization and Literature Library of National Institute of Metrology. On the standardized and scientific workflow management plan courses covering literature collection, resource processing, Data storage, online-service and ST literature data analysis, and the principals as centralized procurement, standardized processing, unified on-line release and resource sharing, NSTL addresses its resource collection, archives and application services regarding science, technology, agriculture, medical literatures and information and provided the nation-wide services in ST literature information.

Most domestic libraries and the literature & information institutions provided service for readers mainly in the light of acquisition of cheap photocopies of the international literature until China became one of formal members of WTO. As China joined WTO in 2001, the Chinese Publication Import & Export Corporations had to stop their reprinting business consequently with the prompt of high demands for foreign literature in China. NSTL rapidly complement subscription of the literature which is no-longer photocopied and released the online document delivery service on the nation-wide scope. As the result of the strong guarantee from NSTL, universities and colleges in China began to decrease demands for printed journal and turned to purchase online literature. Meanwhile, the National Library adjusted the resource development construction, reduced the purchasing acquisition of international ST literature and greatly expanded the list of foreign social works. Almost all the institutes of information and service centers in China no longer purchased foreign ST literature and made their work turn to collecting local literature information and characteristic resources, and extended their services upon the advantages of NSTL resources. In a word, the establishment of NSTL national ST literature guarantee system has significantly contributed to integrated structured deployment optimization of China's international literature resource.

Supported by National Eleventh-Five Year Research Program of China (2011BAH10B05)

1.2 Organizational Structure of NSTL

NSTL implements the director responsibility system under the leadership of the council, which is responsible for the overall leading and decision-making and is constituted by the recognized scientists, information experts and representatives of other relevant departments. NSTL shall operate under the policy instruction and the supervision of the Ministry of Science and Technology (MOST) in behalf of all six ministries. The director's main job is to organize the personnel and work implementation. A virtual office set takes charge of management of the sharing system of the ST literature information resources. The consultation and guidance for the relative service operation may launch from the Information Resource Experts Committee and the Computer Network Service Expert Committee.

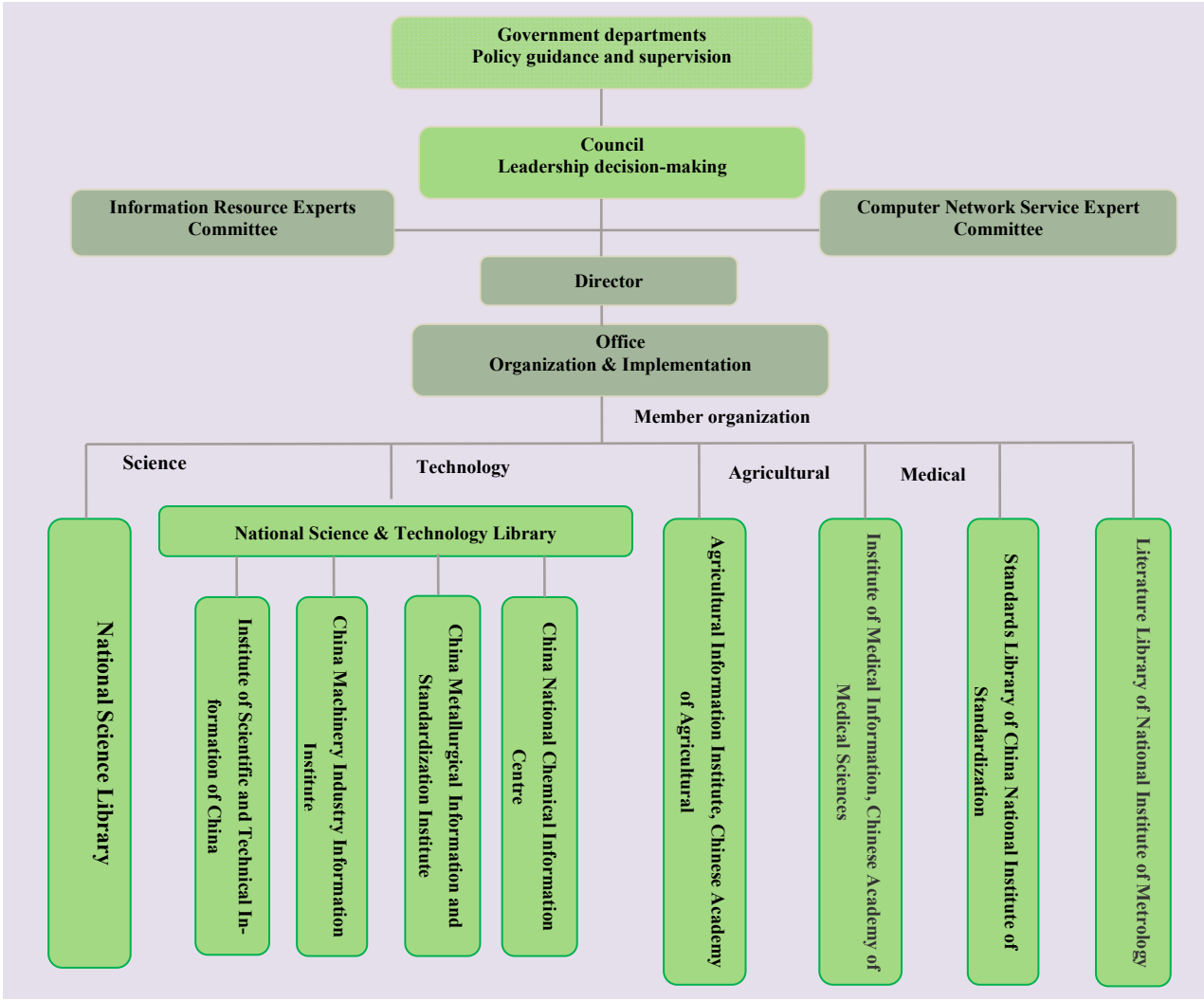


Figure 1. NSTL Organizational Structure

Table 1. Main Organizations of Scientific and Technical Literature Service in China

	The National Science Library	public library	CALIS	NSTL
Supporters	Chinese Academy of Sciences	Ministry of Culture & regional government	Ministry of Education	Ministry of Science and Technology
Resource type	Mainly natural sciences, journals, proceedings, dissertations, dictionary Mainly purchases database resources, basically does not purchase print resources, and through joint procurement to achieve co-building &Sharing	Mainly humanities, social sciences, books, local history, ancient books. Mainly prints, is digitalizing the prints.	Humanities, social sciences, natural sciences are equal importance; main collections are journals, proceedings, dissertations. Mainly purchase database resources, basically, do not purchase print resources.	The main collection of tech-related literature, journals, proceedings, science and technology reports, standards, patents, dissertations. Mainly purchase print journals, supplemented by electronic resources, institutions process and index their own prints to digital resources.
Service mode	Retrieval literature data, according to the permission to download. For no permission to provides ILL	Mainly serve in the library, network service function is limited	Retrieval literature data, according to the permission of universities to download, for no permission to provide ILL.	Retrieve bibliographic data, issued request orders for literature services to, processing orders manually, by email, or cache way to download, provide scanned documentation.
distribution of service user	Serve for al Institutes of Chinese Academy of Sciences, is a closed services	Service for worldwide users	Serve universities joined CALIS system, is a closed services.	Serve the region of mainland China.

Service fee	purchase of database services, the users not need to pay, direct access and download documents	free viewing in the library, copy fee is 0.3 Yuan / page	As the purchase of database services, the users of major colleges and universities within CALIS do not pay the fee, direct access to download documents.	Document delivery charges by types, 0.3 Yuan / page The users of region of western China have 50% off discount.
Advantage	Serve for China's leading institutes, can directly download the retrieved documents, can centralized resource allocation, coordinate funding arrangements. The system experts, academicians can quickly trace the latest scientific research progress.	Rich in resources are On China's history, culture etc. as a national libraries, provide services to the world, in cooperation with many countries, the National Library is completed, advanced facilities.	Combined all the key universities, to form a unified service that can be downloaded directly retrieved all kinds of literature, colleges and universities build system quite different, in the standardization of construction, the building CALIS has well done to promote better outcomes, in cooperation with other software developers.	Types and quantity of science and technology resources are most comprehensive, provide services to scientific research and education for the country, is the highest level of domestic network, the system is most open service platform, rely on central stations to build a nationwide service system, unified management, service system, unified management, Uniform deployment of resources.
disadvantage	All the resources are in the database providers or publishers, subject to constraints of network conditions and social conditions, if not renewed, documentation and data can no longer continue to serve. A large number of resources only serve research personnel of the acandemy. Other institutions of society cannot be served.	Can provide the network information service under the conditions enough, most are the library service, in the field of literature information, is lack of natural sciences collections. Because a large number of collections books, ancient books and other resources, circulation is small	System serve institutions joined CALIS, have more resources, share resources, those who are not key or general colleges and universities is difficult to have these services, due to various types of colleges and universities differ from the administration and software systems, document delivery efficient is in lower. Also subject to network conditions and social conditions.	Document delivery need staff in background to process, cannot be real-time download, and customers feel inconvenient.

It was shown that the main organizations of scientific and technical literature service in China. In this table, different characteristics of organizations were compared.

2 Resources Types and Architecture of NSTL

2.1 Resources Type

Till now, the resources of ST literature collected by NTSL include Chinese, Japanese and Russian ST journals, conference proceeding in Chinese and English, English ST reports, and dissertations in Chinese and English; the patents of China in Chinese, Seven Countries (Australia, German, France, English, Japan, Russia, and America) and Two Organizations (EPO and WIPO); those of Chinese and Foreign standards, and as measurement specifications as well. While stabilizing the construction for the printed resources, NSTL starts actively to purchase the on-line & full-text editions of journal in institutes and associations for the nationwide users on the one hand; to lay a great emphasis on taking the state buyout of major international retrospective literature databases as the key resource construction project to provide services for countrywide science researchers to compensate the great missing of foreign ST literature in Chinese national history. At the same time, NSTL increasingly strengthens its revelation and service ability in the aspect of open resources on the Internet.

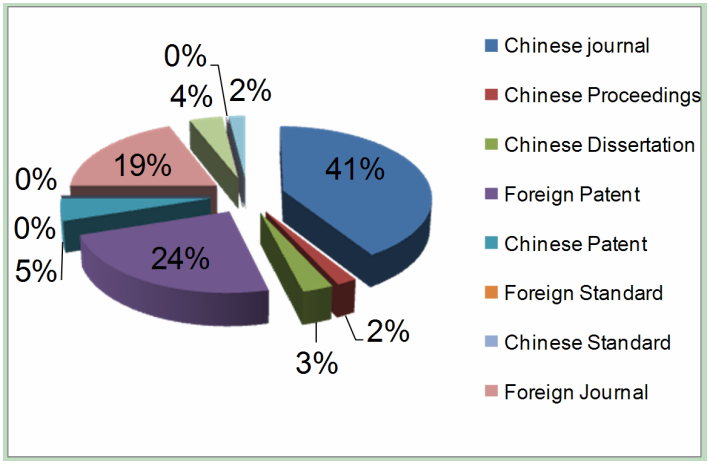


Figure 2. The proportion of a variety of data for online services system

978-1-935068-76-1 © 2011 SciRes.

2.2 Architecture of the NSTL System

NSTL embodies a whole set of standardized and scientific operation workflow as well as working mechanism, which starts from literature-collection, resources processing, data storage, networking services and data analysis of ST literature as the detailed following as:

---literature-collection: to organize and study the evaluating indicator system, to establish the selected database of NSTL literature resources, to unify budget and procurement guided by the development of the country and the society;

---data processing: to distribute the collected literature to each unit on the basis of rules, and process the metadata information during the standardized time-limit and with the accuracy examination on the overall process;

---data storage: to establish unified and centralized data storage, as the foundation to supply Web services, data-mining and information extraction, uniformly store all types of metadata;

---on-line service: NSTL online service system has regulated the evaluation mechanism of limitation period & service quality specific to full-text service. The system provides 24-hour service to the nationwide users, with the requirement that the service of document delivery shall be finished within one working day, and the online service platform has put into practice the incessant service for 7*24 hours. Additionally, NSTL has also erected the service station mode to speed up the construction and popularization of the central service system, and to improve the application scale and efficiency of the system resources

---The data analysis of scientific and technical literature: The analysis and process to the citation numbers from foreign journals done by NSTL facilitate the deep-down discovery of the scientific information for future use.

3 NSTL Service Scope and Its System Construction

3.1 Brief Introduction to Online Service

NSTL provides scientific researchers and educators with nationwide information via Internet, whose service includes meta-data retrieval of ST literature, compositive revealing, document delivery, checking and borrowing, advisory & consultation, citation service, retrospective data service, Chinese preprint service and domestic hotspot scientific gateways and so on. The detailed of the above is described as:

---meta-data retrieval: through the NSTL network service platform, all types of meta-data resource of NSTL can be retrieved in the integrated pattern, and the conditions on local resource of the user can be determined through the assistance of knowledge base, which can facilitate users to reach the scientific literature on basis of the familiar condition;

--- compositive revealing: to establish the registration system for the national ST resources with the standard service interfaces, carrying out the cross-database search on the resources registered by the domestic information services, which is easy for readers to get the ST information resource;

---document delivery: AS the response to the full-text requirement of ST literature resulted in the meta-data retrieval can be submitted to the online service platform, NSTL servers is supposed to finish the delivery within specified time. Users can also send their demands to the NSTL for all types of ST literature through the service of "checking and borrowing done by the replacer", while servers would transferred their applications to the artificial service to provide assistance in the seeking of the desired literature;

---advisory & consultation: NSTL provides virtual consultation service on network, with consultants from nine membership ministries to offer on-line & real-time advice in turn or the non-real-time consultation by e-mail;

---citation service: NSTL processes and analyzes the citation information on the international journals and conference resources, to supply the domestic web users with comprehensive information mining service;

---retrospective data service: As the retrospective database resources purchased from the internationally renowned publishers by means of state buy-out, NSTL to offer the related retrieval and full-text download services to the nationwide scientific researchers and educators charge without any charge;

---Chinese Preprint service: There is a Chinese preprint service center in NSTL construction, accepting research papers from domestic users respectively;

---domestic ST hotspot gateways: NSTL has built its domestic hotspot scientific gateways in 17 academic fields according to the current ST development, collecting research hotspots and dynamic information in this area for the convenience of scientific researchers.

3.2 NSTL Nationwide Service System

NSTL online service platform not only provides WEB central station service in Beijing, but set up mirror image websites in other 8 cities to ease the access from the nationwide users. In addition, for the user's convenience, NSTL has developed service station mode, extending NSTL system to all institutions to intensify the function of service system.

3.3 Service Mode of NSTL

NSTL online service has transferred from the simple mode of directly-to- reader to the multi- service mode, supporting efficiently the business of other domestic information service institutions. Many types of Web Service Interfaces (including the retrieval interface, document delivery interface and the interface for the replacer so on) offered, NSTL online service system opens to

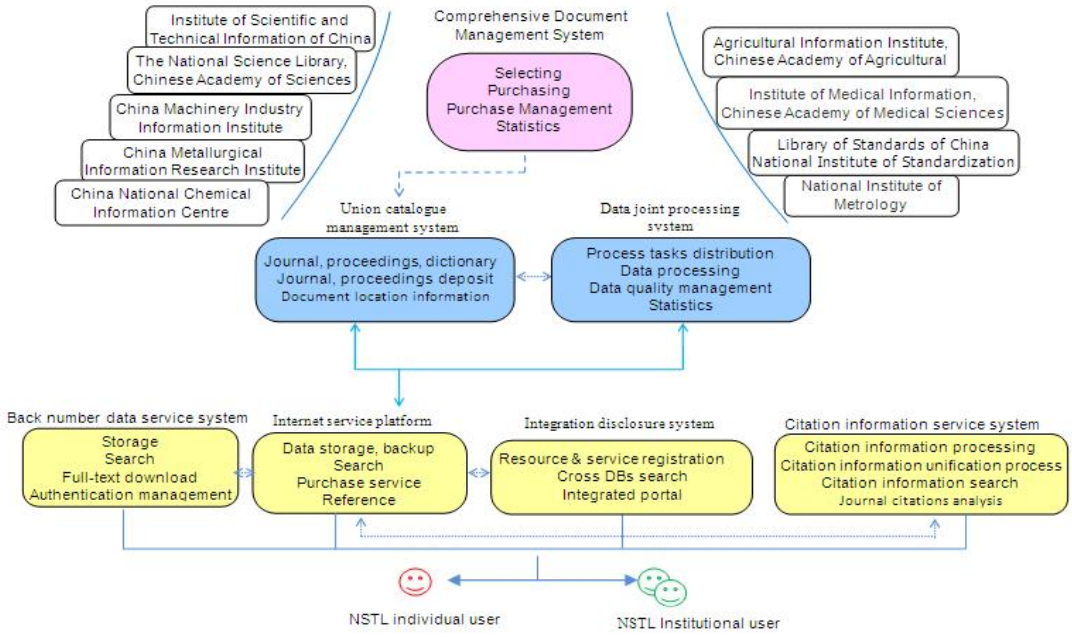


Figure 3. NSTL digital service platforms

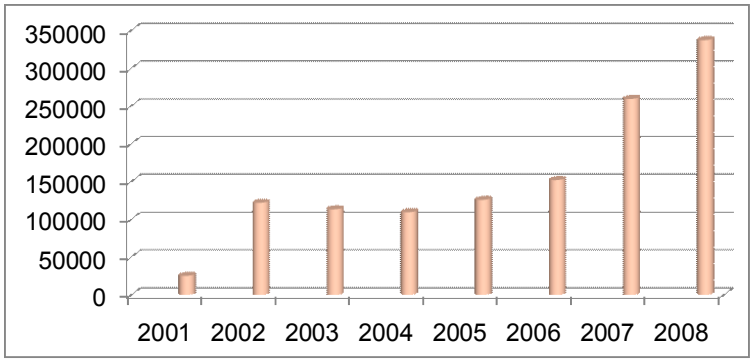


Figure 4. Comparing the number of document delivery service from 2001-2008

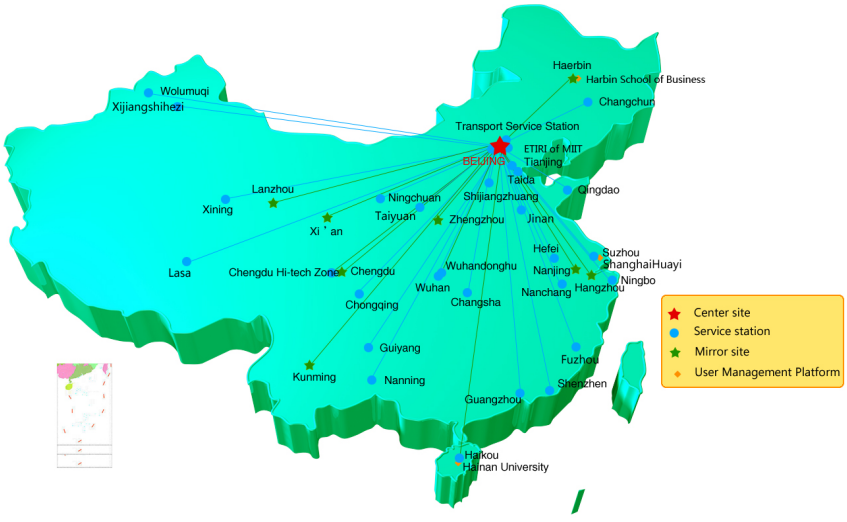


Figure 5. NSTL nationwide services

978-1-935068-76-1 © 2011 SciRes.

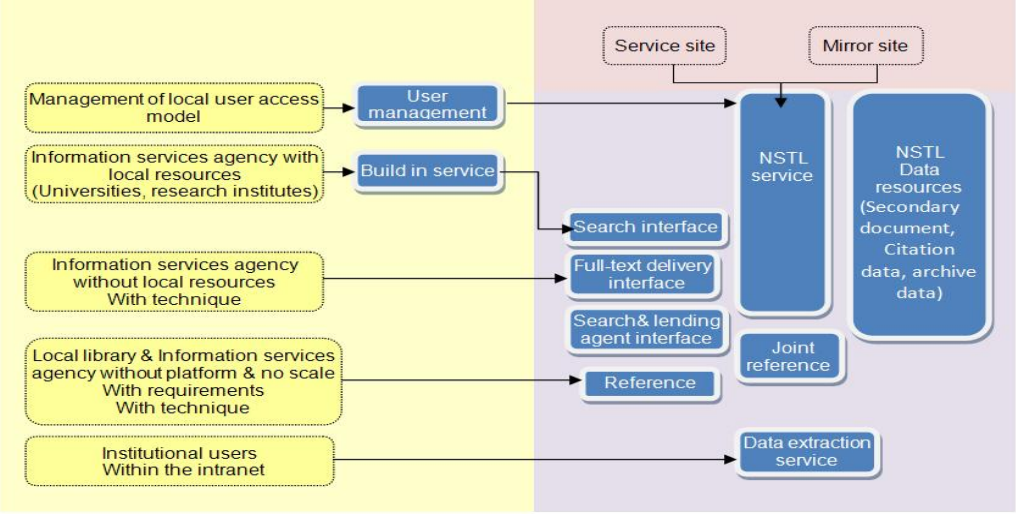


Figure 6. NSTL Service Models

individual readers, as well as domestic information service agencies. It completely supports the NSTL resource calling programme of the third-party information service providers. Those with inadequate resources and funds can enjoy the opportunity to get access to NSTL resources more easily, supplementing the efficient residual support of technology and resource to the nationwide science and technology service institutions.

The knowledge base, NSTL network service system constructed, can be deployed in NSTL' center sites, mirror sites, or service sites. The institutions with the provision of services jointly with NSTL, may register its own resources into the NSTL base and reveal compositively them through NSTL retrieval module, resulting in the development of own resource discover and service. During the information retrieval, users will be revealed the local resources and resources from NSTL in a centralized way by the system and be led to get more convenient literature service from the local sources as possible as they can. Meanwhile, the role NSTL plays is not only an information provider, but an infrastructural supporter of resource revealing and service safeguard. It is possible for users to choose whatever mode of service as they'd like and NSTL would be the last channel and assurance of the literature resources, once others are not available. That NSTL integrates roles either firstly as the discoverer of literature resources or high hierarchical guarantor of the related utility system further indicates the assurance and supporting efforts of NSTL in domestic ST literature service.

4 Future

The information assurance system of ST literature is an important portion of the national innovative system. With the rapidly growing development of science and technology in China, NSTL will, with no doubt, undertake increasing challenges and be entrusted the historical tasks to widen the service scope and improve the service quality. In the new situation, to continue the perfection and development of service, NSTL should:

1) Establish a high-efficiency center service system and improve the ability to provide integrated service

---Keep updating technology and reform, enhancing the retrieval function and cementing further the construction of knowledge structure system;

--- Deepen what is revealed and the correlation service, focusing on the knowledge mining ability based on the content of literature;

2) make overall services online and open completely available with support to the 3rd party serving ability development;

---Continue to support the 3rd party information service institutions and provide joint service, linking center service with other institutions concerning the retrieving of local literature, literature delivery and information consulting, make NSTL service a part of the local literature service and maximize the convenience of NSTL service for end-users, support the secondary development and in-depth tailor-made of NSTL resources and services in the forming new service ability in literature information providing.

3) Organize and support members and the 3rd parties to popularize NSTL service, encourage them to make use of NSTL resources and launch value-added service, and level up the operation of Guarantee system of China national ST Literature service.

References

[1] Yuan Haibo. Practice of the Joint Construction and Sharing of Information Resources--the Construction and Development of National Science and Technology Library. JOURNAL OF THE CHINA SOCIETY FOR SCIENTIFIC AND TECHNICAL INFORMATION , 2002, 21(1):57-62.

[2] Zhang Xiaolin. The Effectiveness and its Evaluation for NSTL. LIBRARY AND INFORMATION SERVICE , 2008,52(3):62-65.
[3] Qiao Xiaodong . The Change for Service Mode of National Science and Technology Library in the New Era. DIGITAL LIBRARY FORUM , 2008(12):50-52.
[4] Liang Bing. The Development of NSTL Science and Technology Literature Service System in the New Era of Independently Innovates. The Conference of Service and Development of Special Library in the New Era of Independently Innovates, Guangdong, China, 2006.

978-1-935068-76-1

Semantic Bibliography Based on Ontology and Linked Data

Haiyan BAI
Information Technology Center, Institute of Science and Technology of Information of China, Beijing, China
Email: bhy@istic.ac.cn

Abstract: In digital environment, information objects appear more and more multiplicity of forms, variability in lifespan and complexity of hybrid digital objects, so it is needed to provide tools and methods to identify, describe, and access these characters and other bibliography relationship. However, traditional information organization is not good at extracting, expressing and sequencing of bibliography relationship. In this paper we describe the techniques used to make NSTL Union Catalog part of Semantic Web and Linked data and introduce the method to organize amount of structured, high-quality data into semantic bibliography based on entity and entity relationship. And this paper proposes a basic idea of deeply sequencing of information that builds semantic linking based on ontology, describes semantic information accordantly using RDFS/ OWL, accesses unitely using SPARQL. The focus is on construction of organization mechanism to improve the ability of semantic describing and provide semantic retrieval based on objects relationships.

Keywords: library catalogue; semantic web; linked data; persistent identifiers; frbr; sparql

1 Introduction

It is a hot issue that semantic technology is applied in field of information organization. The study of semantic bibliography based on ontology and linked data arises from the needs of bibliography organization and integration of National Science and Technology Library (NSTL) of China workflow. NSTL, established in 2000 by the Ministry of Science and Technology, is a library federation and STM literature sharing platform consisting of nine special national libraries serving basic sciences, agricultural sciences, medical sciences, and engineering. Its workflow are comprised of the Union Acquisition System, Union Catalog System, Union Data (abstracts, contents and citations) Processing System, Central Repository and the Union Service System supporting cooperative collection building, integrated abstracts retrieval and contents browsing, and interlibrary loan service. The union collection emphasizes on academic resources in fields of science, technology and medicine, including 20,000 journals and 100,000 proceedings, which contain many related information objects but the relationships are not identified fully or described obviously. At the same time, information objects appear more and more multiplicity of forms, variability in lifespan and complexity of hybrid digital objects. So it is necessary to build a mechanism to identify, describe and organize the characters and relationships of all kinds of information objects and provide tools and service for end users to access and browse information objects and their properties expediently by following the linked web. While semantic technology provide a corresponding basic framework that allows machine to express and understand semantic relationship. So we will attempt the application of ontology and linked data in bibliography organization.

Problem Section analyses the problems and localizations of information organization and retrieval systems by some OPAC cases and defines the issues of this study.

Procedure Section proposes basic idea of building semantic linking based on ontology, describing semantic information accordantly with RDFS/OWL and accessing consistently with SPARQL to improve the semantic degree and sequencing extend. Method is the key part of Procedure Section and composed of construction of NSTL bibliography ontology, transformation of organization model and publishing of linked data.

Results Section provides several query procedures and results of semantic bibliography retrieval demo system to show the improving of semantic degree of bibliography organization. Specially, the results demonstrate that it is an effective method to resolve the problems and localizations of traditional bibliography organization.

2 Problems

2.1 Problem Statement

Traditional bibliography organization method provides a liner and single-dimensional organization mode based on MARC (Machine Readable Catalog) and relationship database. The advantage is high structural degree with fields and sub-fields to markup characters of physical and content description. But its disadvantage is low semantic degree and low describing commonality. The idea of organization based on MARC provides index points and access points based on MARC fields or subfields, but it do not distinguish information objects extracted from bibliography and do not show the hierarchical or related relationships between them. So there is lacking a mecha-

nism to identify and describe the bibliography relationships and lacking method to sequence bibliography relationships such as the characters of multiplicity of information forms, variability in information life and complexity of hybrid digital object.

The main problem of traditional bibliography organization express as following

1) Lacking method to support the centralization and clustering based on manifestation and expression of bibliography objects;

2) Lacking way to extract information objects;

3) Lacking description between information objects and organization based on relationship of objects;

4) Lacking indication of relationship of deriving and changing of information objects and organization based on changing relationship.

2.2 Case Analysis

(1) Case of multiplicity of information object

Information object has multiplicity forms. For example, the book of Harry Potter and the Goblet of Fire has many different media forms, edited versions (full or brief works), translation forms, binding forms (hard cover or simple package), publication forms (collected works or offprint) and forms of literature (novel, film or audio reading). In Amazon web site, the result of keyword query equaled Harry potter and goblet of fire is a set of 916 records. Although Amazon like other OPACs facets the result in language, media and format and other form property, it is still difficult to find out an audio reading in Arabic quickly. For multiplicity of information objects, facet query is a good method to divide a large set into small sets according to all kinds of multiplicity. But how to organization so many properties of information object is still a problem. It is key issue to realize dynamic facet based on specific query phase and property of specific information object.

(2) Case of variability of information object

Information object has variability. For example, journal Atlantic Monthly has changed its journal title for 5 times, changed ISSN for many times and there are many changes in fields of publication frequency, media forms and publication place etc. In OPAC of Library of Congress, the result of the query of title equals Atlantic Monthly is a set of 47 items, but the result set do not included the another journal Atlantic and do not show the changes and the two different journal titles are same journal. The same query is realized in OCLC WorldCat, the results are better than LC OPAC's, two titles- journal Atlantic Monthly and Atlantic are all hit. Although the recall is higher than LC OPAC, the problem of identity of information object is still not solved.

(3) Case of complexity of information object

Complexity of bibliography mainly represents as multi-volume documents, collection and its single articles, proceedings and its sub-proceedings. It is necessary to describe and organization the complexity and hierarchical relationship. For example, the journal BBA is a collection and includes 9 sub-collections. But in NSTL web site, it is not available to get the sub-collections using collection title or get the collection using title of any sub-collection.[1]

3 Procedures

3.1 Research Design

1) Obtain entity and semantic relationship by distinguishing the levels of bibliography, extract attributes of entity, property of entity and relationship between entities; Describe entity and relationship explicitly and machine-readably using RDF graphic of node and lines;

2) Construct bibliography ontology which provide concept specification and semantic relationship specification using vocabulary set and restriction rules integrating from DC, FRBR, MARC and OAI-ORE etc. to build multidimensional organization mode;[1,2]

3) Construction linking mechanism between bibliography ontology and instance data; Build linked data web with richly structure and semantic data by providing uniform data model, consistent semantic describing method and standard access measure.

3.2 Methodology

(1) Construction of NSTL Bibliography ontology

NSTL Bibliography ontology references following bibliography ontology such as MarcOnt Ontology, Bibliographic Ontology Specification and common vocabularies such as Dublin Core Element Set, RDA Elements, Metadata Management Associates, FRBR Entities, CDLR(Centre for Digital Library Research) and so on. The NSTL bibliography ontology includes 18 classes, 31 object properties and 379 data properties.[3,4]

Referencing the concept model and specification of FRBR, OAI-ORE, DCMI etc., NSTL bibliography distinguishes and extracts bibliography entity as Work, Subwork, Manifestation and Expression, Subject, Agent, Item etc. and abstract them into 18 classes. The relationship between classes is described by object properties (see Table 1) and the attributes of object are described by data properties.

Table 1. Data Property of NSTL bibliography ontology

Relationship Name	domain	Range
isOwnerBy	Item	CorporateBob
isPublisherOf	CorporateBody	Work
isCreatedBy	Work	Person
isSucceedby	Work	Work
IssueedWith	Subwork	Subwork
isPartof	Subwork	AggregateWork
isSubSeriesOf	Work	SupperWork
IsExemplarOf	Item	Manitestation
IsRealizedBy	Expression	Person
isPublishedBy	Work	CorporateBody

978-1-935068-76-1 © 2011 SciRes.

IsCreatorof	CorporateBody	Work
isAReproductuion	Manitestation	Mantiestion
isProducerOf	Person	Manifestation
isRealizedThrough	Work	Expression
isOwnerOf	CorporateBody	Item
hasAReproduction	Manitestation	Manifestation
isSeriesOf	SuperWork	Work
hasAtranslation	Expression	Expression
isProducedBy	Manifestation	Person
isExemplifiedBy	Manifestation	Item
isAReconfigurationOf	Item	Item
isATransiationOf	Expression	Expression
Succeedto	Work	Work
hasReconfiguration	Item	Item

Figure 1(a) demonstrates the hierarchical relationship between main class and subclass, figure1 (b) demonstrates the relationship between classes.

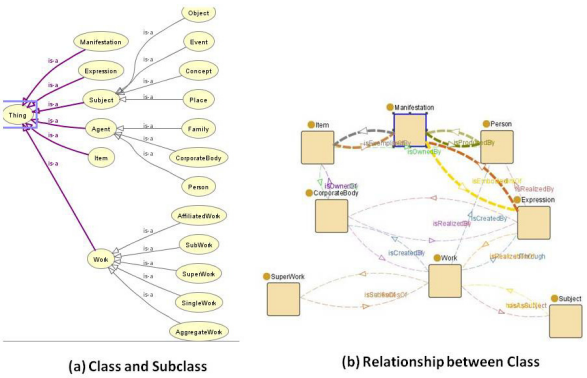


Figure 1. Classes and relationship of NSTL bibliography ontology

(2) Transformation of organization pattern

D2R is a popular open source software to publish linked data. We transform relationship database pattern into semantic organization pattern based on D2R.[5,6]

① Architecture of D2R

D2R Platform consists of three parts:

- D2R Server, an HTTP server that can be used to provide a Linked Data view, a HTML view for debugging and a SPARQL Protocol endpoint over the database.
- D2RQ Engine, a plug-in for the Jena and Sesame Semantic Web toolkits, which uses the mappings to rewrite Jena and Sesame API calls to SQL queries against the database and passes query results up to the higher layers of the frameworks.
- D2RQ Mapping Language, a declarative mapping language for describing the relation between an ontology and an relational data model.

② Operation Mechanism of D2R

D2R Server uses a customizable D2RQ mapping to map relationship database content into this format, and allows the RDF data to be *browsed* and *searched* – the two main access paradigms to the Semantic Web. A D2RQ mapping specifies how resources are identified and which properties are used to describe the resources.

An ontology is mapped to a database schema using d2rq:ClassMaps and d2rq:PropertyBridges. The central object within D2R and the object to start with when writing a new D2RQ map is the ClassMap. A ClassMap represents a class or a group of similar classes of the ontology. A ClassMap specifies how instances of the class are identified. It has a set of PropertyBridges, which specify how the properties of an instance are created.

First we processed NSTL Marc data in relationship database and build table for every entity extracted from single dimension and linear MARC data. For example, see figure2 which shows AggregateWork, SubWork and Subject there entities. The restrictions between the table presents the semantic relationship. The subjects of Aggregate Work come from the table Subject and Aggregate Work is composed by Sub-Work.

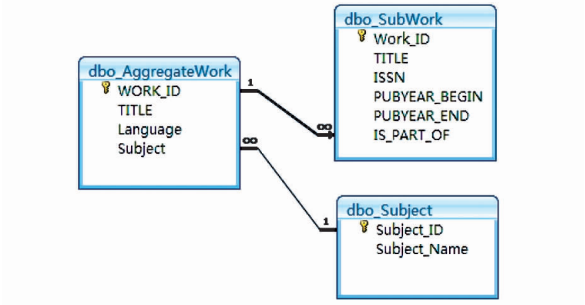


Figure 2. Organization Structure in Relationship Database

Second, we build mapping from relationship database pattern to RDF Schema by mapping files which transfer data structure, restriction into semantic concept and rule. Mapping files are executed by script GenerateMapping.bat provided by D2R Server.[7]

D2rq:ClassMap includes some properties such as d2rq: Class,d2rq:UriPattern;D2rq:PropertyBridge composed by d2rq:belongsToClassMap,d2rq:property, d2rq: column, d2rq:refersToClassMap, d2rq:join. In our mapping file, AggregateWork, SubWork, Subject tables are mapped into classes by ClassMap and Property Bridges of ClassMap prescribe a property-set which is generated by mapping from the columns of tables.

Figure 3 shows the class mapping and property mapping pattern from the relationship database.

(3) Building and publishing linked data

Executing script 2r-server.bat to run the above mapping file and start up the linked data web server which provide three pages of HTML View, RDF View and SPARQL Endpoint. In HTML View, three tables are mapped into three different entities. The class of AggregateWork in page has an instance of work BBA and BBA's attributes including title, language mapped from columns of table in relationship database. And the property SubWork_IS_PART_OF links to all sub-works and

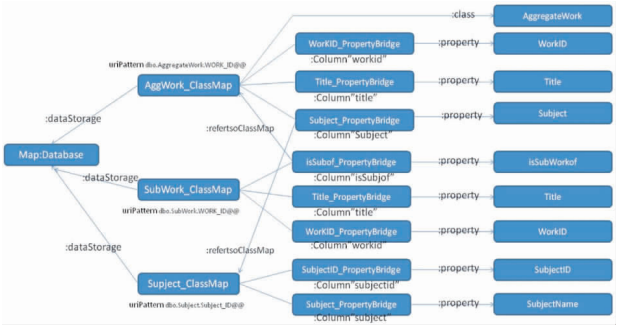


Figure 3. Mapping instance of class and property

the property Subject links to its subjects. So the entity property and entity relationship in ontology are published into RDF and linkage. The SPARQL Endpoint provides a tool to query RDF data using SPARQL language so it can support query and browse based relationships between classes. For example, HasPart can be used as query condition to find out all collection work which has sub-work.[8]

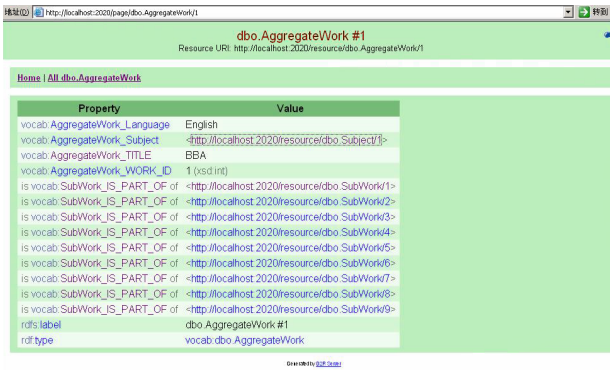


Figure 4. Publishing linked data

4 Results

(1) Demo of Dynamic Facet Query

Traditional facet query is a kind of static query which has characters as following:

1) Property used as facet condition is pre-defined and extracted and indexed beforehand;

2) Do not support to extract property and relationship of data dynamically;

3) Do not support the organization based on properties;

Dynamic facet query method has advantages as following:

1) Extract property or relationship as facet condition in any phase of query;

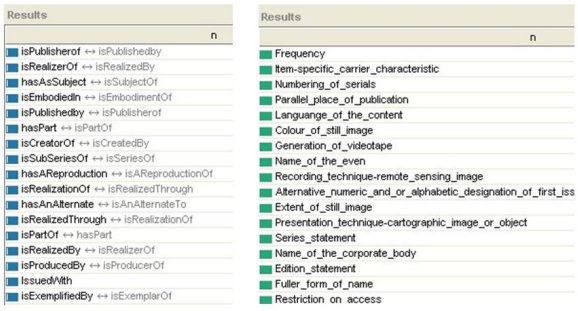
2) Organization property and relationship into hierarchical facet;

Because bibliography data is organized based on ontology and information objects is described by classes and properties, it is easy to extract attributes of a specific class, relationship between classes and it is available to facet dynamically based on specific object in any phase of query.[9]

For example, there is a SPARQL script (see table 2) used to extract all object properties (see figure 5(a)) and data properties (see figure 5(b)) and we can add some filter to limit specific object. Figure 5 shows the results of all property of class Work which can be used as a facet conditions to browse the class of Work.

Table 2. Script of query of object property and data property

```
prefix istic:
<http://www.istic.ac.cn/ontologies/2009/10/5/ontology_bib.owl>
SELECT ?n
WHERE { ?n rdf:type owl:ObjectProperty}
Select ?n
Where {?n rdf:type owl:DatatypeProperty}
```



(a) Object Property (b) Data Property

Figure 5. Demo of Dynamic Facet Query

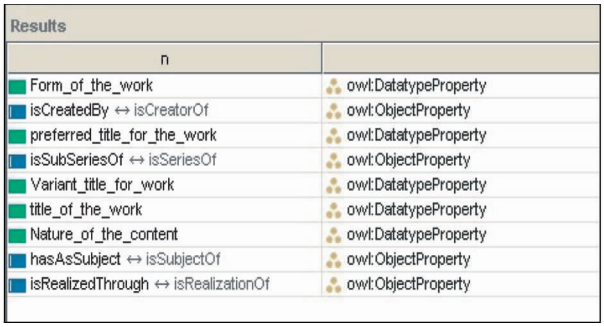


Figure 6. Result of property of class Work

(2) Demo of Semantic Relationship Query

Traditional query is based on the attributes of object such as title, author, ISBN came from index to field of MARC data. Because bibliography ontology expresses the relationship between objects using formalized describing from ontology, it is available to query based on object property and browse relationship between objects.[10]

For example, there is a script list to query all aggregates works and their sub-works. The query keyword is 'hasPart'. The query results are journal BBA and its sub-work. Figure 7 and Figure 8 demonstrate the relationship between BBA and its 9 sub-works.

978-1-935068-76-1 © 2011 SciRes.

Table 3. Script of relationship query (1)

```
prefix                                                    istic:
<http://www.istic.ac.cn/ontologies/2009/10/5/ontology_bib.owl>
SELECT ?title ?sub_title
WHERE {?n :hasPart ?o.
?n :title_of_the_work ?title.
?o :preferred_title_for_the_work ?sub_title.}
```

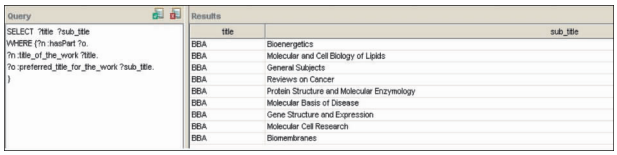


Figure 7. Result of relationship query

Another query example (See Table 4) is about evolvement of Japanese journal whose title is Journal of Anatomy. This journal changed and succeeded to English version. The result shows the start publication of the journal is 1928 and it began to publish English version and Japanese version in 2005. Figure 9 shows the visualization of change relationship.

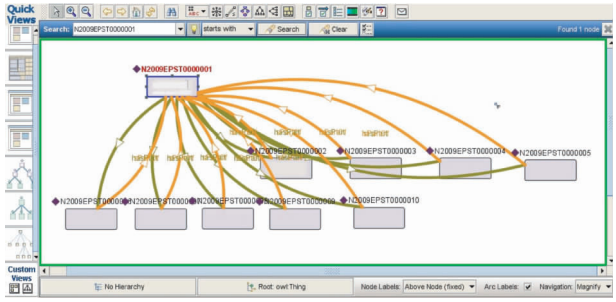


Figure 8. Visualization of relationship of result(1)

Table 4. Script of relationship query (2)

```
SELECT ?title1 ?date1 ?Succeed_to ?title2 ?date2
?isRealizedThrough ?othertitle ?language
WHERE {?n :Succeed_to ?o.
?n :title ?title1.
?o :title ?title2.
?n :data ?date1.
?o :data ?date2.
?o :isRealizedThrough ?other.
?other :title ?othertitle.
?other :Language_of_expression ?language}
```

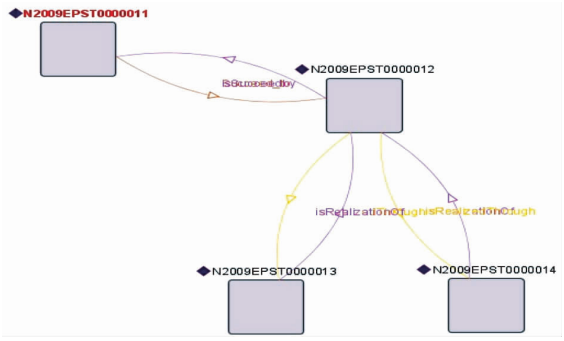


Figure 9. Visualization of relationship of result (2)

(3) Demo of Complex Relationship Query

Based on relationship between objects and properties and characters of SPARQL, complex relationship query can be realized. SPARQL is a query language for pattern matching against RDF graphs, so it support relationships matched and combined and it is helpful to complex relationship query. For example, there is a request to find out journal whose start publication has over 50 years and publish company has history over 100 years which published in field of science, technology and medical science. This journal is about medical science and has electronic and print versions. This request query includes many relationships match and pattern match. Table 5 shows the query script.

Table 5. Script of complex relationship query

```
SELECT ?s ?o ?p ?d ?n ?x ?media
WHERE { ?s :isPublishedby ?o.
?s :pubdate ?y.
?s :hasAsSubject ?subject.
?subject :Term_for_the_concept ?n.
?o :Name_of_the_corporate_body ?p.
?o :Date_of_establishment ?d.
?s :isEmbodiedIn ?x.
?x :Media_type ?media.
filter (?d<= '1910' && ?y <='1960' &&
( regex(?n,"Science","i")
 || regex(?n,"Medical","i")
  || regex(?n,"Surgery","i") || regex(?n,"Agriculture","i"))
 &&                regex(?media,"print","i")                ||
regex(?media,"electronic","i") )}
```

The result is shown as Figure 10 that ID N2009EPST0000015 of a journal was published by Elsevier Company which was created in 1900 and its subject is Surgery and it has two Manifestations of print and electronic version.

Figure 10. Result of complex relationship query

5 Conclusion

This study experiments the application of ontology and linked data in bibliography organization. The basic idea of this paper focus on deeply sequencing of information and the methods emphasize on building semantic linking based on ontology, describing semantic information unitedly by RDFS/ OWL and accessing conformably using SPARQL. The demos of some queries has proved the techniques we applied can improve the semantic degree and sequencing extent and overcome the limitation of traditional organization in describing syntactic rules and data model of organization.

In further work we will attempt deeply sequencing techniques in wider scope, for example, the organization of merging data from integrated systems, the organiza-

tion of knowledge data based on data nodes and data relationships and so on.

References

[1] The OAI-ORE Interoperability Framework in the Context of the Current Scholarly Communication System. [EB/OL].[2010-02-20].http://www.slideshare.net/hvdsomp/the-oaiore-interoperability-framework

[2] Eric Childress.FRBR and OCLC Research. [EB/OL].[2010-02-20].www.oclc.org/research/presentations/childress/20060410-uncch-sils.ppt.

[3] Integration Ontology for Bibliographic Description Formats. [EB/OL].[2010-03-25]. http://www.marcont.org/marcont/pdf/DC2005skmskz.pdf.

[4] Marcin Synak , Sebastian Ryszard Kruk .MarcOnt Initiative the Ontology for the Librarian World. [EB/OL]. [2010-03-25]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.114.3986&rep=rep1&type=pdf.

[5] Eric Prud'hommeaux,Andy Seaborne.SPARQL Query Language for RDF. [EB/OL]. [2010-04-05]. http://www.w3.org/TR/rdf-sparql-query/

[6] Chris Bizer, Richard Cyganiak, Tom Heath. How to Publish Linked Data on the Web. [EB/OL].[2010-02-22]. http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/20070727/

[7] Chris Bizer, Richard Cyganiak. D2R Server, Publishing Relational Databases on the Semantic Web. [EB/OL]. [2010-06-16]. http://www4.wiwiss.fu-berlin.de/bizer/d2r-server/index.html.

[8] Martin Malmsten. Making a Library Catalogue Part of the Semantic Web [EB/OL].[2009-11-12]. http://dcpapers.dublincore.org/ojs/pubs/article/view/927/923

[9] Zeng Lei, Linked libray data. [EB/OL].[2010-02-22]. http://202.114.9.60/dl6/pdf/26.pdf.

[10] Chris Bizer, Richard Cyganiak.The D2RQ Platform v0.7 - Treating Non-RDF Relational Databases as Virtual RDF Graphs. http://www4.wiwiss.fu-berlin.de/bizer/d2rq/spec/#specification.

978-1-935068-76-1

Study on Fusion Classification Model of WEB Page on Multi-Media Information

Xiaodan ZHANG, Bing LIANG, Li WANG, Haiyan BAI, Shijiong LV, Jing XIAO, Ziqi ZHOU
Institute of Scientific and Technical Information of China, Beijing, China, 100038
Email: zhangxd@istic.ac.cn

Abstract: For higher WEB page classification precision, a general feature layer fusion classification model and algorithm are proposed, which based on model theory of information fusion, adopting multi-Media information of the WEB page for different classification, text and image information are used in the paper. The model includes two layers, one is feature layer, which deals with different Media information with different classification algorithm, and inputs the classification results into the higher layer fusion centre separately. The other is decision layer, which deals with the results from the feature layer, and concludes the final classification result. The experiment expresses the fusion model can improve the text classification precision effectively.

Keywords: web claasification; feature layer fusion model; multi-media; information fusion

1 Introduction

WEB page classification is the important hot problem in the field of data mining, which is a process to distribute a WEB page into a given class correctly. Although there are many sophisticated classification algorithms, such as KNN, SVM, and Bayes, but how to improve the WEB page classification precision is the main study topic now[1,2].

On the WEB page, multi-Media information is more and more, such as, text, image, audio, video and etc., how to utilize them to improve WEB page classification precision is the important problem in this paper.

Based on the model theory of information fusion, a general feature layer fusion classification model is proposed, which includes two layers, one is feature layer, which deals with different Media information with different extracting method, pre-proceeding method and classification algorithm, finally inputs the different classification results into the higher layer. The other is decision layer, which processes the results of the feature layer, and inference the final classification result.

Text and image are the general information on the WEB page[3-6], which are adopted in the fusion algorithm and the experiment. The fusion algorithm is proved that the fusion model can improve classification precision effectively than other usual WEB page classification algorithms

2 Feature Layer Fusion Classification Model

Information fusion is the assessment process, which deals with kinds of information with kinds of methods to get more precise assessment. In fusion model theory, there are three kinds of fusion layer and two kinds of fusion structures, which are data layer, feature layer, decision layer and series connection structure, parallel connection structure.

WEB page classification can be seen as an assessment problem, which is a process of classifying a WEB page into a given class correctly[7-12]. Based on fusion model theory, a feature layer fusion classification model with multi-Media information is built showed as Figure 1.

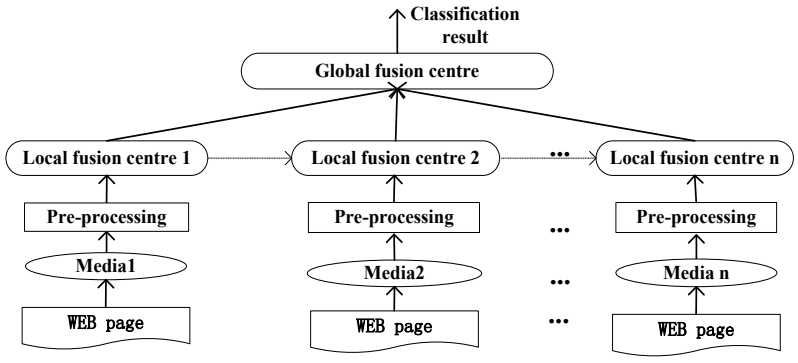


Figure 1. Multi- Media information fusion classification model

From Figure 1 we can see the multi-Media information fusion classification model has two layers, feature layer and decision layer, named local fusion center and global fusion center too. The specific process includes information extract, pre-processing, local classification and the global classification. After the different type information, for example, text, image, audio and video, is extracted from Web pages respectively, they are pre-processed by different method, and the processed data are input into the corresponding local fusion center. The different local fusion center accomplished different media information classification. The classification results of feature layer are input into the global fusion centre and the associated local fusion center.

The global fusion center gets final classification after all local fusion results input into it. And the final classification result is output from the fusion model.

3 Feature Layer Fusion Classification Algorithm

Text and image information are the general resource on the WEB page, so we select them as the Media data of the fusion model. For determining feature layer classification algorithm, we contrast KNN, SVM, Bayes and BP Net with the same image training data set and text training data set under the same experiment condition. Experiment data is the same as that of the chapter 4 of the paper. From the experiment, KNN is exceeding to the other algorithm in text classification, and SVM is prior to the other algorithm in image classification. So KNN and SVM are as the feature layer algorithms of the fusion model to deal with text and image separately. D-S Evidence Theory is adopted as the decision layer fusion algorithm in the fusion model，which deals with the feature layer results of SVM and KNN.

The fusion classification algorithm is showed in Figure 2,

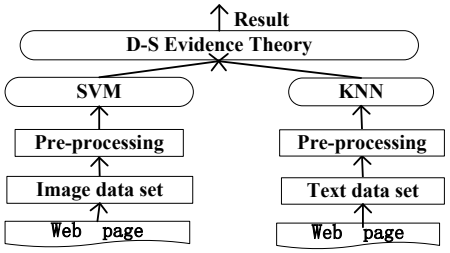


Figure 2. Feature layer fusion classification algorithm

From Figure 2 we can see, KNN and SVM are adopted as feature layer fusion algorithm, processing text and image data classification separately, and D-S Evidence Theory algorithm is used in decision layer classification, processing the results output from the feature layer.

The steps of the fusion classification algorithm are as followed:

Step 1. Training text data are extracted from the WEB pages, the obtained data are pre-processed, and input into KNN, and then the KNN classifier is determined.

Step 2. Training image data are extracted from the WEB pages, the obtained data are pre-processed, and input into SVM, and then the SVM classifier is determined.

Step 3. The real text data are input into KNN classifier after extracted and pre- processed, the text classification result is got from KNN.

Step 4. The real image data are input into SVM classifier after extracted and pre- processed, the image classification result is got from SVM.

Step 5. The results of **Step 3** and **Step 4** are input into decision layer, which is D-S Evidence Theory algorithm.

Step 6. The final classification result is output from D-S Evidence Theory.

The pre-processing of image data includes image denoising, feature extraction and vector express. And the pre-processing of text data includes segmentation, feature extraction, weight calculation and vector express.

If $X = \{X^{(k)} \mid k = 1,2,\cdots,N\}$ is the non-empty mode set of n -dimensional vector space $U \subset R^n$, $S = \{S^{(m)} \mid m = 1,2,\cdots,c\}$ $(1 \le c \le N)$ is a subset family of X, if meeting the formula (1):

$$\begin{cases} S^{(m)} \neq \phi \\ S^{(i)} \cap S^{(j)} = \phi, (i \neq j) \\ X = \bigcup_{i=1}^{c} S^{(i)} \end{cases} \tag{1}$$

Then m can be got according to combination the initial belief function *BEL* is the classification result.

If the size of true value is reduced, or the true value is found, then the true value can be got by Minimum point rule.

4 Experiment

The feature layer fusion classification model is realized in JAVA development plot. To test the text classification precision, we contrast KNN, SVM and the feature fusion method with the same training data set and testing data set.

The text dataset is from Sogou net station. Randomly 6 classes are chose as dataset, including Education, Computer, Environment, Traffic, Economy, Military affairs. 1049 documents are as training, and 520 documents are as testing set.

The image dataset in the experiment is from the Ground Truth Database Team of Object and Concept Recognition for Content-Based Image Retrieval of University of Washington. The image format is JPG. The image database is departed into 6 classes, that are Education, Computer, Environment, Traffic, Economy,

978-1-935068-76-1 © 2011 SciRes.

Military affairs, 1700 images in total. And every class is departed into training set and test set, which Proportion is 2:1.

For Assurance the effectiveness of the experiment, we choose the images data as the followed rules.

1, the image needs maximum for describing the image feature, and being easy to distinguish from other image.

2, the image of the same class is not similar to each other, and the image should be more as more kinds.

The contrast result of the three methods is showed as Table 1.

Table 1. Contrast of three algorithms

Methods\ precise	Recall rate	precision	F1
KNN	80.3%	90.2%	84.8%
SVM	76.4%	92.6%	83.7%
Fusion model	85.6%	95.2%	90.7%

In Table 1, we can see the precision and recall rate of the fusion method higher than the other classification methods.

From Figure 3, we can see, the precision of the fusion method is superior to KNN. When the number of the sample is 5.5 and 6.4, with the number of samples is more, the precision of the two methods is the higher, and the precision of the fusion method is higher than the KNN.

From Figure 4, we can see, the precision of the fusion method is superior to SVM with image date set.

The precision of the fusion method is superior to SVM. When the number of the sample is 2, 4.4 and 6.2, with the number of samples is more, the precision of the two methods is the higher, and the precision of the fusion method is higher than the SVM.

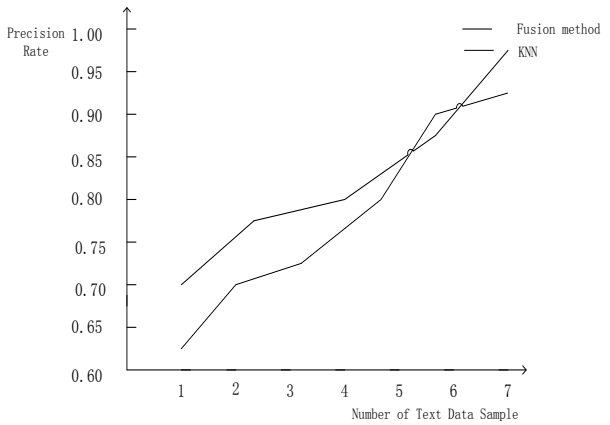


Figure 3. Contrast of the fusion algorithm and KNN algorithm with text data set

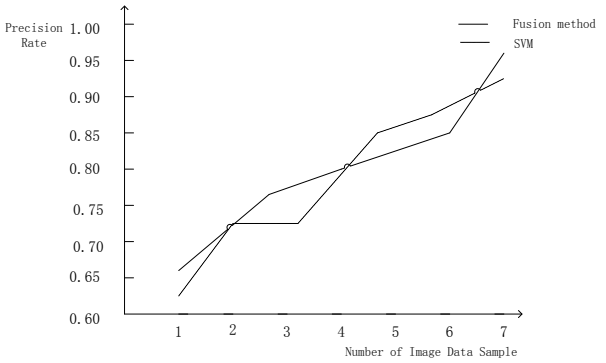


Figure 4. Contrast of the fusion algorithm and SVM algorithm with image data set

5 Conclusion

With the development of Media information on the net, how to use the information to improve the precision of WEB page classification is the main study point in the paper. According to the model theory of information fusion, a general feature layer fusion classification model is proposed, which adopts multi- Media information on the feature layer, and their classification results are input into decision layer. Then the final classification is output from the global fusion center. Text and image information are adopted in the feature layer fusion classification algorithm because of their generality on the net. At last, the fusion algorithm is realized and done comparison with other classification algorithms in the experiment. From the contrast, we can conclude that the fusion model proposed in the paper can improve the WEB page classification precision effectively.

Acknowledgment

The author acknowledges the support of the Natural Science Foundation 60803050 of P. R. China.

References

[1] Zhang, B., Chen, Y., Fan, W., Fox, E. A., Goncalves, M., Cristo, M. & Calado, P.(2006a). Intelligent GP fusion from multiple sources for text classification. In Proceedings of the 14th ACM international conference on Information and knowledge management (pp. pp. 477-484). : ACM Press, Bremen, Germany.

[2] Zhang, G. P. (2007). Avoiding Pitfalls in Neural Network Research. Systems, Man and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on, 37, pp. 3-16.

[3] Zhang, L., Zhu, J. & Yao, T. (2004). An evaluation of information fusion techniques. ACM Transactions on Asian Language Information Processing (TALIP), 3,pp. 243-269.

[4] Zhang, Y., Zincir-Heywood, N. & Milios, E. (2005c). Narrative fusion classification for automatic key phrase extraction. In Proceedings of the 7th annual ACM international workshop on Web information and data management (pp. pp.51-58). : ACM Press, Bremen, Germany.

[5] Zheng, Z. & Webb, G.I. (2000). Lazy Learning of Bayesian Rules. Machine Learning, 41, pp. 53-84.

[6] Xu L Y, Du Q D. Application of neural fusion to accident forecast in hydropower station. Proceedings of the Second International

Conference on Information Fusion. Vol2 Sunnyvale, 1999, pp. 1166-1171.

[7] Schapire, R. E., Singer, Y. & Singhal, A. (1998). Boosting and Rocchio applied to text filtering. In Proceedings of SIGIR-98 21st ACM International Conference on Research and Development in Information Retrieval (pp. p. 215--223). : ACM Press New York US.

[8] Sebastiani, F. (2002). Machine learning in automated image categorization. ACM Computing Surveys, 34, pp. 1-47.

[9] Shih, L. K. & Karger, D.R. (2004). Using urls and table layout for web classification tasks. In Proceedings of the 13th international conference on World Wide Web (pp. pp.193-202). : ACM Press, New York, NY, USA.

[10] Rish, I. (2001). An empirical study of the naive Bayes classifier on image. In IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence (pp. pp. 41-46). : T.J. Watson Research Centre, Seattle, Washington.

[11] Riloff, E. (1995). Little words can make a big difference for text classification. In Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval (pp. pp. 130-136).: ACM Press, Seattle, Washington, United States.

[12] Zhang Xiao-dan. A system of file Automated classification[P], 201020200043, China: 2010.12.

978-1-935068-76-1

Two-Pass Based System Combination for Machine Translation

Yanqing HE, Huilin WANG
Information Technology Support Center, Institute of scientific and Technical Information of China, Beijing, China
Email: heyq,wanghl@istic.ac.cn

Abstract: Nowadays Multiple system combination has been a hot topic in machine translation, which combines the outputs of different machine translation systems to get a better translation performance. System combination usually operates on three levels: sentence, phrase and word level. In this paper we give a new approach of system combination based on two-pass. In the first pass we generate the word alignment using the method of word level system combination. A phrase table is extracted from the word alignment. In the second pass we implement a phrase-based re-decoding similarly to phrase level system combination. In this way we can guarantee some phrases with greatest possibility of candidate translations. Furthermore we can use more features to select the final translation. We test the method on two test sets and illustrate the effectiveness of the proposed mechanism.

Keywords: machine translation; system combination; phrase extraction; two-pass

1 Introduction

In recent years, there have been a lot of paradigms for machine translation, ranging from example based, rule based, phrase based, hierarchical and syntax based machine translation. Each translation model can be associated with various kinds of information and present the different characteristic. Even one translation system's performance will have significant difference when it uses difference parameters. So it's hard to say which translation model has overwhelming advantage on others. The idea of combining the outputs of multiple machine translation systems to get a better performance is a cleaver choice.

Multiple translation systems combination usually operates on three levels: sentence, phrase and word level. Sentence level system combination [1] is a re-ranking of translation results which employ extra features to score all the translation outputs from multiple systems and choose the result with highest score as its combination output. This kind of approach integrates new features other than those used in each translation system to help choosing translation. It doesn't generate new translation hypothesis and has many parameters to be estimated. Phrase level system combination [1] employs the word alignment between source sentence and each candidate translation to re-extract phrase pairs. Based on the new phrase table it re-decodes to generate its final translation result, which is similarly to a phrase-based statistical machine translation system. This kind of system combination generates new translation hypotheses but it is hard to guarantee the quality of the phrase table. Nowadays word level combination approaches using network [2,3] are widely adopted. They introduce Minimum Bayes Risk decoding [4] to find a hypothesis as its backbone from all the translation hypotheses, then construct consensus network by aligning all the other translation hypotheses against the backbone, finally generate a path with highest score from the consensus network. One major difference in these approaches is how to obtain the word alignment between its backbone and a hypothesis translation. Ref. [5] adopts WER (Word Error Rate) to align them. Some heuristics measures [6] are used to modify [4]. In [7] alignment models are trained by considering all hypothesis pairs as a parallel corpus using GIZA++ [8]. Translation edit rate (TER) based alignment is proposed in [4]. In [9] arbitrary features are added log-linearly into the objective function, thus allowing language model expansion and re-scoring. A couple of other approaches, such as the Indirect Hidden Markov Model [10] and the Incremental HMM alignment [11], are recently proposed with better results reported. In short word level system combination guarantees the maximum probability of newly generated translation hypothesis but the result may be ungrammatical due to alignment errors.

We integrate multiple machine translation systems in a translation platform and build a system combination system based on two-pass. In the first pass, we generate the word alignment using the method of word level system combination and a phrase table is extracted from the word alignment. In the second pass we implement phrase-based re-decoding similarly to phrase level system combination to get the final translation. In this way we can guarantee some phrases with greatest possibility of candidate translations. Furthermore we can use more features to help choosing the final translation. This method is similarly to [12]. The chief difference is [12]'s phrase table is a subset of the phrase table here. Its phrase table is in fact a diction-

The research work has been partially funded by Discipline Construction Research Project (NO. XK2010-6), Key Task Project (NO. ZD2010-3-3) and Pre-Research Project (NO. YY-2010020) of ISTIC in 2010.

ary which source phrase and target phrase have only one word. In this paper we don't need to construct confusion network but extract a larger phrase table from the word alignment similarly to a phrase-based statistical machine translation. Our phrase table is not between different languages. We test the two methods on two test sets and illustrate the effectiveness of the proposed mechanism.

The remaining of this paper is organized as follows: First the key algorithm of our method of system combination is described in section 2. Experimental results are presented in section 3. Finally section 4 gives the conclusion.

2 Description of the Algorithm

Our framework of system combination is given in Fig. 1. We take two passes to get the final combination result. In the first pass we generate the word alignment between the backbone and other candidate hypotheses by using the method of word level system combination. We extract a phrase table from the word alignment. Finally In the second pass a phrase-based re-decoding is implemented to find the output translation.

2.1 First Pass

Five translation systems are collected to obtain candidate translations: three rule-based translation engines and two statistical translation engines (a phrase-based and a hierarchical statistical machine translation system). In order to guarantee a more robust combination result we choose 1best translation from the system with the best performance among three rule-based translation systems as the backbone. After getting all the translation results from the five systems, the word alignment between the backbone and all the other translation hypotheses is implemented. Here GIZA++ is used to train word alignment. Grow-Diag-Final heuristics are used to extend the result of GIZA++ to get the final word alignment. Please refer to [12] to find the detail of building the word alignment.

From the word alignment between the backbone and each candidate hypothesis, we extract a phrase table.

Fig.2 gives the detail of extracting algorithm. For the English-to-Chinese translation task, our source sentence is in English and the backbone is in Chinese. In the word alignment a_n for the backbone and hypothesis translation h_n, we collect all aligned phrase pairs that are consistent with a_n: The words in a legal phrase pair are only aligned to each other, and not to words outside. For each index in the backbone we enumerate and check all the possible target phrases. The algorithm is similarly to [13]. Here we don't takes into account possibly unaligned words at the boundaries of the target phrases. In Fig. 2, $h_{n\,j_1}^{\;\;j_2}$ is a word sequence whose position is from j_1 to j_2 in the n[th] candidate hypothesis and $e_{i_1}^{i_2}$ is a word sequence whose posi-

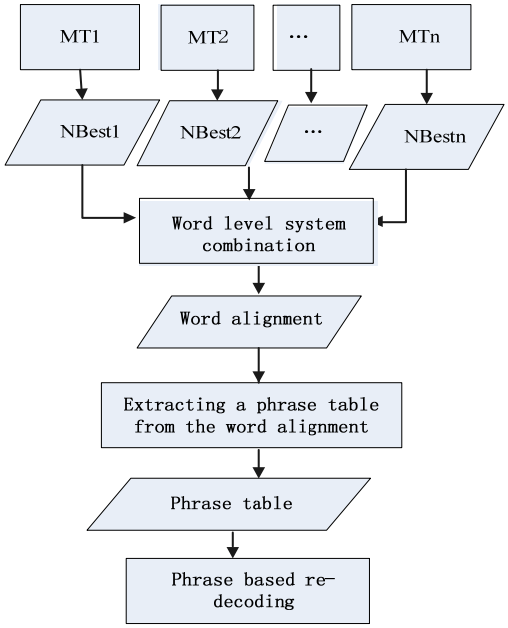


Figure 1. Framework of our algirithm

Input: the Backbone $c_1c_2\cdots c_I$, candidate hypotheses $\{h_1,h_2,\cdots,h_N\}$ and their word alignments $A=\{a_1,a_2,\cdots,a_N\}$.

Output: Phrase table P

1: $P=\varnothing$;
2: for (each a_n in A) {
3: while ($i_1 \le I$) {
4: $i_2 := i_1$;
5: while ($i_2 \le I$) {
6: $TP := \{j \mid \exists i : i_1 \le i \le i_2 \wedge (i,j) \in a_n\}$;
7: $j_1 := \min(TP)$;
8: $j_2 := \max(TP)$;
9: $SP := \{i \mid \exists j : j_1 \le j \le j_2 \wedge (i,j) \in a_n\}$
10: if($SP \in \{i_1, i_1+1, \cdots, i_2\}$) {
11: $P := P \cup \{(e_{i_1}^{i_2}, h_{n\,j_1}^{\;\;j_2})\}$;
12: add 1 to its frequency;
13: }
14 }
15: }
16: }
17: for (each phrase pair $pp \in P$){
18: calculate $prob(pp) = \dfrac{frequency(pp)}{\sum_{pp_s = e_{i_1}^{i_2}} frequency(pp)}$;
19: }

Figure 2. Algirithm for extracting phrases

978-1-935068-76-1 © 2011 SciRes.

tion is from i_1 to i_2 in the backbone. pp_s means the source phrase of a phrase pair pp. The probability of each target phrase is calculated according to its frequency. Each target phrase's frequency in phrase pair is calculated as its posterior probability by voting. Identical phrase is counted repeatedly.

2.2 Second Pass

In the second pass we use a phrase-based re-decoding to get the final translation which is similarly to a phrase level system combination. Here the source sentence is the backbone b. A log linear model is executed to search for the target translation $c*$ with highest probability:

$$c* = \sum_{k=1}^{K} \lambda_k h_k(b,c)$$

where $h_k(b,c)$ is the feature function， λ_k is the corresponding weight. The features are listed as follows:

- Posterior probability of phrases;
- Language model;
- Distance-based phrase reordering model;
- Word penalty.
- Phrase penalty.

Here the language model feature is calculated as [14]. The phrase reordering feature is easily modeled as a distance-based probability: $P_{d(a_k - b_{k-1})} = |a_k - b_{k-1} - 1|$ where a_k is the starting position of the source phrase for the kth target phrase and b_{k-1} is the ending position of the source phrase for the k-1th target phrase. The word penalty feature is the size of the words and the phrase penalty feature is the size of the phrases in the sentence. A beam searching is implemented to find the 1best translation for combination output. We perform the maximum BLEU training [15] on a development set to train the feature weighs.

Our combination strategy in the first pass is different from word level system combination in 1) We don't construct confusion network but extract a phrase table from the word alignment 2) Our decoding algorithm can use more features than confusion network decoding although the features used in this paper may be not more than those in confusion network decoding. Different from phrase level system combination, the source side and the target side of our phrase table is in a same language. In such way we can guarantee the maximum possibility of some target positions' candidate translations.

3 Experiments

This section gives the experiments with English-to-Chinese translation task in science technical domain. Our evaluation metric is case-insensitive BLEU-4 [16].

We use two test sets to demonstrate the performance of our system combination. Each sentence has 4 references. The first test set is from computer science domain. Some sentence pairs in the second test set come from computer science domain, but some come from news domain. 0.9 million parallel sentences from computer science domain, 1,210,050 parallel sentences from news domain and a dictionary with 70850 entries comprise the training data. The statistics of all the data is listed in Table 1 where A.S.L. is average sentence length. SRILM toolkit [14] is used to train a trigram language model with the English sentences in the training data.

Table 1. Coupus statistics

Set	Language	Sentence	Vocabulary	A.S.L
Training set	*Chinese*	2,180,900	215081	20.1
	English	2,180,900	266514	22.4
Test set 1	*Chinese*	6364	6223	21.7
	English	1591	5024	21.6
Test set 2	*Chinese*	984	2677	26.1
	English	246	2302	22.6

Three rule-based, one phrase-based, one hierarchical system is combined in the experiments to illustrate the performance of our method. All systems are tuned on a development set with 4 references. After collecting the five 1-best translation results from all translation engines, we implement two-pass system combination (TPSC). Here the result from rule-based translation engine Rule-System1 is selected as the reference translation. We compare our method against a baseline system combination (WPSC) [12].

Table 2 lists the translation performance of the five translation engines and two system combinations. It can be seen that both TPSC and WPSC improve the translation performance of the five translation engines. Compared with the best system SMTS2, TPSC yields about 2.71% gain in BLEU in the first test set and the BLEU score is improved by 1.88 points in the second test set. It also shows that TPSC outperforms WPSC in both test sets. In the first test set TPSC yields about 1.45% gain in BLEU. In the second test the BLEU score is improved by 0.27 points.

Table 2. Translation Performance

System	Test set 1		Test set 2	
	BLEU%	NIST	BLEU%	NIST
Rule System 1	33.03	9.0305	28.87	7.9905

Rule System 2	22.14	7.7995	22.89	7.2995
Rule System 3	21.12	7.6577	21.34	6.9344
SMTS1	27.98	8.8290	23.63	7.5931
SMTS2	33.94	9.7105	29.77	8.4171
WPSC	34.36	9.7131	31.38	8.4296
TPSC	34.86	9.8364	31.65	8.9842

The analysis of translation results shows that our method of system combination improves the translation results from each system involved. The reason is that some phrase in the reference hypothesis will have more translation candidates by extracting the phrase table from word alignment. Phrase-based re-decoding integrates more features, such as language model and phrase reordering model to ensure the translation performance.

4 Conclusions

In this paper we propose a novel approach of system combination based on two passes, which generate a phrase table from the word alignment in the first pass, then employ a phrase-based re-decoding to get the final combination translation. Our phrase table effectively extends the possibility of the candidate phrase for each phrase in the backbone. In the re-decoding more features can be integrated to choose the final translation. We test the method on two test sets and illustrate the effectiveness of the proposed mechanism.

Our approach still has some weaknesses. There is much room for further improvement. The strategy of choosing backbone is too simple to guarantee a good translation order. We only select the best rule-based translation. Our word alignment still needs to be improved. GIZA++ strongly depends on the size of parallel corpus. This may produce a bad word alignment. We only use five features in the phrase-based re-decoding. More features will be added to implement the re-decoding, such as syntactical features. And also we will try to incorporate other systems into our combination to improve the translation performance.

References

[1] Antti-Veikko I. Rosti, Necip Fazil Ayan, Bing Xiang, Spyros Matsoukas, Richard Schwartsz and Bonni J. Dorr. Combining outputs from multiple machine translation systems. In *Proc. of NAACL-HLT,* 2007, pages 228-235.

[2] Jonathan.G. Fiscus. A post-processing system to yield reduced word error rates: recognizer output voting error reduction (ROVER). *IEEE Workshop on Automatic Speech Recognition and Understanding*, 1997, pages 347-354.

[3] Holger Schwenk, Jean-Luc Gauvain. Improved ROVER using language model information. *ISCA ITRW Workshop on Automatic Speech Recognition: Challenges for the new Millenium*, 2000, pages 47-52.

[4] K.C. Sim, W.J. Byrne, M.J.F. Gales, H. Sahbi and P.C. Woodland. Consensus network decoding for statistical machine translation system. In *IEEE International Conference on Acoustics, Speech, and Signal Processing*, 2007, pages 105-108.

[5] Srinivas Bangalore, German Bordel, Giuseppe Riccardi. Computing consensus translation from multiple machine translation systems. In *Proceedings of ASRU*, 2001, pages 351-354.

[6] Shyamsundar Jayaraman and Alon Lavie. Multiengine machine translation guided by explicit word matching. In *Proc.EAMT*, 2005, pages 143-152.

[7] Evgeny Matusov, Nicola Ueffing, Hermann Ney. Computing consensus translation from multiple machine translation systems using enhanced hypotheses alignment. In *Proceedings of EACL*, 2006,33-40.

[8] Franz J. Och and Hermann Ney. 2003. A systematic comparison fo various statistical alignment models. Computational Linguistics, 29(1): pages 19-51.

[9] Antti-Veikko I. Rosti, Spyros Matsoukas and Richard Schwartz. Improved word-level system combination for machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 2007, pages 312-319.

[10] Xiaodong He, Mei Yang, Jianfeng Gao, Patrick Nguyen and Robert Moore. Indirect-HMM-based Hypothesis Alignment for Combining Outputs from Machine Translation Systems. In Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, 2008, pages 98-107.

[11] Chi-Ho Li, Xiaodong He, Yupeng Liu and Ning Xi. Incremental HMM alignment for MT system combination. In *Proceedings of the 4th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP*, 2009, pages 949-957.

[12] Yanqing He, Junsheng Zhang and Huilin Wang. “Combining Multiple Translations based on Words and Phrases”. Journal of the China Society for Scientific andTechnical Information, in press.

[13] Franz Josef Och and Hermann Ney. The Alignment Template Apporach to Statistical Machine Translation. *Computational Linguistics*, vol. 30 (2004), pp. 417-449.

[14] Stolcke Andreas. SRILM-an extensible language modeling toolkit. In *Proceedings of International Conference on spoken language processing*, Volumn 2, 2002, pages 901-904.

[15] Franz Josef Och. Minimum error rate training instatistical machine translation. In Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1. Sapporo, Japan, 2003.

[16] Kishore Papineni, Salim Roukos, Todd Ward and Wei-Jing Zhu. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the ACL, 2002, pages 311-318.

978-1-935068-76-1

Multiple Perspective Research Profiling: Illustrated for Dye-Sensitized Solar Cells

Alan L. PORTER[1,2], Tingting MA[2,3], Ying GUO[2,3]

[1]*Search Technology, Inc., Norcross GA 30092, USA*

[2]*Technology Policy and Assessment Center, Georgia Institute of Technology, Atlanta, GA 30332-0345, USA*

[3]*School of Management and Economics, Beijing Institute of Technology, 5 Zhongguancun South Street, Haidian District, Beijing, China, 100081*

Email: alan.porter@isye.gatech.edu, matingtingmay@gmail.com, violet7376@gmail.com

Abstract: This paper illustrates how "tech mining" can gain valuable intelligence from Science, Technology & Innovation ("ST&I") information resources. It uses Dye Sensitized Solar Cells (DSSCs) as a case example, analyzing data sets drawn from four major databases. Cross-database comparisons provide insight into stages of development. A range of analyses help answer "who, what, where, and when" questions about such an emerging technology. "What" questions enable a topic of interest. We illustrate how to identify the leading countries and institutions doing R&D to illustrate "who" questions. We also perform "key players by technologies" analysis based on keyword extraction and classification, to determine the leading technologies under development for each dimension and address "who is doing what" questions.

Keywords: TechMining; bibliometrics; knowledge diffusion; science mapping; dye-sensitized solar cells

1 Introduction

Science, Technology & Innovation ("ST&I") information resources present wonderful opportunities to extract knowledge. Such intelligence can help guide research toward the best opportunities and synergistic collaborations. On a more macro level, research profiling can also be tuned to inform institutional or national research policy [1].

But ST&I information also has value "downstream from research toward innovation." Management of Technology concerns many facets of technological development. In the private sector, managers must deal with issues such as intellectual property protection and licensing, open innovation, and mergers & acquisitions. All demand keen understanding of what is happening with respect to related ST&I activities. In serving public sector needs, such knowledge is essential to stimulate innovation while protecting societal interests (e.g., early identification of potential unintended side-effects of technological innovations). Real time technology assessment is a prime focus of one of the projects supporting this research [2]. "TechMining" is our set of tools to help extract useful ST&I intelligence [3].

This paper combines analytical findings from multiple databases – namely, WOS, EI Compendex, Derwent World Patent Index ("DWPI"), and Factiva. By doing analyses in this way we gain richer understanding of ST&I activities.

Our focus is on downstream technological advances and applications. In this paper we draw on data concerning Dye-Sensitized Solar Cells ("DSSCs") – a Newly Emerging Science & Technology ("NEST"). We will sometimes compare DSSC patterns with those for another NEST that we have been studying – Nano-BioSensors ("NBS"). Our interest here is scholarly, not for direct application. We seek to advance tools to enrich ST&I intelligence on an emerging technology. We orient that intelligence (i.e., our analytical questions) toward Forecasting Innovation Pathways ("FIP") [4].

Research knowledge diffusion can be considered in several variations:

a. Among researchers – especially as evidenced by research publications citing others' work
b. To inventors – especially as evidenced by patents citing non-patent research
c. Translational – especially for bio-medical research contributing to clinical medical practice
d. From research to major innovations – as in studies that search back from important technological innovations to identify contributing research

Drawing upon multiple databases enables us to pursue "b," as well as "a" (within WOS). Tracking the flow of knowledge from science to technology is especially important in an era of increasing science-based innovation (e.g., biotechnology, nanotechnology) [5]. Narin et al. [6] found that literature citation by patents increasing mark-

This research was undertaken at Georgia Tech drawing on support from the U.S. National Science Foundation (NSF) through the Science of Science Policy Program —"Measuring and Tracking Research Knowledge Integration" (Georgia Tech; Award No. 0830207), and the Center for Nanotechnology in Society (Arizona State University -- Award Numbers 0531194 and 0937591).

The findings and observations contained in this paper are those of the authors and do not necessarily reflect the views of the National Science Foundation.

edly from 1987 to 1994. Analysts associated with CHI, Inc. examined patent citation of scientific papers for the U.S. National Science Foundation ("NSF") [7].

Our group at Georgia Tech has been compiling extensive sets of nanotechnology ("nano") R&D records from several databases and performing multiple analyses (c.f., http://www.nanopolicy.gatech.edu/) [8]. Since 2009, we have focused on particular sub-topics within nano. This paper reflects an ongoing effort to understand nano-enhanced solar cells. Solar cells can be characterized in three developmental generations [9,10]. First Generation -- "Conventional Solar Cells" -- made in crystalline silicon, account for ~90% of the market, but these are expensive. Second Generation -- "Thin-film Solar Cells" -- can be divided into two groups: silicon thin-film and compound semiconductor thin-film. Third Generation solar cells are classified in different ways. We note two groups: "Compound Semiconductor Thin-film Solar Cells" that employ quantum dots to enhance efficiency, and "Dye-sensitized Solar cells" (DSSCs).

DSSCs, invented by O'Regan and Grätzel, constitute perhaps the most promising and, so far, the most efficient of all solar cells that employ nanotechnology[11,12]. Although DSSC commercialization is still in its infancy, many technical papers anticipate fascinating prospects [11-14]. This situation highlights the complexity of NEST innovation processes. Some emerging technologies are totally new, with no existing markets. Others advance improvements into an existing market. DSSCs are just entering the market (so data are minimal), yet the solar cell market does have a track record.

2 Data and Methods

We chose a modular, Boolean term search approach to identify DSSC technology in four databases -- WOS, EI Compendex, DWPI, and Factiva [8]. Our approach involves four major steps. First, we create search terms to capture a variety of DSSC technical terminology. This shows high retrieval and is defined as our main search term. Second, we enrich those search terms according to different expressions of DSSCs and closely related technical structures. Third, we check retrieval results of those complementary search terms by excluding their retrieval overlap with the main search. We revise the terms by adding limitations – such as adding restriction to select records only if they also appear in certain International Patent Classifications (IPCs) for the DWPI database. We add exclusion terms for the publication databases. Finally, we combine the terms and evaluate them by randomly testing and assessing retrieval results, and then further revise our search terms. We create search algorithms somewhat tailored for each of the four databases based on our approach just described. Data-cleaning methods are then applied to further process the data that are downloaded from the four databases [8]. We also create search terms for NBS using a similar approach. The data are then downloaded to support our occasional comparative analyses with DSSCs.

3 Results

We begin by showing trends based on the annual number of records from WOS (using its Science Citation Index -- SCI), EI Compendex, DWPI, and Factiva. We sought to use this more comprehensive perspective to gain richer understanding of ST&I activities. SCI focuses more on fundamental research compared to Compendex, a prominent engineering-oriented database. DWPI is a patent database, which reflects invention. And Factiva mainly contains business information (e.g., trade magazine items, press releases).

In Fig. 1, it is clear that the research publications drawn from the SCI and Compendex databases keep growing and show similar trends. This suggests that fundamental research on DSSC continues to increase essentially exponentially. Possibly in part because the Japanese government canceled solar energy subsidies in 2006, which affects DSSC commercialization, the data from DWPI and Factiva both show a small peak in 2005 and suddenly decrease in 2006. Actually the data from Compendex also show slower growth in 2006. After 2006, DWPI patents resume growth, although the data in 2009 and 2010 are not completely collected by Thomson Reuters yet. After 2008, the Factiva records suddenly climb quicker than does activity in the other databases. The rapid growth of DWPI and Factiva data suggests that DSSC technology is becoming more mature, and is possibly entering into an era of rapid commercialization.

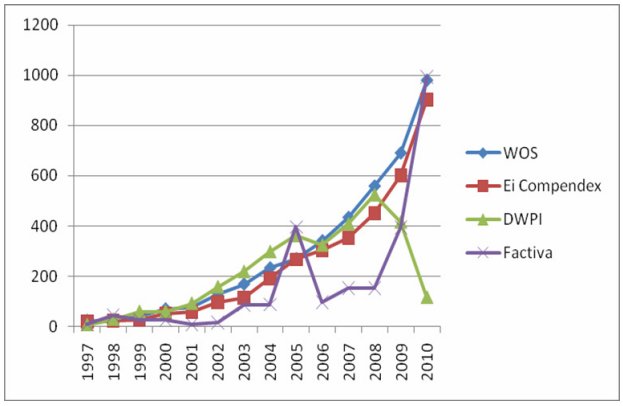


Figure 1. Trends of Annual DSSC Activity

3.1 Characterizing the Research Field

Locating the research among the disciplines: Web of Science data enable us to explore which disciplines contribute heavily to a research field. WOS denotes Subject Categories ("SCs") to which they assign journals. Based on the extent of cross-citation among these, we have developed a science base map, over which we can overlay the publications (or citations) of a target body of

978-1-935068-76-1 © 2011 SciRes.

research.[1] Fig. 2 locates DSSC research activity in 14 "macro-discip- lines."

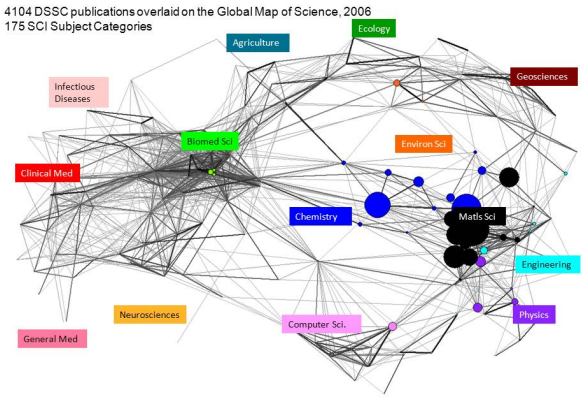


Figure 2. Locating DSSC publications over a base map of science

Research field coherence: We probe the DSSC research field structure further through examination of citations. DSSC research appears to be a highly coherent (tightly connected, unified) field. Here we probe citations by our recent DSSC papers. We find an incredibly dense citation pattern. For the 1691 papers published in 2009 or later, we find 1 reference cited by 987 (58%) of them [15]. Furthermore,

- 23 additional references are cited by at least 100 of the papers (i.e., 6% of the 1691 recent DSSC set)
- 980 references are cited by at least 10 of the papers.

In comparison, for a larger set of 2009-10 NBS papers (4396), we see much less concentrated citation. The leading source is cited 192 times -- i.e., by 4% of the papers vs. 58% for that leading DSSC reference. Not one paper is cited by as many as 6% of the NBS set, whereas 24 DSSC references were cited at least that broadly by the recent DSSC papers.

Fig. 3 and 4 reflect the density of coupling among recently active researchers in the two fields, based on shared bibliography (i.e., references in common).[2] For these research network maps, we have selected 54 DSSC first authors publishing 2 or more papers since 2009 and mapped their similarity in sharing the 980 references cited by 10 or more of the 1691 papers. We contrived a roughly comparable subset of the NBS researchers – 49 first authors with 5 or more papers since 2009, and mapped their similarity based on commonality among them in terms of the 950 references cited by 15 or more (of the 4396 papers). The difference in connectedness is striking.

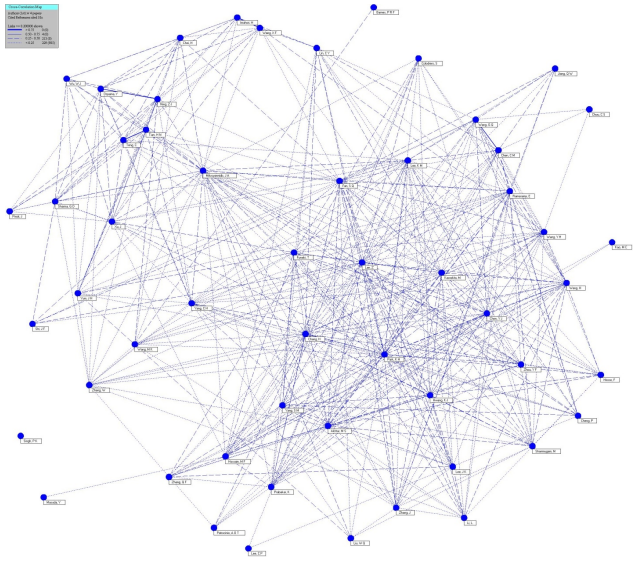


Figure 3. DSSC Research Network Map

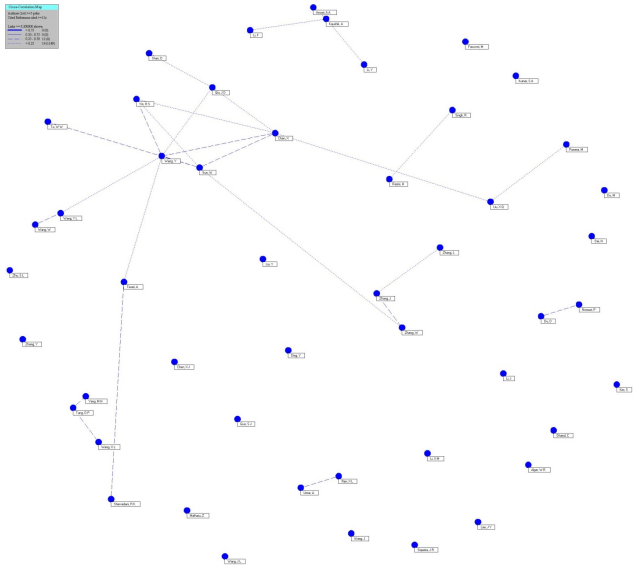


Figure 4. NBS Research Network Map

Both maps are set to show links using the same cutoff of 0.2 link similarity. Counting the resulting links appearing in the maps,

- 48 of 54 DSSC researchers (89%) show with 3 or more connections (Fig. 3), compared to just

[1] Our science overlay maps have been described elsewhere [16-19]. The WOS journal-to-journal cross-citation matrix is converted to SC-to-SC. This map keys on SCI, distinguishing 175 SCs. Factor analyses of that matrix (as cosine values) extracts 14 factors that we label based on the prominent SCs grouped together. For examples and to make your own, see: www.interdisciplinaryscience.net or www.idr.gatech.edu/.

[2] These are Multi-Dimensional Scaling (MDS) visualizations generated by TDA. Location on the X and Y axes has no inherent meaning. More similar nodes are located closer together, but for complete accuracy the rendition would require N-1 dimensions, not the 2-dimensional figure drawn. We thus add links using a Path Erasing Algorithm, with an adjustable threshold; heavier lines indicate stronger relationship.

- 3 of 49 NBS researchers (6%) with 3 or more connections.

One research field is highly coherent; one is quite diffuse.

3.2 Key Players

Countries: Those publishing the most on DSSC in WOS-indexed sources since 2009 appear in Table 1. Note China's prominence in publishing in these international journals and in producing highly cited, recent articles. The US and Switzerland show relatively high percentages of their publications receiving high citation rates.

Table 1. Leading Countries publishing on DSSC in WOS since 2009

Countries	#	# of 135 hi cited	% hi cited
Peoples R China	440	29	7%
South Korea	267	10	4%
Japan	192	18	9%
Taiwan	181	10	6%
USA	167	26	16%
Switzerland	101	21	21%
India	81	2	2%
Germany	65	11	17%
England	59	12	20%
Spain	56	10	18%
Sweden	56	11	20%
TOTAL globally	1691	135	8%

Scientific Activity for Leading Organizations: Focusing on WOS, we can identify the leading DSSC research publishing organizations. Table 2 shows the "top ten." Note that none are companies. Such "research profiles" are readily generated in the software,[3] with the analyst selecting which variables to include.

The Chinese Academy of Sciences ("CAS") leads the list. Of course, CAS includes some 100 discrete units, so for some analyses we separate those. Second is a university – the Swiss Federal Institute of Technology in Lausanne (also appearing as Ecole Poly technique Federale Lausanne – we'll use "EPFL" as shorthand). This is where DSSC work originated, led by Grätzel and colleagues. Notably present are Asian organizations, with 6 of the top 10. Notably absent are any U.S. organizations. The U.S. National Renewable Energy Lab ("NREL") ranks 13th with 58 DSSC publications, but only 7 from 2009 on.

Browsing through this "research profile," we can ascertain a number of important features about these organizations. This information could be useful in identifying prospective partners or individual targets with whom to pursue collaboration. For instance, Column 2 shows the countries, other than their own; this indicates the extent and direction of their international collaboration. For instance, note that relatively few of CAS' 286 publications involve co-authors from other countries (led by 18 with Switzerland, all with EPFL). In contrast, the Swedish Royal Institute of Technology collaborates very heavily – 61% of its WOS papers involve at least one foreign co-author; 34%, involve Uppsala University (in Sweden; and these overlap with international papers).[4]

Column 3 shows a few of the prominent keywords and phrases. These give a glimpse into the technical emphases of each organization. "Efficiency" is prominent for all but Korea University. "Conversion" appears as a top keyword for all except Imperial College. Conversely, some topics appear more specialized – e.g., "electrolyte" appears in these top terms only for CAS and Osaka University. So, if one were seeking expertise in that area, these might be good candidate institutions to approach. [One could peruse full lists and identify the researchers pursuing particular topics within TDA.]

Column 4 shows the top authors from these institutions for the 4104 papers. This can help get a sense of the size of the research team at each organization (recall that CAS includes many separate organizational units). Note the huge numbers of papers by the leading EPFL researchers and the obvious extent of overlap – suggesting a research team. We sometimes spotlight one organization such as this and map its internal and external research networks (not shown).

The last column shows the percentage of these institutions' WOS publications since 2009. The Asian organizations are especially active recently.

Fig. 5 plots the trend in WOS publication for these top ten DSSC organizations. Previously, EPFL led; now CAS leads in recent years.

3.3 Leading DSSC Companies

We now look across our search results from all four databases together. How might companies likely pursue commercialization of DSSC? We gain three vantage points on companies engaging DSSC: those publishing DSSC research; those patenting, and those active in the business arena. In an imperfect process, we have identified leading companies in each dataset to see the commercialization activities, then checked in the others as well. Table 3 pre-sents the results. This gives a snapshot

[3] These analyses draw heavily on text mining software. Two related versions are VantagePoint and Thomson Data Analyzer ("TDA") – see http://science.thomsonreuters.com/support/faq/tda/ and www.theVantagePoint.com. TDA is most popular and accessible in China as it is supported strongly by Thomson Reuters staff in Beijing. Hence, we will mention the software here as TDA.

[4] Information not presented in the Table, but easily discerned in TDA.

978-1-935068-76-1 © 2011 SciRes.

Table 2. Research Profiling: Top 10 DSSC Research Organizations

Author Affiliations (Name Only)	Countries	Combined Keywords + Phrases	Authors	Publication Year
%Selection%	Research partners	Top 5	Top 5	% since 2009
Chinese Acad Sci (CAS) [286]	Switzerland [18] Japan [9] India [4] USA [4]	EFFICIENCY [90] CONVERSION [83] FILMS [71] PERFORMANCE [58] ELECTROLYTE [56]	Dai, S Y [64] Lin, Y [48] Hu, L H [43] Wang, K J [39] Xiao, X R [35]	43% of 286
Swiss Fed Inst Technol (EPFL) [259]	UK [29] China [25] South Korea [20] Italy [19]	CONVERSION [71] EFFICIENCY [62] FILMS [60] RECOMBINATION [60] PERFORMANCE [52]	Gratzel, M [226] Zakeeruddin, S M [106] Nazeeruddin, M K [88] Humphry-Baker, R [41] Wang, P [36]	35% of 259
Natl Inst Adv Ind Sci & Technol [137]	Taiwan [29] China [5] India [4]	TiO2 [41] FILMS [35] CONVERSION [30] EFFICIENCY [30] LIGHT [26] PERFORMANCE [26]	Hara, K [38] Arakawa, H [29] Sugihara, H [26] Sayama, K [22] Katoh, R [17] Yanagida, M [17]	45% of 137
Korea Inst Sci & Technol [94]	India [9] Switzerland [2]	EFFICIENCY [26] conversion efficiency [20] effect [19] FILMS [19] TiO2 [19] TRANSPORT [19]	Park, N G [38] Kim, K [25] Do Kim, Y [13] Ko, M J [13] Kim, J Y [12]	50% of 94
Uppsala Univ[94]	China [8] Switzerland [7] France [5] Germany [3] Peru [3]	dye [36] TiO2 [30] FILMS [29] EFFICIENCY [25] CONVERSION [23]	Hagfeldt, A [72] Boschloo, G [38] Sun, L C [15] Lindquist, S E [13] Rensmo, H [13]	33% of 94
Osaka Univ[93]	Sri Lanka [5] China [3] Spain [3] Brazil [2] South Korea [2]	EFFICIENCY [39] FILMS [34] CONVERSION [25] photocurrent [21] ELECTROLYTE [19]	Yanagida, S [77] Wada, Y [53] Kitamura, T [43] Masaki, N [19] Kubo, W [16]	9% of 93
Korea Univ[91]	Switzerland [16] India [9] China [7] Spain [2] USA [2]	CONVERSION [29] LIGHT [25] photocurrent [24] cell [23] fill factor [23]	Kim, J K [43] Ko, J [33] Choi, H [21] Vittal, R [21] Kang, M S [18] Kang, S O [18]	44% of 91
Natl Taiwan Univ[84]	India [6] Japan [2]	EFFICIENCY [30] CONVERSION [25] PERFORMANCE [22] TiO2 [20] ELECTRODES [13] increase [13] LIGHT [13]	Ho, K C [57] Lee, K M [26] Chen, J G [22] Hsu, C Y [14] Chen, C Y [13] Wu, C G [13]	65% of 84
Univ London Imperial CollSciTechnol& Med[83]	Spain [18] Switzerland [15] Netherlands [8] Brazil [5]	RECOMBINATION [33] EFFICIENCY [25] TiO2 [24] FILMS [23] NANOCRYSTALLINE TIO2 FILMS [23] TIO2 FILMS [23]	Durrant, J R [63] Haque, S A [26] Palomares, E [24] O'Regan, B C [19] Gratzel, M [12]	27% of 83
Royal InstTechnol[79]	China [26] Peru [4] France [3] Switzerland [3]	TiO2 [24] FILMS [22] EFFICIENCY [21] RECOMBINATION [20] CONVERSION [19] dye [19] LIGHT [19]	Hagfeldt, A [57] Boschloo, G [37] Sun, L C [33] Yang, X C [16] Hagberg, D P [15] Kloo, L [15] Marinado, T [15]	51% of 79

a. Based on Web of Science publication.

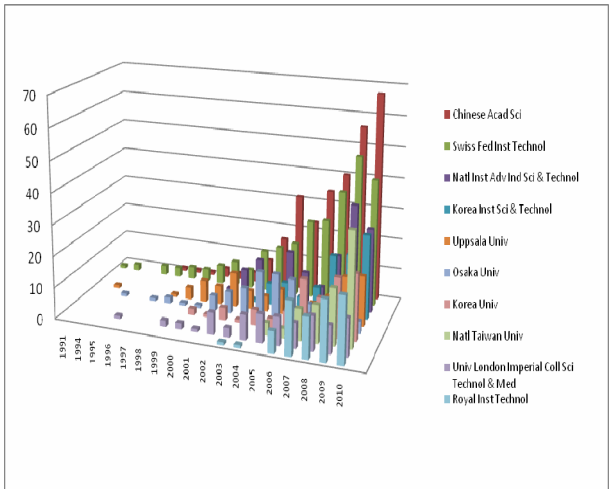


Figure 5. DSSC Publications in Web of Science by the Top 10 Organizations over Time

of activity in scien-tific research WOS), engineering-oriented research (Compendex), patenting (DWPI),[4] and business activity (Factiva).

Table 3 shows Japanese companies leading DSSC activity. Note the marked variation in organizational presence in these four different data sources. For instance, Samsung is the leading patent assignee and research publisher (in this compilation just for companies) on DSSC, but has not been frequently mentioned in conjunction with business actions (Factiva). The use of multiple information sources in conjunction with each other enriches perspective on how a NEST is being developed and the roles of various important actors.

Table 3. Leading DSSC Companies across Databases

	SCI	Compendex	DWPI	Factiva
Samsung SDI Co LTD	52*	38	65*	4
Sharp Co Ltd	27*	24	17*	4
Nippon Oil Corp	15*	35	27*	10*
Hayashibara Biochem Labs Inc	14*	9	0	0
Fujikura Ltd	12*	8	17*	9*
Chemicrea Co Ltd	10*	8	0	0
Sumitomo Osaka Cement Co Ltd	10*	3	3	2
Toshiba Co Ltd	9*	7	2	1
Konarka Technologies Inc	7*	11	11*	9*
DONG JIN SEMICHEM CO LTD	0	1	16*	8*
SONY CORP	10	10	17*	17*
Evonik Degussa GmbH	0	0	0	15*
STMicroelectronics NV	0	0	0	12*
Data Systems & Software Inc	0	0	0	8*
Dongjin Semichem Co Ltd	0	1	0	8*
Dyesol Ltd	3	3	2	8*

[4] This tally for DWPI is restricted to patent families that include at least one non-native country. This reduces the total set from some 3100 to about 800 records. This process aims to focus on those patents with highest perceived value (sufficient for the assignees to seek intellectual property protection in other countries). It also helps compensate for the higher counts of Japanese patenting as the Japanese Patent Office follows different practices (tending toward more discrete, rather than compound, patents and not examining in depth unless a patent is challenged).

Some leading research institutions are not notable in commercialization activity, so we just show and compare the data for companies in Table 3. However, we also explored the leading three research organizations from WOS (CAS, EPFL, and AIST) in those other three databases. Results were highly uneven. In Compendex, we do find significant publication activity: CAS – 183 papers; EPFL – 58; and AIST – 50. We do not find much activity by any of them in DWPI and Factiva, but this may be related to problematic coverage and searching in Factiva. In DWPI, we do find about 60 CAS patent families in the full set, but none in the set limited to families with foreign filings.

3.4 Technology Components & Key Players

A dye-sensitized solar cell is generally composed of four component *dimensions: the photoanode, sensitizer, electrolyte, and counter-electrode*. To gain a sense of which are the leading technologies (or the one with most potential) for each DSSC dimension, we conduct an in-depth analysis for the four technology dimensions based on extracted keywords.

In order to show how we can get more information about technical development, we do trend analysis of the dye technology dimension in the four databases. What we can see in Fig. 6 is that the numbers of organic dye records climb up quicker than metal complex dye in recent years. In WOS, Compendex, and DWPI, the numbers of organic dye records rocket upward since 2007. While in Factiva, the numbers rocket since 2009 – i.e., two years later. And the numbers of records on organic dye exceeded metal complex dye in 2009 and 2010 in the 4 databases. This may suggest that organic dye is a more promising dye in future development and commercialization of DSSC.

To help interpret "who is pursuing what," we devise a "key players by technologies" analysis. It is based on our keywords extraction and classification. Key terms relating to the four technical dimensions are first extracted from the record sets, and then cleaned. We present these to two knowledgeable technical professionals (see Acknowledgements) for review.

978-1-935068-76-1 © 2011 SciRes.

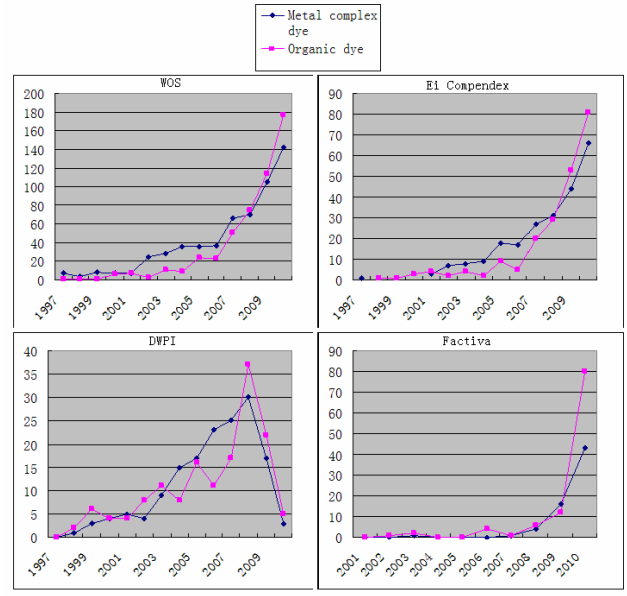


Figure 6. Trend analysis of dye research in 4 databases

Based on the resulting term sets, we compare key players' emphases and intents by counting patents or publications which contain those different keywords. To gain perspective, we visualized these emphases by creating a players-technologies network map. Two types of nodes separately represent players (companies) and technologies. Lines demonstrate linkages between players and technologies – the thicker the line, the stronger the linkage between player and technology. This provides an approach to analyze and visualize organizations' R&D emphases and intents.

Fig. 7 shows the top players who have more than 10 publications on dye technologies in WOS. We classified dyes into two classes, with experts' assistance -- metal complex dye and organic dye. We can clearly see that China, Japan and Switzerland focus more on dye research. There are 6 Chinese organizations, 4 Japanese organizations, and 2 Swiss organizations. More players focus on organic dye than on metal complex dye. Meanwhile, we find that the top 13 players who have the most DSSC publications are apt to focus on metal complex dye research. They are mostly earlier players in the field so that they have done a lot of research on metal complex dye. On the contrary, the relatively recent players focus more on organic dyes. This implies that organic dye is becoming more and more popular in recent years. In Fig. 7, we also can see that players in China show more interest in organic dyes. Dalian University of Technology, Nankai University, University of Science and Technology of China, and Xiangtan University are all universities in China. And the Chinese Academic of Sciences is the leader in China which publishes the most papers on both of the two classes of dyes in China.

Also, we analyzed the other three databases to see if there is any more information. In DWPI, we find that there are more Japanese companies than in WOS (of the top 13 players, 9 are in Japan). The Chinese Academic of Sciences applied for more patents on organic dye than on metal complex dye. It may illustrate that its research on organic dye is more innovative than metal complex dye. And in Factiva, players show more differences. The business records on organic dye are almost twice that on metal complex dye (130 vs. 69). We also find that the TDK Corporation is one of the top players in DWPI; it has only one patent on organic dye. The other top players in WOS and DWPI don't have business information on dye research.

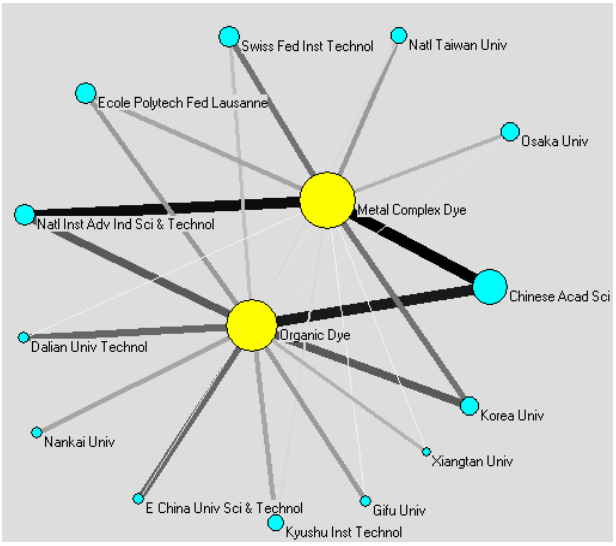


Figure 7. Players-technologies Network Map

To better understand leading institutions' research emphases on DSSCs, we create a spider map (Fig. 8) to look into the leading DSSC institutions' R&D activities on those important technologies in four technology dimensions. Fig. 8 shows the percentage publication shares of the top two institutions in WOS on each technology. The percentage is calculated as the ratio of the publications in specific DSSC technology affiliated with one institution to the total number of publications on DSSCs affiliated with that institution. The Chinese Academy of Sciences (CAS) and Ecole Polytechnique Federal Lausanne (EPFL) are the top two institutions, with more than 100 publications in WOS. In Fig. 8, we see that both CAS and EPFL show more interest in TiO_2, dye, and electrolyte research. This is not too surprising because semiconductor, dye, and electrolyte are the most important components of DSSCs that mainly impact performance. When we pay attention to the semiconductor dimension, we can see that EPFL emphasizes TiO_2 more than CAS, while CAS does more research on ZnO than EPFL. For the other dimensions, we also can see differ-

ences. CAS shows relatively more interest in organic dye, gel electrolyte, and Platinum counter electrodes (considering each of the four technology dimensions), while EPFL focuses more on metal complex dyes and solid electrolytes. Such analyses may help us identify technical strengths and weakness of institutions, and further understand their research emphases in each technology dimension. Such intelligence can help point an organization to desirable partners in new research or technology development initiatives.

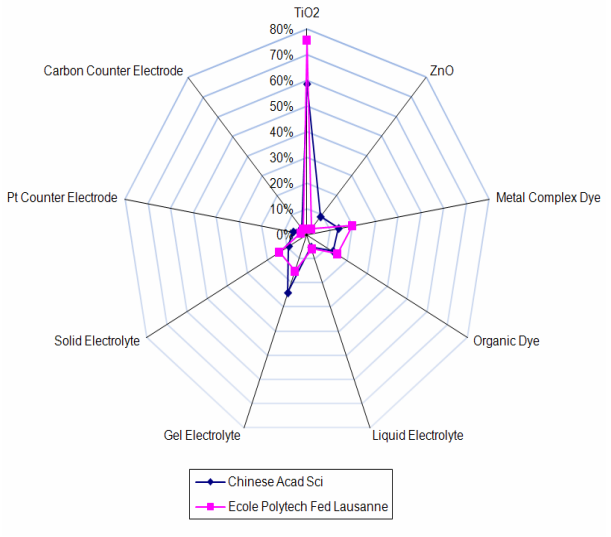


Figure 8. Publication Ratios of Leading Institutions on Technologies in Four Tech Dimensions

4 Discussion

We illustrate a range of analyses of ST&I information resources. Text mining software enables one to answer "what" questions about a topic of interest. We identify which fields are heavily engaged in the R&D efforts and map these through a science overlay map. This can help identify capabilities vital to compete in the field. We show analyses and network visualizations that help learn how tightly interconnected the research field is.

We also work to answer "who" questions. We illustrate how to identify the leading countries doing R&D on an emerging technology under scrutiny. We then focus down on leading institutions, breaking out areas of emphasis and trends over time. We also compare companies' activity across the four databases to gain insights into relative strengths in research, development, invention, and business initiation concerning DSSCs.

Tech mining can continue to probe more deeply. We illustrate by addressing "who is doing what" questions. We perform key term analyses to determine the leading technologies under development for each of four key DSSC dimensions. Bi-partite research network plots show which institutions emphasize which technologies. Spider plots provide different comparisons.

Such knowledge is vital for prudent R&D management to know the surrounding "mountain" of topics and researchers. Richer awareness of that research landscape can identify complementary research methods, theories, findings, and potential applications of one's own work as well.

Acknowledgments

We especially thank Dr. Jud Ready and PhD candidate, Chen Xu, of Georgia Tech for their helpful guidance.

References

[1] Porter, A.L., Kongthon, A., Lu, J-C. Research Profiling: Improving the Literature Review, *Scientometrics*, 2002, 53, 351-370.

[2] Guston, D H, Sarewitz, D. Real-time technology assessment. Technology *in Society* , 2002, 24, 93-109.

[3] Porter, A. L., Cunningham, S. W.*Tech mining: exploiting new* technologies *for competitive advantage*.Wiley, New York, 2005.

[4] Robinson, D.K.R., Huang, L., Guo, Y., and Porter, A.L. (under submission) Forecasting Innovation Pathways for New and Emerging Science & Technologies, *Technological Forecasting& Social Change*.

[5] Chen, C., and Hicks, D.Tracing knowledge diffusion, Scientometrics, 2004, 59 (2), 199-211.

[6] Narin, F., Hamilton K. S., Olivastro D. The increasing linkage between U.S. technology and public science, *Research Policy*, 1997, 26, 317-330.

[7] Carpenter, M., Cooper, M., Narin, F. Linkage between basic research literature and patents,*Research Management*, 1980, 13 (2), 30-35.

[8] Porter, A.L., Youtie, J., Shapira, P., Schoeneck, D. Refining search terms for nanotechnology. *Journal of Nanoparticle Research*, 2007, 10(5), 715-730.

[9] Conibeer, G. Third-generation Photovoltaics. *Materialstoday*, 2007, 10, 42-51.

[10] Green, M.A. Thi*rd Generation Photovoltaics: Ultra-High Efficiency at Low Cost*. Springer-Verlag, Berlin, 2003.

[11] Grätzel, M. Dye-sensitized solar cells. Journal of Photochemistry and Photobiology C: *Photochemistry Reviews*, 2003, 4, 145-154.

[12] Aydil, E.S. Nanomaterials for Solar Cells.*Nanotechnology Law & Business*, 2007, 4(3), 275-292.

[13] Lenzmann, F.O., Kroon, J.M. Recent Advances in Dye-Sensitized Solar Cells,*Advances in OptoElectronics, 2007*, 10.

[14] Snaith, H.J., Schmidt-Mende, L. Advances in Liquid-Electrolyte and Solid-State Dye-Sensitized Solar Cells. *Advanced Material*, 2007, 19, 3187-3201.

[15] O'Regan, B.,and Grätzel, M. A low-cost, high-efficiency solar-cell based on dye-sensitized colloidal TiO2 films, *Nature*, 1991, 353 (6346), 737-740.

Appendix I: WOS Search Terms

No.	Records	Term	Annotation
#1	4000	TS= (((dye-sensiti*) or (dye* same sensiti*) or (pigment-sensiti*) or (pigment same sensiti*) or (dye* same sense)) same (((solar or Photovoltaic or photoelectr* or (photo-electr*)) same (cell or cells or batter* or pool*)) or	It is various expressions of Dye sensitized solar cell (pigment sensitized solar cell is a kind of DSSC).

978-1-935068-76-1

		photocell* or (solar-cell*)))	
#2	1204	TS=((DSSC or DSSCs) not ((diffuse cutaneous Systemic sclerosis) or (diffuse cutaneous SSc) or (diffuse SSc) or (distributed switch and stay combining) or (Distributed Static Series Compensator*) or (decoupled solid state controller*) or (Active Diffuse Scleroderma*) or (systemic sclerosis) or (diffuse scleroderma) or (Deep Space Station Controller) or (Data Storage Systems Center) or(decompressive stress strain curve) or (double-sideband-suppressed carrier) or (Flexible AC Transmission Systems) or (DSS induced chronic colitis) or (Dynamic Slow-start) or (dextran sulfate sodium) or (disease or patient* or QSRR)))	It is the papers which include 1) DSSC and relate to Dye sensitized solar cell and exculde 2) noisy data.
#3	330	TS=((((dye- Photosensiti*) or (dye same Photosensiti*) or (pigment- Photosensiti*) or (pigment same Photosensiti*)) same ((solar or Photovoltaic or photoelectr* or (photo-electr*)) same (cell or cells or batter* or pool*))) not (melanocyte* or cancer))	It is 1) various expression of Dye photo-sensitized solar cell, and use 2) (melanocyte* or cancer) to exclude noisy data.
#4	188	TS = (((((dye adj (sensiti* or photosensiti*)) and (conduct* or semiconduct*)) same electrode*) and electrolyte*) not (wastewater or waste-water or degradation))	It search term searches DSSC papers according to1) the component of DSSC and use 2) (wastewater or waste-water or degradation) to exclude noisy data.
Total	4104	#1 or #2 or #3 or #4	Combined search terms.

Appendix Ⅱ: DWPI Search Terms

No.	Records	Term	Annotation
#1	2783	TS= (((dye-sensiti*) or (dye* same sensiti*) or (pigment-sensiti*) or (pigment same sensiti*) or (dye* adj sense)) same ((solar or Photovoltaic or photoelectr* or (photo-electr*)) same (cell or cells or batter* or pool*)))	It is various expressions of Dye sensitized solar cell (pigment sensitized solar cell is a kind of DSSC).
#2	1	PN= CN101271775	It is the patent which includes in DSSC but not includes #1 and relates to Dye sensitized solar cell.
#3	231	TS= (((dye- Photosensiti*) or (dye same Photosensiti*) or (pigment- Photosensiti*) or (pigment same Photosensiti*)) same ((solar or Photovoltaic or photoelectr* or (photo-electr*)) same (cell or cells or batter* or pool*))) and IP=(H01G* or H01M* or H01L* or G03C*)	It is 1) various expressions of Dye photo-sensitized solar cell, and use 2) IPC to exclude noisy patents.
#4	275	TS= (((dye- optoelectri*) or (dye same optoelectri*) or (pigment- optoelectri*) or (pigment same optoelectri*) or (dye- opto-electri*) or (dye same opto-electri*) or (pigment- opto-electri*) or (pigment same opto-electri*)) same ((solar or Photovoltaic or photoelectr* or (photo-electr*)) same (cell or cells or batter* or pool*))) and IP=(H01G* or H01M* or H01L* or G03C*)	It searches the DSSC patents which include optoelectri* but not include sensiti* and photosensiti*, and use 2) IPC to exclude noisy patents.
#5	651	TS = (((dye and (conduct* or semiconduct*)) same electrode*) and electrolyte*)	Itsearches DSSC patents according to the component of DSSC.
Total	3097	#1 or #2 or #3 or #4 or #5	Combined search terms.

Appendix Ⅲ: Ei Compendex Search Terms

No.	Records	Term	Annotation
#1	3547	((dye-sensiti* OR (dye* NEAR sensiti*) OR pigment-sensiti* OR (pigment NEAR sensiti*) OR (dye* NEAR sense)) AND (((solar OR Photovoltaic OR photoelectr* OR photo-electr*) AND (cell OR cells OR batter* OR pool*)) OR photocell* OR solar-cell*)) WN AB OR ((dye-sensiti* OR (dye* NEAR sensiti*) OR pigment-sensiti* OR (pigment NEAR sensiti*) OR (dye* NEAR sense)) AND (((solar OR Photovoltaic OR photoelectr* OR photo-electr*) AND (cell OR cells OR batter* OR pool*)) OR photocell* OR solar-cell*)) WN TI	It is various expressions of Dye sensitized solar cell (pigment sensitized solar cell is a kind of DSSC) in abstract and title.
#2	1082	((DSSC or DSSCs) not ((diffuse cutaneous Systemic sclerosis) or (diffuse cutaneous SSc) or (diffuse SSc) or (distributed switch and stay combining) or (Distributed Static Series Compensator*) or (decoupled solid state controller*) or (Active Diffuse Scleroderma*) or (systemic sclerosis) or (diffuse scleroderma) or (Deep Space Station Controller) or (Data Storage Systems Center) or(decompressive stress strain curve) or (double-sideband-suppressed carrier) or (Flexible AC Transmission Systems) or (DSS induced chronic colitis) or (Dynamic Slow-start) or (dextran sulfate sodium) or (disease or patient* or QSRR))) WN AB or ((DSSC or DSSCs) not ((diffuse cutaneous Systemic sclerosis) or (diffuse cutaneous SSc) or (diffuse SSc) or (distributed switch and stay combining) or (Distributed Static Series Compensator*) or (decoupled solid state con-	It is the papers which include 1) DSSC and relate to Dye sensitized solar cell and exculde 2) noisy data in abstract and title.

		troller*) or (Active Diffuse Scleroderma*) or (systemic sclerosis) or (diffuse scleroderma) or (Deep Space Station Controller) or (Data Storage Systems Center) or(decompressive stress strain curve) or (double-sideband-suppressed carrier) or (Flexible AC Transmission Systems) or (DSS induced chronic colitis) or (Dynamic Slow-start) or (dextran sulfate sodium) or (disease or patient* or QSRR))) WN TI	
#3	109	(((((dye- Photosensiti*) or (dye near Photosensiti*) or (pigment- Photosensiti*) or (pigment near Photosensiti*)) and ((solar or Photovoltaic or photoelectr* or (photo-electr*)) and (cell or cells or batter* or pool*))) not (melanocyte* or cancer)) WN AB or ((((dye- Photosensiti*) or (dye near Photosensiti*) or (pigment- Photosensiti*) or (pigment near Photosensiti*)) and ((solar or Photovoltaic or photoelectr* or (photo-electr*)) and (cell or cells or batter* or pool*))) not (melanocyte* or cancer)) WN TI	It is 1) various expression of Dye photo-sensitized solar cell, and use 2) (melanocyte* or cancer) to exclude noisy data in abstract and title.
#4	181	((((dye near sensiti*) or (dye near photosensiti*)) and ((conduct* near electrode*) or (semiconduct* near electrode*)) and electrolyte*) not (wastewater or waste-water or degradation)) WN AB or ((((dye near sensiti*) or (dye near photosensiti*)) and ((conduct* near electrode*) or (semiconduct* near electrode*)) and electrolyte*) not (wastewater or waste-water or degradation)) WN TI	It search term searches DSSC papers according to1) the component of DSSC and use 2) (wastewater or waste-water or degradation) to exclude noisy data in abstract and title.
Total	3097	#1 or #2 or #3 or #4	Combined search terms.

Appendix IV: Factiva Search Terms

No.	Records	Term	Annotation
#1	2408	(((dye-sensiti*) or (dye* and sensiti*) or (pigment-sensiti*) or (pigment and sensiti*) or (dye* adj sense)) same (((solar or Photovoltaic or photo-electr* or (photo-electr*)) and (cell or cells or batter* or pool*)) or photo-cell* or (solar-cell*)))	It is various expressions of Dye sensitized solar cell (pigment sensitized solar cell is a kind of DSSC).
#2	1586	((DSSC or DSSCs) not ((diffuse cutaneous Systemic sclerosis) or (diffuse cutaneous SSc) or (diffuse SSc) or (distributed switch and stay combining) or (Distributed Static Series Compensator*) or (decoupled solid state controller*) or (Active Diffuse Scleroderma*) or (systemic sclerosis) or (diffuse scleroderma) or (Deep Space Station Controller) or (Data Storage Systems Center) or(decompressive stress strain curve) or (double-sideband-suppressed carrier) or (Flexible AC Transmission Systems) or (DSS induced chronic colitis) or (Dynamic Slow-start) or (dextran sulfate sodium) or (disease or patient* or QSRR)))	It is the news which include 1) DSSC but not includes #1 and relate to Dye sensitized solar cell and exculde 2) noisy data.
#3	65	(((((dye- Photosensiti*) or (dye and Photosensiti*) or (pigment- Photosensiti*) or (pigment and Photosensiti*)) same (cell or cells or batter* or pool*)) and (solar or Photovoltaic or photoelectr* or (photo-electr*))) not (melanocyte* or cancer))	It is 1) various expression of Dye photo-sensitized solar cell, and use 2) (melanocyte* or cancer) to exclude noisy data.
#4	50	(((((dye adj (sensiti* or photosensiti*)) and (conduct* or semiconduct*)) same electrode*) and electrolyte*) not (wastewater or waste-water or degradation))	It search term searches DSSC news according to1) the component of DSSC and use 2) (wastewater or waste-water or degradation) to exclude noisy data.
Total	4052	#1 or #2 or #3 or #4	Combined search terms.

978-1-935068-76-1

Session 3

Engaging with the Information Environment

分论坛 3

信息环境构建

Influential Factors on Digital Consciousness of China Communities

——An Empirical Study Based in Beijing, Shanghai and Guangdong

Hui YAN

Dept. of Information Resource Management, Business School, Nankai University, Tianjin, China

Email: yanhui@nankai.edu.cn, hyanpku@gmail.com

Abstract: By employing China Family Panel Studies (CFPS) 2008 database and Logistic Regression method, this paper measures the impacts of economical capital, cultural capital and social factors on digital consciousness of China's digital poor communities, and concludes with implications for central government policies (including information policy) for a digitally-equal society.

Keywords: Digital consciousness; digital inequality; digital equity; ICT; information policy; digital poor communities

中国社群数字化意识的影响因素研究

——基于京沪粤三地调研的实证分析

闫慧

南开大学商学院信息资源管理系，天津，中国，300071

摘　要：论文采用逻辑斯蒂回归分析方法和 CFPS2008 数据库，分析经济资本、文化资本和社会因素对社会主体数字化意识的影响程度，为数字不平等和数字公平视角下的意识贫困社群处于数字贫困状态提供实证与政策方面的解读。

关键词：数字化意识；数字不平等；数字公平；信息通讯技术；信息政策；数字贫困社群

1 引言

信息通讯技术（Information and Communication Technologies, ICT）的重大革命为社会结构的变迁拓展了新的维度，社会不平等的表现和演变更加复杂。19 世纪后半叶和 20 世纪，电报、广播和电视等传统 ICT 加速了原有生产关系和社会阶层在地域间的调整和新兴经济模式和社会形态的建立，社会不平等基础上的信息分化现象呈现出加剧的态势；计算机和互联网络等数字化 ICT 在延续新经济生命力的同时，也在刷新人类不平等的纪录。本项研究的宗旨在于探索数字不平等（Digital inequality）表现维度之数字化意识出现差异的深层原因，为建立旨在实现数字公平（Digital equity）的公益信息制度提供实证经验。

2 研究综述

国际数字不平等的研究已经有将近 15 年的历史。据笔者调查，最早提出“数字不平等”概念的是弗吉尼亚理工大学政治学教授蒂莫西·鲁克（Timothy W. Luke），他于 1997 年秋在《新政治学》杂志上发表一篇题为“数字不平等的政治学：虚拟空间的获取/接入，能力和分配”[1]。他认为，数字化不平等的标志是，历史上的阶级斗争在新时代转变成企业所有者和工人之间、生产者与消费者之间、知情者与不知情者之间、拥有技术接入机会的人和没有这些机会的人之间、网络素养具备者和不具备者之间的“信息战争”。国内外相关学者认为，数字不平等的表现维度主要有心理不平等（Purpose / Motivation / Attitude）、ICT 设施的接入（Access）与所有权（Ownership）的不平等、技术使用状况（Usage/use）、技能不平等（Skills）等。心理状态的不平等具体表现在目的[2]、动机和兴趣[3]、态度[4]、数字化意识[5]等。

相关学者对影响数字不平等的因素研究较多，其选择取决于这些学者的研究问题和研究对象。如果研究个人层面的数字不平等现象，他们则倾向于选取人口统计学的指标以及个人所拥有的各类政治、经济、文化、技术、社会等资本；如果研究地区层面的数字

978-1-935068-76-1

不平等，则会选择与地区发展相关的要素如地区经济发展、城乡差异等作为分析对象；如果研究国际层面或国家之前的数字不平等，则会聚焦在各个国家政府政策和经济发展水平、全球信息经济、政治经济要素等问题上。数字不平等的影响因素可以分为两大类——人口统计指标和资源要素；前者包括性别、年龄、种族、健康程度、就业、婚姻、生育、家庭地位等八个因素；后者包括地理、经济、技术、社会、文化、人力、政治等七大类资源或资本。经济资本的一般要素包括收入[6-15]、阶层[16-18]、经济发展[19-20]等。文化资本对数字不平等的影响途径包括文化角色[10-11][19-20]、语言[17]、文化空间[21]、大众媒介的支持[9]、信息内容[22]、教育[2][4][7-8][11-13][15-16][18][23-25]等。社会位置以及机构[2][6-7][10-11][19][21-22]、个人网络[26]、社会网络[6][9-10]等因素构成社会资本因素。这些变量对于社会主体数字化意识的影响并没有得到充分的验证，特别是在中国情境下。

3 术语界定

社群是指社会中拥有共同利益、共同的经历或历史、共同的道德价值观、认同和共同的期望的个体，通过血缘、地缘、社会关系和社会网络或特定社会组织所形成的集合体。常见的社群包括亲友、地理社区、政治社群、社交圈子等。本研究关注的社群主要是在中国信息化社会中处于不同社会层次上的社群。数字不平等是指不同的国家和地区、组织、社群和个人在数字化 ICT 接入和使用以及信息资源的开发和利用实践活动中形成的多样化的信息差距。据此，信息社会可以划分为数字贫困社群（含意识贫困社群、物质贫困社群等）、数字中产社群、数字富裕社群和数字精英等五个阶层。

数字化意识是指社会主体对电脑、互联网以及其他用于获取网络信息的设备和技术重要性的认识，这也是数字贫困社群脱离贫困状态的基本前提条件。经济资本是用来描述社群及其成员通过各种就业或创业渠道和方式，赚取并积累的物质财富，尤其是以货币化收入来衡量的财富；测量指标不仅包括货币化收入的绝对值，也包括社群成员对自身收入、家庭收入等在整个社会中的水平和地位。文化资本主要是对社群及其成员的受教育程度、语言能力等的概括；社会因素（包括社会资本）的操作化定义是社群及其成员以其自身性格和实际需求为出发点，建立并保持的社会关系，测量标准的例子包括：是否与别人容易相处，人缘关系如何等等。

4 研究设计

4.1 调查样本

本研究所采用的数据是北京大学中国社会科学调查中心 2008 年收集并于 2009 年在有限范围内开放的“中国家庭动态跟踪调查”（China Family Panel Studies, CFPS）数据库。2008 年问卷调查的样本选取方法是：在北京、上海、广东各抽取了 8 个区/县的 32 个村/居的 800 户家庭，每户家庭所有成员都是受访对象；共完成 24 个区/县、95 个村、居、2375 户、7214 个人的访问，共计获得 9708 份问卷，620 多万个数据点，有效问卷为 7212 份，其中含成人问卷 6093 份，少儿问卷为 1119 份。该数据集合的特点是主题的综合性强、样本量大、以及对发达地区的代表性比较好等，其中也涉及到 ICT 的问题，尤其是电脑、手机和互联网的使用问题。

4.2 变量选取与分析方法

CFPS2008 年数据库中能够反映数字化意识的问题是“您认为互联网络在您的学习/工作中（的重要程度是，笔者注）”（Q5.9.2）。该变量作为数字化意识回归分析的因变量。由于因变量是定序变量，不能做一般线性回归分析，但能够支持罗吉斯蒂回归分析（Logistic Regression），Logistic 函数属于一种概率函数[27]，反映的是一些自变量对因变量所代表的“行为”发生概率的影响。这也是目前涉及到二分变量（定类变量分为两种情况）的研究中最常用方法之一。

根据笔者对相关文献的综述结果，选取以下几个变量作为自变量进行罗吉斯蒂回归分析：（1）受访者的收入，对应于数据库中的问题“去年您的个人收入总计”，由于数值较大，采用社会科学研究中的常规处理办法，即以 1000 元为单位，并取其自然对数（为了线性变换而转换），即转化成 ln(收入/1000)，其含义是如果 ln(收入/1000)从原来的 X 增加 1 个单位，那么收入（以千元为单位）增加到原来水平的 e 倍（即大约 2.718281828 倍）；（2）受访者对其自身社会地位的感知；（3）受访者的受教育水平，根据数据库中受访者接受不同阶段教育的时间统计数据及各阶段平均时间而计算出来的，1-6 年为小学水平或学历，并作为参照变量，7-11 年为初中，12-14 年为高中或中专，15-18 年为大专或大学本科，19 年以上为研究生水平；（4）受访者的年龄；（5）受访者在工作日上网的时

间与相同工作日内休闲时间的比重，这个指标是为了从受访者的实际行为中识别他们对互联网络的相对重视程度（与工作日的休闲时间相比较，可以发现该指标），比值越大，说明互联网络对受访者的学习/工作/生活来说更加重要；反之亦然。

数字化意识的程度与上述各个自变量的相关关系如表 1 所示，按相关程度（即相关系数的绝对值）从大到小排列的自变量依次是受教育水平、工作日上网时间与休闲时间之比、收入、年龄、社会地位等，而且这些相关系数在统计上都是显著的，能够为数字化意识的变化提供了显著的解释水平。年龄与数字化意识的程度是成弱负相关关系，即年龄越大，其对互联网络重要性的评价会越低；其他四个变量均与数字化意识成正相关关系。这为逻辑斯蒂回归分析提供了前提条件。

5 研究发现

按照罗吉斯蒂回归分析的步骤，首先将定序性质的因变量（数字化意识）转换为二分虚拟变量，即将选项 1（不重要）、2（不太重要）和 3（一般）合并为“0”（不重要），将选项 4（重要）转变为“1”（重要）；依次重新归类，认为互联网络对其学习或工作重要的受访者人数 847 人，占到“非缺省”受访者总数（1802 人）的 47%；认为“不重要”的人数为 955 人，比例为 53%。随后将 5 个自变量纳入到分析过程中。结果显示，该罗吉斯蒂回归模型的截距模型中似然函数值（likelihood）相应值（-2log likelihood）较大，说明模型的拟合程度一般。但该概率模型的 wald 值（反映偏回归系数显著程度）为 30.233，说明几个自变量的综合作用比较显著；方程的总体显著度（表示错误地拒绝了一个真实的无关假设的概率）小于 0.01，说明这些自变量的综合作用非常明显。

Table 1. Correlations between digital consciousness and different influential factors
表 1. 数字化意识与各个自变量的相关关系

		互联网络重要性的认知	年龄	社会地位的感知	受教育水平	ln(收入/1000)	工作日上网时间与休闲时间之比
互联网络重要性的认知	Pearson Correlation Sig. (2-tailed) N						
年龄	Pearson Correlation Sig. (2-tailed) N						
社会地位的感知	Pearson Correlation Sig. (2-tailed) N						
受教育水平	Pearson Correlation Sig. (2-tailed) N						
ln(收入/1000)	Pearson Correlation Sig. (2-tailed) N			4387	4156	4449	4313
工作日上网时间与休闲时间之比	Pearson Correlation Sig. (2-tailed) N	.211** .000 1740	-.376** .000 5832	.000 .985 5478	.436** .000 5420	.285** .000 4313	1.000 .000 5832

** Correlation is significant at the 0.01 level (2-tailed).

Table 2. Influential factors on digital consciousness
表 2. 影响数字化意识的自变量

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	上网时间与工作日休闲时间之比	1.100	.263	17.470	1	.000	3.003
	Ln(收入/1000)	.347	.076	20.919	1	.000	1.416
	受教育水平	.462	.082	31.697	1	.000	1.587
	年龄	-.019	.005	13.798	1	.000	.982
	社会地位感知	.188	.081	5.360	1	.021	1.206
	Constant	-2.034	.370	30.233	1	.000	.131

a. 进入罗吉斯蒂回归方程的自变量包括：时间、收入、受教育水平、年龄、社会地位感知

978-1-935068-76-1 © 2011 SciRes.

表 2 显示了各个自变量对受访者数字化意识的影响程度。如果受访者的工作日上网时间与休闲时间的比值上升 1 个单位，他（她）改变“互联网络对学习和工作并不重要”观点的发生比为 3.003，是原来时间比水平上发生概率的 3.003 倍，新的发生概率增加了 2.003 倍；如果受访者的受教育水平提高 1 个档次，如从小学水平变为初中水平时，他（她）改变“互联网络对学习和工作并不重要”观点的发生比为 1.587，比原来教育水平上该观点发生概率增加了 0.587 倍；如果受访者的收入（以千元为单位）增加到原来的 e 倍（约 2.7182 倍），即 ln(收入/1000)增加 1 个单位，他（她）改变“互联网不重要”观点的发生比为 1.416，发生概率增加 0.416 倍；如果受访者对自身社会地位的感觉和评价提高 1 个档次，其改变“互联网不重要”观点的发生比为 1.206，新水平上的发生概率增加 0.206 倍。年龄对受访者数字化意识的影响是负面的，年龄越大，受访者改变“互联网不重要”观点的概率越低。这五个变量的 sig 值小于 0.05，它们对受访者数字化意识的影响都具有统计意义，均能提供显著的解释水平。有关数字化意识的罗吉斯蒂回归方程为：

$$P=\frac{1}{1+e^{-(0.462*E+0.347*\ln(I/1000)+0.188*S+1.1*T-0.019*A-2.034)}}$$

其中，E 为受教育水平，I 为收入，S 为社会地位感知，T 为上网时间与工作日休闲时间之比，A 为年龄。

6 结论与建议

数字贫困社群中的意识贫困群体受到了收入、受教育水平、社会地位感知、年龄等的影响，“时间比”变量仅对那些有条件使用电脑或上网、但认为电脑或互联网不重要且偶尔去实践的人群有作用。如果能够尽可能提高他们的受教育水平和实际收入，从而改变他们对自身社会地位的认知，应该能够帮助他们改变对电脑或互联网重要性的认识。数字意识贫困社群所拥有的经济资本、文化资本和社会地位等均影响着其所处的数字贫困状态。后续研究还需要收集全国范围内的调研数据，来验证和推广京沪粤三地的研究发现。

如从政策角度来理解研究发现，我们可以得到如下结论：为提高意识贫困社群的数字化地位，应该从根本上尽可能提高其受教育程度，在普及九年义务教育基础上，推广高中和高等教育以及信息素养教育；拓宽该群体的增收途径，增加其可支配收入；建立各项有利于维护其尊严的制度如社会保障制度、民主制度等。此外，中央政府应着手建立公益信息制度，为数字贫困社群提供持续的“信息福利”，以实现数字公平的愿景。

致　谢

感谢北京大学赖茂生教授对本研究的指导；感谢北京大学中国社会科学调查中心授权笔者使用 CFPS2008 数据库。

References (参考文献)

[1] LUKE T. The Politics of Digital Inequality: Access, Capabilities, and Distribution in Cyberspace[J]. *New Political Science*, 1997, 41/42: 121-144.

[2] HARGITTAI E. The digital reproduction of inequality[M]// GRUSKY D.B. Social stratification: class, race, and gender in sociological perspective. Philadelphia: Westview Press, 2008: 936-944.

[3] VAN DIJK J. A.G.M. Digital divide research, achievements and shortcomings[J]. *Poetics*, 2006, 34: 221-235.

[4] MOSSBERGER K., TOLBERT C. J., STANSBURY, M. Virtual Inequality: Beyond the Digital Divide[M]. Washington DC: Georgetown University Press,2003:1-14.

[5] YAN H. Digital inequality in communities and information institution for public welfare[D]. Beijing: Peking University, 2010.
闫慧. 社群数字不平等问题及公益信息制度研究[D]. 北京大学博士学位论文, 2010.

[6] CARTIER C., CASTELLS M., QIU, J. L. The Information Have-Less: Inequality, Mobility, and Translocal Networks in Chinese Cities[J]. *Studies in Comparative International Development*,2005,40(2): 9-34.

[7] DIMAGGIO P., HARGITTAI E., CELESTE, C., et al. Digital inequality: from unequal Access to differentiated use[M]// NECKERMAN K.M. Social Inequality. New York: Russell Sage,2004: 355-400.

[8] AZARI R., PICK J.B. Understanding global digital inequality: The impact of government, investment in business and technology, and socioeconomic factors on technology utilization: Proceedings of the 42nd Annual Hawaii International Conference on System Sciences, Waikoloa, January 5-8, 2009[C]. Hawaii: IEEE Computer Society:1-10.

[9] HSIEH P. A. Leverage points for addressing digital inequality: An extended theory of planned behavior perspective[D]. Atlanta: Georgia State University, 2005

[10] KVASNY L. M., KEIL M. The challenges of redressing the digital divide: a tale of two US cities[J]. *Information Systems Journal*, 2006,16:23–53.

[11] VAN DIJK J. A.G.M. The deepening divide: inequality in the information society. Thousand Oaks: Sage, 2005: 9-14.

[12] WILLIS S.,TRANTER B. Beyond the 'digital divide': Internet diffusion and inequality in Australia[J]. *Journal of Sociology*, 2006,42(1):43-59.

[13] ALVAREZ A.S. Behavioral and Environmental Correlates of Digital Inequality[J]. *IT&Society*,2003,1(5): 97-140.

[14] GIBSON C. Digital divides in New South Wales: A research note on socio-spatial inequality using 2001 Census data on computer and Internet technology[J]. *Australian Geographer,* 2003, 34(2): 239-257.

[15] GIL-DE-ZUNIGA, H. Reshaping digital inequality in the European Union: How psychological barriers affect internet adoption rates[J]. *Webology*,2006,3(4):32.

[16] CASTELLS M. The internet galaxy: reflections on the internet, business, and society[M]. Oxford: Oxford University, 2001: 247-274.

[17] FUCHS C. The role of income inequality in a multivariate cross-national analysis of the digital divide[J]. *Social Science Computer Review*,2009,27(1):41-58.
[18] RIDEOUT V. Digital inequalities in Eastern Canada[J]. *Canadian Journal of Information and Library Science*, 2003, 27(2): 3-31.
[19] HO C.C., TSENG S.F. From digital divide to digital inequality: the global perspective[J]. International *Journal of Internet and Enterprise Management,* 2006, 4(3): 215.
[20] KVASNY L. M. Cultural (RE)production of digital inequality in a US community technology initiative[J]. Information Communication and Society, 2006,9(2):160-181.
[21] LEVINSON, N.S., HERVY A.C. Digital Inequalities: Technology, Development and Cross-National Alliances: Annual Meeting of American Political Science Association, Chicago, IL, September 2-5, 2004[C]. Washington, DC: APSA, 1-22.
[22] KVASNY L. M. A Conceptual Framework for Studying Digital Inequality: BANKER R.D., CHANG H, KAO Y.C.: Proceedings of the Eighth Americas Conference on Information Systems (AMCIS 2002), Dallas, TX, August 9-11, 2002[C]. Dallas: University of Texas at Dallas, 1798-1805.
[23] HUSING T., SELHOFER H. DIDIX: A Digital Divide Index for Measuring Social Inequality in IT Diffusion [J]. *IT & Society*, 2004, 1(7): 21-38.
[24] ONO H., ZAVODNY M. Digital inequality: a five country comparison using microdata[J]. *Social Science Research*, 2007, 36(3): 1135-1155.
[25] JUNG J., QIU J., KIM Y. Internet Connectedness and Inequality: Beyond the "Divide"[J]. *Communication Research,* 2001, 28(4): 507-535.
[26] HSIEH P. A., RAI A., KEIL M. Understanding digital inequality: comparing continued use behavioral models of the socio-economically advantaged and disadvantaged [J]. *MIS Quarterly*, 2008, 32(1):97-126.
[27] GUO Z.G. Social statistic analysis methods: applications of SPSS. Beijing: Renmin University of China Press, 1999: 182-183.
郭志刚. 社会统计分析方法：SPSS 软件应用[M]. 北京: 中国人民大学出版社, 1999: 182-183.

978-1-935068-76-1

Optimization of the Evaluation Indicators of Scholars' Research Impact and Comparative Analysis between National and International Academic Behaviors of Researchers: A Perspective of Scientific Age

Jian DU, Bin ZHANG*, Yang LI, Xiaoli TANG, Peiyang XU
Institute of Medical information, Chinese Academy of Medical Sciences, Beijing, 100005

Abstract: Taking 73 doctoral supervisors as the sample who devoted on clinical research and now are working in Chinese Academy of Medical Sciences, the paper analyzed the difference of research performance on several indicators, and tried to assist in optimizing the evaluation indicators by using the statistical methods of ANOVA and correlation analysis. The optimized evaluation indicators include number of articles, total cited times, cited times per paper, h-index, A-index, g-index, and accumulated impact factor of journals. Average counts related indicators (e.g. cited times per paper and A-index) are more suitable for assessing China's high level of academics (e.g. Academicians). The indicator of m quotient is not suitable for evaluating international impact of Chinese scholars. The evaluation of China's researchers should pay attention to the factor of scientific age.

Keywords: Research Impact; China's Scholars; Scientific Age; Citation Analysis

学者学术影响力评价指标的优选与学术行为特点的国内外比较

杜建，张玢*，李阳，唐小利，许培扬
中国医学科学院医学信息研究所，北京，100005

摘　要：以中国医学科学院 73 位从事临床科研的博士生导师为样本，分析不同学术年龄组在各学术影响力指标上的表现差异，并采用方差分析和相关性分析辅助优选指标。得到的优选指标包括发文量、总被引次数、篇均被引次数、h 指数、A 指数、g 指数和累计影响因子，其中与"平均量"相关的篇均被引次数、A 指数更适合于评价中国高水平学者的学术表现；而 m 熵指数不太适合于评价中国学者的国际影响力；中国科研人员的评价应该注意学术年龄这一因素。

关键词：学术影响力；学者；学术年龄；引证分析

1 引言

应用文献计量学指标对宏观和中观层面（国家、地区、大学、机构、期刊等）的评价研究较为广泛也更为人们接受。然而，微观层面（如学者、研究团队）评价的复杂性和各种评价指标的劣势使得该层面的评价总是存在着诸多矛盾和争议[1-2]。目前，所有的基于引证分析的学者学术影响力评价指标，如总被引次数、篇均被引次数、h 指数等都有缺点，因此单独依靠一种指标进行评价是不科学的[3]。Van Leeuween 早在 2003 年就强烈建议若干个指标的联合应用[4]。但迄今为止，也未见在微观层面联合应用若干个文献计量学指标的具体方法和相关实证。另外，这些学术影响力评价指标又与科学家从事学术生涯的时间长短（即学术年龄）密切相关。一般来说，越年老的科学家的总被引次数和 h 指数越高。实际上，美国物理学家 Hirsch JE 在提出 h 指数的同时也提出了用学术年龄对 h 指数进行划分的思想[5,6]。为此，本文选取中国医学科学院 73 位从事临床科研的博士生导师为样本，以学术年龄（最新论文与最早论文的发表年之差）为主要控制变量，采用方差分析和相关性分析方法，探讨不同学术

基金项目：中国医学科学院医学信息研究所中央级公益性科研院所基本科研业务费"基于引证分析的学术影响力理论与实践研究"（项目编号：09R0216）
*张玢为通讯作者，E-mail: zhang.bin@imicams.ac.cn

年龄组的学者在各个指标上的表现差异，对基于引证分析的学者学术影响力评价指标进行优选，并分析不同学术年龄组在国内外学术行为的不同特点，以期为科研人员学术影响力综合评价的实证研究提供理论基础和事实依据。

2 样本与方法

2.1 样本选择

考虑到h指数及h型指数不适合考察年轻科学家的学术影响力，按照姓名列表从中国医学科学院/北京协和医学院423位博士研究生导师列表中选取73位临床科研人员为研究对象。分别根据其教育和工作经历，获取这73位学者发表的国内外论文及被引用数据。

2.2 数据来源

国内论文来源于中国生物医学文献数据库（CBM）、CNKI和万方数据，并利用CNKI中国引文数据库和中国科学引文数据库（CSCD）检索其被引用数据。国外利用Web of Science中的SCI-E和2009版《期刊引证报告》(JCR)检索73位学者发表的国际论文、被引用数据和论文所在期刊的影响因子。

2.3 评价指标

根据上述检索结果计算得到 73 位学者在国内外的发文量、总被引次数、篇均被引次数、h 指数、g 指数、A 指数、m 熵指数、SCI 论文所在期刊的累计影响因子、平均影响因子等指标数据，应用这些指标评价学者的国内和国际学术影响力。对相关指标的定义简介如下。

（1）h 指数

由美国物理学家 Hirsch JE 于 2005 年提出，表示一位作者发表的论文中，有 h 篇至少被引用了 h 次[6]。与传统的引文分析法相比，h 指数兼顾个人科学产出的质量和数量[7]。

（2）g 指数

2006 年，比利时著名科学计量学家 Egghe L 提出了 g 指数[8]，即将论文按被引次数由高到低排序，将序号平方，被引次数按序号层层累加。当序号平方等于其对应累计被引次数时，该序号则为 g 指数。如序号平方不是恰好等于而是小于对应的累计被引次数，则最接近累计被引次数的序号即为 g 指数。与 h 指数相比，g 指数只对一篇或几篇突出的、高影响力的论文敏感。

（3）A 指数

2007 年，金碧辉和 Rousseau R 提出了 A 指数[9]，即纳入 h 指数的论文的篇均被引次数，衡量的是纳入 h 指数的论文的平均被引用强度。A 指数可以克服 h 值几年不变的呆滞局面以及出现大量相同的 h 值而导致区分度不大的问题。

（4）m 熵指数（m quotient）

即年均 h 指数，由 Hirsch JE 提出。它根据科学家从事学术生涯的年份数对 h 指数进行划分，从而为不同资历的研究者之间的学术评价提供了可能，但 m 熵指数不能甄别出高被引论文的影响力[10]。

（5）累计影响因子和平均影响因子

由于引用活动本身的时滞特征，总被引次数很难应用在较近的出版物上。JCR 提供了期刊的影响因子(IF)用以反映该期刊刊载论文的平均影响力，可以根据论文发表的期刊来推断论文可能产生的被引情况，即期望被引次数，从而使用累计影响因子(Cumulative Impact Factor,CIF)和平均影响因子(Average Impact Factor,AIF)来替代被引次数指标。CIF 和 AIF 的计算公式如下：

$$CIF = \sum_{i=1}^{n} IF_i \qquad AIF = \frac{1}{n}\sum_{i=1}^{n} IF_i$$

其中 n 代表该学者的论文总数，IF_i 是第 i 篇论文所在期刊的影响因子。

需要说明的是，期刊不同年份的影响因子会有一定差别。本文使用的影响因子数据来源为 2009 版 JCR，少数未被收录的期刊，则按倒序顺序分别 2008、2007、…、2001 年版 JCR 的数据。而

3 结果分析

3.1 学术年龄及其分区

本数据集中，73 位学者都是在大学毕业或读研究生期间开始发表第一篇论文，最晚发表论文年均为 2010 年，目前在科研产出上仍很活跃。图 1 表示了学术年龄的频数分布。

图1可见，学术年龄在整体上趋近于正态分布（Kolmogorov-Smirnov Z= 1.102，p= 0.176>0.05）。平均学术年龄为25.42年，最小值为13，最大值为46。考虑到73位学者的特征，以25年为分界点，按学术年龄分为4个组，见表1。第1组学者为1994年后开始发表文献，第2组1987年后开始发表文献，第3组1977年后开始发表文献，将学术年龄最大的院士群体单独划分在第4组。

978-1-935068-76-1 © 2011 SciRes.

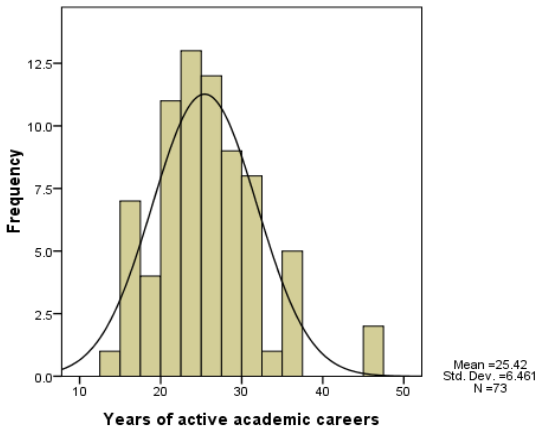


Figure 1. Distribution of frequency of years of active academic careers
图1. 学术年龄的频数分布图

Table 1. Groups divided by years of active academic careers
表1. 学术年龄分组

组别	学术年龄(年)	学术生涯的起始年	频数	比例(%)
1	13-17	1994-1998	8	11.0
2	18-24	1987-1993	28	38.4
3	25-34	1977-1986	30	41.1
4	35-46	1965-1976	7	9.6
总和			73	100.0

上表可见，除“最年轻”的第1组外，其他组的学术年龄跨度均约为10年。从频数上看，学术年龄呈现出“两头小，中间大”的特征。第1组和第4组的频数均占10%以上，而作为主体的第2组和第3组则分别约占40%。另外，第4组是比较特殊的群体，他们在1976年之前开始学术活动，且大多为中国科学院/中国工程院院士。本数据集中的6位院士，其中5位在此组。

3.2 基于引证分析的学者学术影响力评价指标的优选

3.2.1 主要引证分析指标的优选

应用单因素方差分析方法探讨不同学术年龄的学者在各评价指标上表现的差别。结果表明：国内只在“论文发文量”这一指标上有显著性差异（$p=0.014<0.05$），而在“总被引次数”、“篇均被引次数”、“h指数”、“g指数”、“A指数”、“m熵指数”6个评价指标上的表现均没有显著性差异（$p>0.05$）。国外部分，9个指标在组间差异的p值均远大于0.05，说明学术年龄对上述指标的影响没有显著性差异。朱平2008年对中国科学院资源环境领域某研究所的全体研究员的SCI论文发表和h指数等情况的分析发现，也认为年龄与学术影响有一定的联系，但影响并不大，因为指数之间并无显著的统计差异[11]。

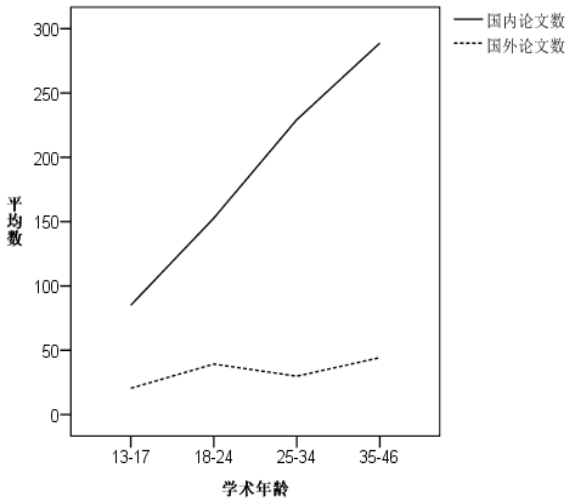


Figure 2. The mean number of publications of 4 groups divided by years of active academic careers
图2. 不同组别学术生涯年份数的学者在国内外发文量的平均数图

图2可见，随着学术年龄的增加，国内发文数和国外发文数在整体上呈现出上升趋势，尤其在国内表现得最为显著。年龄越大，发文数越多，这一点符合国内“学术生涯越长，科研产出越多”的现象。但在SCI发文量上，“最年长”的35-46岁组和“较为年轻”的18-24岁组大致持平，突显了年轻学者在SCI学术产出方面的优势。

综合考虑上述国内和国外的方差分析结果，以下仅从平均数（Mean）的角度，分析不同组别学术年龄的学者在各个指标上的表现差异，见图3（包括6个子图）。

图3可见，关于总被引次数、h指数和g指数3个指标的表现，国内最占优势的是“25-34岁”组（实心三角形标识），而国外最占优势的则为“18-24岁”组（空心三角形标识）。国内的表现均呈现出前3个年龄段一直升高，最后年龄段下滑的趋势；国外则正好相反，从第1个至第2个年龄段升高，后2个年龄段下滑。因此，这3个指标对于不同学术年龄学者在国内外的评价结果在总体上是一致的。由于总被引次数是传统的基于引证分析的评价指标，而h指数和g指数是针对传统不足提出的新指标，且g指数又是对h指数的一种弥补。因此，我们认为这3个指标都应保留。

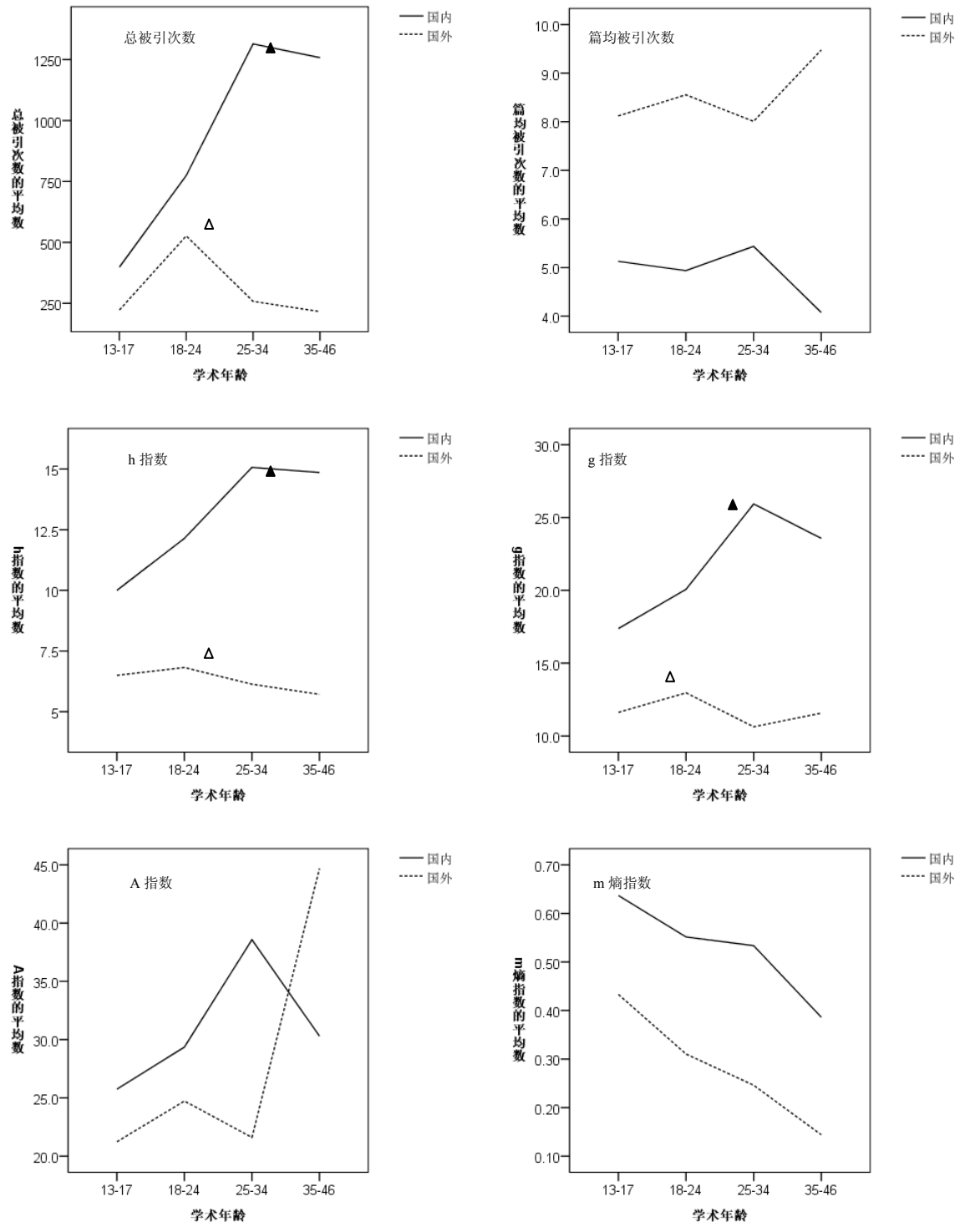


Figure 3. The mean performance of 6 indicators of 4 groups divided by years of active academic careers
图3. 不同组别学术年龄的研究人员在6个指标上表现的平均数图

学术年龄最大（35-46岁）的院士群体在总被引次数、h指数和g指数上的优势地位均不明显。但值得注意的是，科学界公认的学术影响力最高的院士的优势主要体现在篇均被引次数和A指数上，尤其是SCI论文方面。由于篇均被引次数表征学者发表论文的平均影响力，A指数主要反映了纳入h指数的论文的篇均被引

978-1-935068-76-1 © 2011 SciRes.

次数，且对高影响力的论文比较敏感，说明院士群体可能只拥有少量的、但影响力高的、创新性强的代表作，且这些作品几乎都是“精品”，一直被人们引用和借鉴。篇均被引次数和A指数在一定程度上更能评价出高影响力学者的学术表现。

m熵指数考虑了学术年龄对h指数的影响。图中可见随着学术年龄的增加，m熵指数显著下降。但从h指数和g指数两图可得，对于中国学者，较年轻的科研人员就已经比年长的学者获得了更高的h指数和g指数。再用学术年龄的长短对h指数进行划分显然不太科学。

3.2.2 附加指标的优选：累计影响因子和平均影响因子

图4为学者发表的SCI论文所在期刊的累计影响因子和平均影响因子的平均数图。

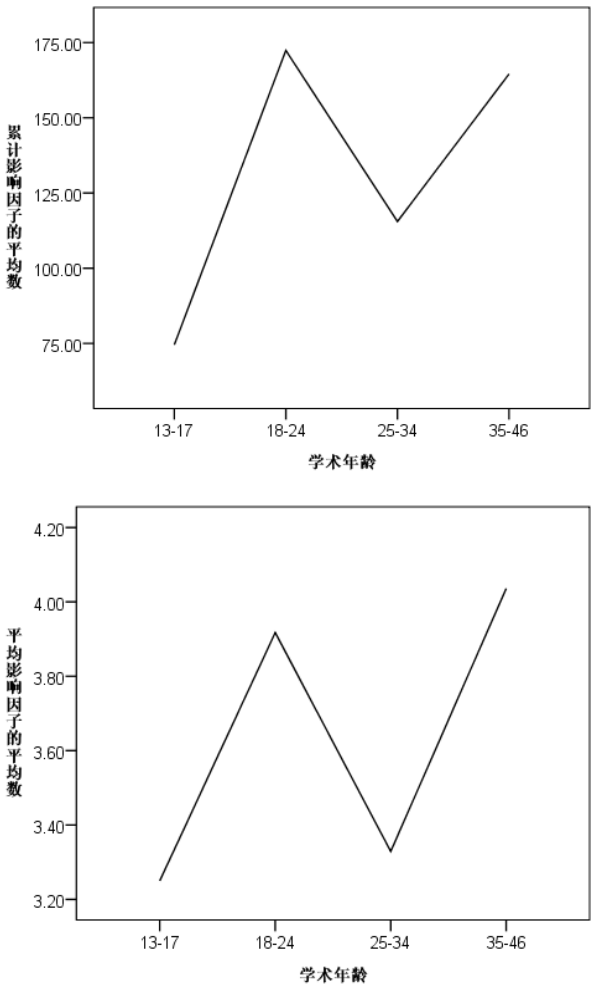


Figure 4. The mean number accumulated impact factor (CIF) and average impact factor (AIF)
图4. 累计影响因子和平均影响因子的平均数图

Table 2. Spearman correlation analysis of ‘the validated indicators’, CIF and AIF
表2. “已验证的指标”与CIF和AIF的Spearman相关性分析

	累计影响因子(CIF)	平均影响因子(AIF)
总被引次数	.757**	.337**
篇均被引次数	.127	.443**
h指数	.693**	.122
A指数	.480**	.470**
g指数	.758**	.300**

注：**表示在0.01的水平显著，*表示在0.05的水平显著，没有标注表示不显著

上图可见，对于累计影响因子和平均影响因子，学术年龄为18-24的学者都表现出明显的优势，学术年龄为35-46的学者的平均影响因子也很高。图4显示，累计影响因子对不同学术年龄研究人员的区分更为明显。由于累计影响因子和平均影响因子的计量对象均为期刊，均表征一种“期望的引证影响力”，我们认为不宜过多地引入这类指标。为此，继续分析上述已验证的总被引次数、篇均被引次数、h指数、A指数和g指数5个指标与累计影响因子和平均影响因子的相关性，选择与“已验证的指标”相关性高的指标作为优选指标，结果见表4。

从总体上看，除篇均被引次数外，其他4个“已验证的指标”与累计影响因子均呈现出较强的正相关，且比与平均影响因子的相关性更为显著。因此选择累计影响因子指标。实际上，累积影响因子已在中观层面（大学）的评价中得以应用。程莹和刘念才[12]从机构层面应用SCI-E和SSCI论文所在期刊的累计影响因子对首批建设的9所“985工程”大学进行了评价，与通过师均论文数、论文平均质量等指标的评价结果较一致。

经过上述分析，优选出的评价指标为发文量、总被引次数、篇均被引次数、h指数、A指数和g指数。国际影响力评价时再增加累计影响因子指标。

3.3 不同学术年龄的学者在国内外学术行为的特点

为了解学术年龄与科研人员学术产出、学术影响力的关系，揭示不同学术年龄的学者在国内外学术行为的特点，以学术年龄为控制变量，分析发文量、h指数、g指数、A指数4个指标国内外表现的差异。下图可见，国内外学术产出（发文量）随着学术年龄变化的趋势较为一致，然而在学术影响力的相关指标（如h指数、A指数和g指数）上却出现了差异，结果见图5和图6。

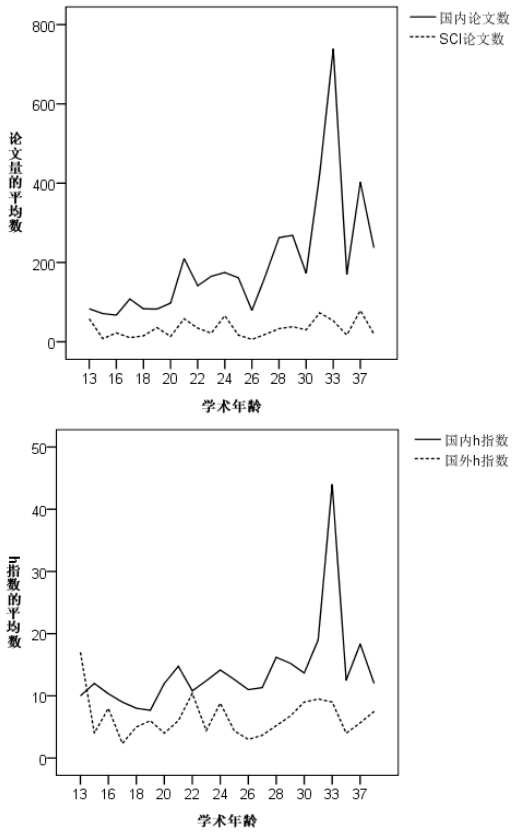
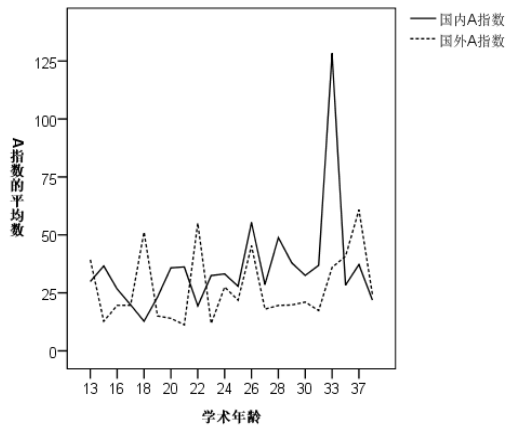


Figure 5. Distribution of the mean number of publications and the mean h-index of researchers with different years of active academic careers

图5. 不同学术生涯年份数的学者在国内外发文量和h指数的分布

图5可见，国内h指数的最高值发生的学术年龄为33，比较符合国内的规律，即从事学术生涯的时间越长，越容易获得较高的h指数。而国外h指数的最高值发生的学术年龄则最小，这在一定程度上反映了国内年轻学者主要在发表SCI论文上占有优势。



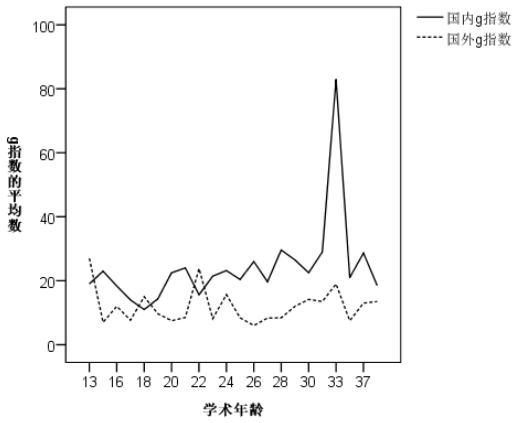


Figure 6. Distribution of the mean g-index and A-index of researchers with different years of active academic careers

图6. 不同学术生涯年份数的学者在g指数和A指数的分布

图6可见，以24岁为分界点（图中红线所示，本数据集学者的平均学术年龄为25岁），A指数和g指数所表征的国内外差异可以将样本明显分为两部分，即年轻学者和年长学者，两部分表现出不同的学术特征。在学术年龄为24岁之前，国内A指数在波峰时，国外A指数却在波谷；国内A指数在波谷时，国外A指数却在波峰。而学术年龄为24-46之间时，其波峰和波谷则呈现出较为一致的趋势。考虑到A指数衡量了少量纳入h指数的论文的平均引用强度，且g指数对于高影响力的论文比较敏感。笔者认为，学术年龄为13-23的年轻学者可分为两类，一类是注重于在国外发表学术论文，其在国内投入的精力则较少；另一类则恰好相反。而对于23-46岁的相对年长的学者，则同时兼顾了国内外的学术产出与交流，在国内外的学术表现较为一致。

4 讨论与结论

4.1 学术年龄对学者学术表现的影响

学术年龄对于学者的学术表现是一个有影响的因素。Falagas M E和Ierodiakonou V等人对生物医学领域的科学家作出重大贡献的最佳年龄的研究显示，学者的学术产出力随着年龄而递减[13]。而Costas R和van Leeuwen TN等人研究发现在西班牙国家研究理事会工作的生物学与生物医学领域的学者发文量和篇均被引次数的年龄分布（以5年为一个跨度）呈现出倒置的“U”型[14]，即先升高、后减少的趋势。国内学者雷二庆，陈红光[15]选择2007年度145位中国科学院院士增选初步候选人为样本，应用SCI-E数据研究发现h指数与年龄之间没有明显相关性，认为用h指数评价中国科

978-1-935068-76-1 © 2011 SciRes.

学家的学术影响力失灵。究其原因，主要是受专业，学科、年龄、姓名，机构等因素的影响。本研究的学术年龄分区考虑了样本的具体特征，对于学术年龄最高的院士群体，主要在表征“平均影响力”的指标（如篇均被引次数、A指数）上占有最显著的优势。而与较之更年轻的学者相比，公认的最年长、影响力最高的院士群体在由总被引次数、h指数、g指数等与发表论文总量密切相关的指标上的表现却不具备优势。笔者认为原因可能有以下几点。

（1）学者逐渐变老的过程中，他们越来越多的参与到除科学研究之外的一些活动中，如行政、教学、研究评估、项目管理，基金资助，指导博士生等。同时，老资历的学者可能逐渐改变了他们的学术交流习惯，如撰写专著、书评、研究报告等，而这些文献类型是Web of Science中检索不到的。

（2）受历史因素的影响，我国60岁以上的科学家（如很多院士）所掌握的第一外语通常是俄语，而不是英语，语种因素限制了老科学家的SCI论文产出要低于年轻科学家，进而老科学家的h指数、g指数通常也不如年轻科学家高。

4.2 评价指标对于中国学者的适用性

h指数、g指数等指标不太适合于评价中国高水平学者的学术影响力，而与“平均量”相关的篇均被引次数和A指数在一定程度上更能评价出我国高影响力学者（如院士）的学术表现。m熵指数不太适合评价国内学者。

4.3 中国学者国内外学术影响力表现的差异

年轻科研人员可分为两类：一类是注重在国外发表学术论文，在国内投入的精力则较少；另一类则把注意力更多的放在国内。而对于学术年龄相对年长的学者，则往往同时兼顾了国内外的产出与交流。因此，科研人员评价应该注意学术年龄这一因素，如“新世纪百千万人才工程国家级人选”的评选对象的年龄均在45周岁以下[16]，国家杰出青年科学基金申请者的年龄要求也是未满45周岁[17]。

本文从学术年龄的视角对基于引证分析的学者学术影响力评价指标进行了优选并对其国内外表现进行了比较，为后期综合利用优选的文献计量学指标对同学科领域的研究人员评价的实证研究提供了理论基础和事实依据。但尚未考虑到多作者合作中作者名誉分配的问题，基于作者贡献的h指数等评价指标的优化以及实证研究将是下一步探讨的主要方向。

References (参考文献)

[1] Costas R, Bordons M. Bibliometric indicators at the microlevel: Some results in the area of natural resources at the Spanish CSIC[J].Research Evaluation, 2005,14(2):110–120.
[2] Sandström U, Sandström E. Meeting the micro-level challenges: Bibliometrics at the individual level. Paper presented at the 12th International Conference on Scientometrics and Informetrics, Rio de Janeiro, Brazil,2009.
[3] Costas R, Bordons M. Is g-index better than h-index? An exploratory study at the individual level [J]. Scientometrics, 2008, 77(2):267-288
[4] Van Leeuwen TN, Visser MS, Moed HF,et al. The Holy Grail of science policy: Exploring and combining bibliometrics in search of scientific excellence [J]. Scientometrics, 2003, 57(2):257–280.
[5] ZHANG Bin, DU Jian, WANG Min, ZHANG Yan-wu, A Lita, LIU Xiao-ting, LI Hai-cun, XU Pei-yang. Review on Citation Analysis Indicators for Evaluation Academic Influence [J]. Journal Of Medical Intelligence, 2010, 31(12): 40-46.
张玢,杜建,王敏,等.评价学术影响力的引证分析指标研究综述[J].医学信息学杂志,2010,31(12): 40-46.
[6] Hirsch J E. An index to quantify an individual's scientific research output [J]. Proceedings of the National Academy of Sciences of the United States of America, 2005,102(46): 16569-16572.
[7] Ding Nan, Pan Youneng. An Empirical Study of H-index and G-index Based on CSSCI [J]. LIBRARY AND INFORMATION, 2008 (2): 79-82.
丁楠,潘有能.h 指数和 g 指数评价实证研究——基于 CSSCI 的统计分析[J].图书与情报,2008 (2):79-82.
[8] Egghe L. Theory and practise of the g-index [J]. Scientometrics, 2006, 69(1):131-152.
[9] Jin Bihui1,Rousseau Ronald. R-index and AR-index: Complementing Indicators to h-index [J]. Science Focus, 2007(3):1-8
金碧辉,Rousseau Ronald. R 指数、AR 指数:h 指数功能扩展的补充指标[J].科学观察,2007(3):1-8
[10] Hirsch JE. An index to quantify an individual’s scientific research output [J]. Proc Natl Acad Sci.2005, 102(46): 16569-72.
[11] Zhu Ping. Applicability of H-index in evaluating research performance of researchers [J]. Science Research Management, 2009, 30(S1):90-93.
朱平. H 指数在科研人员科研成就评价中的适用性分析[J]. 科研管理,2009,30(S1):90-93.
[12] CHENG Ying,LIU Niancai. Mapping the Development of Chinese Top Universities in Recent Years Based on Scientometric Indicators [J]. Science of Science and Management of S.& T, 2007, 28(9): 132-138
程莹,刘念才.从科学计量学指标看我国名牌大学近年来的发展[J].科学学与科学技术管理,2007,28(9):132- 138
[13] Falagas M E, Ierodiakonou V, Alexiou V. At what age do biomedical scientists do their best work? The FASEB Journal, 2008, 22, 1-4.
[14] Costas R, van Leeuwen TN, Bordons M. A Bibliometric Classificatory Approach for the Study and Assessment of Research Performance at the Individual Level: The Effects of Age on Productivity and Impact [J]. Journal Of The American Society for Information Science and Technology,2010,61(8): 1564-1581
[15] Lei Erqing, Chen Hongguang. Applicability of H-index in evaluating Chinese scientists [J]. Science Research Management, 2009,30(S1):84-89.
雷二庆,陈红光. H 指数评价中国科学家的适用性及影响因素分析——以 2007 年中国科学院增选院士候选人为例[J]. 科研管理,2009,30(S1):84-89.
[16] Ministry of Human Resources and Social Security of the People's Republic of China. [EB/OL].[2010-11-23]. http://www.mohrss.gov.cn/Desktop.aspx?path=mohrss/mohrss/InfoView&gid=8cef35c2-429c-4435-8ac6-fd61ff5c167d&tid=Cms_I

nfo
中华人民共和国人力资源和社会保障部, 新一批“新世纪百千万人才工程”国家级人选产生[EB/OL].[2010-11-23]. http://www.mohrss.gov.cn/Desktop.aspx?path=mohrss/mohrss/InfoView&gid=8cef35c2-429c-4435-8ac6-fd61ff5c167d&tid=Cms_Info

[17] National Natural Science Foundation of China. [EB/OL]. (2009-12-11) [2010-11 -23]. http://www.nsfc.gov.cn/Portal0/InfoModule_506/28397.htm
国家自然科学基金委员会, 国家杰出青年科学基金项目管理办法[EB/OL].(2009-12-11)[2010-11-23]. http://www.nsfc.gov.cn/Portal0/InfoModule_506/28397.htm

978-1-935068-76-1

Investigating the Information Searching Behavior of Academic Users

Bai CHEN
Institute of Scientific and Technical Information of China, Beijing China
Email: baichen001@hotmail.com

Abstract: Investigating the behavioral characteristics and the adaptive learning mechanism of users during their information searching is meaningful for the academic database providers to develop more effective search products and online learning environment. Based on the theories and models of reinforcement learning behavior, this research takes the freshmen and senior students from universities as user samples, experimentally observes the explicit behavioral and implicit psychological characteristics of their reinforcement learning in the process of search tasks performing, and further quantitatively simulates their reinforcement learning behavior in information seeking using the Bush-Mosteller model, Borgers-Sarinare model and Cross model. Finally, the paper gives some advices for the academic database providers to improve their service and build e-learning platform to help the users manipulate the search products more effectively and efficiently.

Keywords: reinforcement learning behavior; information searching; academic users; behavior analysis and investigating

1. Introduction

As a primary method for researchers to obtain necessary information, academic search plays an important part in scientific research and knowledge learning. To facilitate access to high quality literatures, many academic database providers have invested a great deal to build comprehensive electronic content collections and offer a variety of search products/functionalities, including traditional simple search (which provides single textbox for quick search), advanced search (which provides multiple textboxes to support more controlled and pinpoint search), expert search (which allow users to form complex queries combined with keywords, field codes, Boolean operators, wildcard or truncations), as well as other special purpose products such as number search, image search and academic trend search. Maximizing the return on these huge investments is the key for the database companies or even the nonprofit digital libraries to survive and develop. To achieve that, the academic database providers have to constantly update the content and improve the service based on their insight into users' characteristics and requirements so as to enable the users to maximize the effectiveness of their knowledge discovery process. Also the database providers should understand whether the electronic resources and search capabilities offered are sufficiently and productively utilized by academic users. In other word, it is of great significance to investigate whether the users can choose the most appropriate search product and correctly use it to obtain the information desired.

Unfortunately, according to our previous research, many academic users (including the novice students, senior students and graduate students in universities, as well as experienced researchers) could not control the elaborately designed search functions effectively. We have organized searching experiments in which some university users with different academic backgrounds were asked to perform some predefined search tasks through a famous academic database system in China, CNKI (http://www.cnki.net). To finish these tasks, the participants might have to use the multi-textbox search function provided by the system. However, our study proved that most participants preferred initially to choose simple search and input keywords in a single textbox to finish searching. Obviously, the functionalities offered by the database websites are not fully understood by the end users, and therefore the values of the database systems are not completely delivered to the users. Thus, it is necessary for the academic users to learn the skills for more effective information seeking. At the same time, it is necessary for the academic database providers to develop an environment more conducive to users' adaptive learning of search skills.

In response to these challenges, this paper focuses on investigating the characteristics of users' information search behavior and the learning mechanism by which the users adjust their strategies during information searching. Probe to these problems is meaningful for the academic database providers to develop online learning platforms and offline services to guide, help or train the users to manipulate the search products more effectively and efficiently. For those potential users who grew up with the Internet and always resort to e-platforms for problem solving, a real-time, convenient and effective e-learning environment for academic search is especially helpful.

The paper comprises three main parts: the first, Theo-

ries and models, explains the basic idea of reinforcement learning behavior, and introduces three representative reinforcement learning models, i.e. Bush-Mosteller model, Borgers-Sarinare model and Cross model; the second, Experiments and analysis, taking the freshmen and senior students from universities as user samples, experimentally observes the explicit behavioral and implicit psychological characteristics of their reinforcement learning in the process of search tasks performing, and further quantitatively simulates their reinforcement learning behavior in information seeking using the three learning models; and the third, Conclusions, makes some suggestions for the academic database providers to improve their service and develop e-learning platform.

2. Theories and Models

2.1. Reinforcement Learning Behavior

According to the learning theory of cognitive behaviorism [1], learning is a process that animals and humans could obtain experience and change their behavior to gain a better feedback. And humans share with other animals a simple way of learning, which is usually called reinforcement learning. This reinforcement learning seems to be biologically inherent. If an action leads to a negative outcome, i.e. a punishment, this action will be avoided in the future. Otherwise, if an action leads to a positive outcome, i.e. a reward, it will reoccur [2].

During the process of search knowledge learning, a user manifests similar reinforcement characteristics: when the user completes an information search task by a certain strategy, he/she will evaluate this process in terms of time cost, quantity of relevant results achieved, and form a tendency to or not to adopt this strategy or switch to other strategies in next tasks.

Consider the evolution of an individual user's information seeking behavior. He/she may experience a process from not knowing how to use, never using, or being not familiar with the information retrieval system to well controlling the system. This sequence of behavior adjustments is a dynamic learning process according to learning theories. The strategy formulation process of information seeking is a process of dynamic alignment, in which a user's strategy in each turn may not be consistent, and the adjustment of behavior is reached from the user's database using experience under the influence of external stimulations (tips on the system interface, training or communication) and based on user's knowledge about the available strategy set. This process tallies with the description of learning behavior in learning theories.

2.2. Models of Reinforcement Learning Behavior

Enlightened by the basic idea of reinforcement learning, a variety of quantitative models were proposed to fit different learning processes in different situations. Considering the characteristics of information seeking process, three classic models are introduced into our research for quantitative analysis: Bush-Mosteller model (BM model) [3], Borgers-Sarinare model (BS model) [4] and Cross model (CR model) [5]. All these models take the reward or punishment as the basis for the adjustment of strategy attraction of next period. Nevertheless, to the rules of this adjustment, they are not the same.

BM Model. In 1955, psychologist Robert R. Bush and Frederick Mosteller started a quantitative research of learning process and set up the BM model, one of the oldest and most famous learning models.

In this model, the attraction of a strategy is defined as a probability variable $P\cdot$, the strategy which is chosen by a user in period t is defined as $d\ t$, and the reward or punishment fed back to the user in period t is $\pi\ t$ (a non-negative $\pi\ t$ means the user gets a reward, otherwise a punishment).

Suppose in perio t, a user chooses the jth strategy from the strategy set, i.e. $j=(\ t)$. Then for this user, the attraction of strategy j is updated under the rules:

$$p(j,t+1)=p(j,t)+\alpha^{BM}\cdot(1-p(j,t)) \qquad j=d(t)\wedge\pi(t)\geq 0$$

$$p(j,t+1)=p(j,t)-\beta^{BM}\cdot p(j,t) \qquad j=d(t)\wedge\pi(t)<0$$

For each strategy k other than (i.e. strategies not chosen by the user), the attraction value is updated according to the following rules:

$$p(k,t+1)=p(k,t)-\alpha^{BM}\cdot p(k,t) \qquad k\neq d(t)\wedge\pi(t)\geq 0$$

$$p(k,t+1)=p(k,t)+\beta^{BM}\cdot(1-p(k,t)) \qquad k\neq d(t)\wedge\pi(t)<0$$

BS Model. Based on BM model, T. Borgers and R. Sarin proposed another model to explain reinforcement learning behavior. Compared to BM model, BS model details the information for evaluating the payoff of a strategy adoption under the supposition that the evaluation of a strategy choice may not directly rely on the absolute value of actual payoff, but on the difference between the actual payoff and the expected one.

Let $P\cdot$, $d\ t$ and $\pi\ t$ be the variables similar to those in BM model. Besides, let $A\ (t)\in[0,1]$ denote the payoff expectation for a user after employing a strategy in period t and $A\ 1$ be the initial expectation for the user before decision-making.

If $(\ t)\geq A(\ t)$, the attraction values of strategies after period t are updated:

$$p(j,t+1)=p(j,t)+\alpha^{BS}\cdot(1-p(j,t)) \qquad j=d(t)$$

$$p(k,t+1)=p(k,t)-\alpha^{BS}\cdot p(k,t) \qquad k\neq d(t)$$

$$\mathrm{A}(t+1)=(1-\beta^{BS})\cdot\mathrm{A}(\mathrm{t})+\beta^{BS}\cdot\pi$$

Else if $(t)<A(t)$, the attraction values are updated as following:

$$p(j,t+1)=p(j,t)-\alpha^{BS}\cdot p(j,t) \qquad j=d(t)$$

$$p(k,t+1)=p(k,t)+\alpha^{BS}\cdot(1-p(k,t)) \qquad k\neq d(t)$$

$$\mathrm{A}(t+1)=(1-\beta^{BS})\cdot\mathrm{A}(\mathrm{t})+\beta^{BS}\cdot\pi$$

CR Model. CR model was proposed by John G. Cross in

978-1-935068-76-1 © 2011 SciRes.

1973. As a modification to BM model, it is one of the most acknowledged reinforcement learning models.

Let $R(t)$ be the reinforcement strength, which is a monotonic function of the payoff $\pi(t)$. Then, in CR model, the attraction values of strategies and the reinforcement strength after period t are updated under the following rules:

$$p(j,t+1) = p(j,t) + A(\pi(t)) \cdot (1 - p(j,t)) \qquad j = d(t)$$

$$p(k,t+1) = p(k,t) - A(\pi(t)) \cdot (p(k,t)) \qquad k \neq d(t)$$

$$R(\pi(t)) = \alpha^{CR} \cdot \pi(t) + \beta^{CR}$$

CR model modifies BM model in the following two points: (i) In CR model, the attraction of a strategy is defined as a linear function of the payoff by configuring the reinforcement strength as a correlated variable to the payoff, while in BM model, the reinforcement strength factors, αBM and βBM, are fixed and independent of payoff. (ii) BM model treats the nonnegative and negative payoff differently, namely assigns them different reinforcement strength, i.e. αB(to a nonnegative payoff) and βBM (to a negative payoff). Comparatively, in CR model, the reinforcement strength assigned to a nonnegative payoff or a negative payoff is the same. According to the viewpoints of Cross [5], it is meaningless to differentiate between the nonnegative payoff and the negative payoff, since the payoff doesn't obey the von Neumann-Morgenstern axioms. For example, 3 units of positive payoff for a user to choose the jth strategy and 3 units of negative payoff for him/her to choose another strategy influences his/her adopting probabilities of strategy j in next period in a same way, as reflected in CR model, the updating rule of reinforcement strength corresponding to the nonnegative and negative payoff is identical.

3. Experiment and Analysis

3.1. Experimental Research Design

Purposes. The overall purpose of our experiments is to observe how the academic users, who have the experience of using general search engine and are used to input keywords in a single textbox to finish searching (hereinafter referred to as "single textbox search"), learn to use multiple textboxes and Boolean droplists so as to logically combine several keywords to finish searching (hereinafter referred to as "multi-textbox search"), namely to observe how they learn search knowledge and adjust their search strategy from simple search to multi-textbox search to perform complex search tasks. More specifically, the intents of our experiments are:

- to reveal the behavioral characteristics of the academic users with different backgrounds during their adaptive learning of search knowledge.
- to observe the belief adjustment process of academic users during their adaptive learning of search knowledge.
- to discover what kinds of behavior adjustment rules govern the process of academic users' adaptive learning, in other word, to find which reinforcement learning model can better fits users' learning behavior in information searching.
- to verify whether the users with different academic backgrounds but sharing the same search preference show different explicit behavioral characteristics and implicit psychological mechanism in the process of search knowledge learning and search strategy adjusting.

Framework of Data Analysis. Analysis to the experiment data involves the following three levels:

- Explicit level. Analysis at this level is to mine users' explicit behavioral characteristics in the process of learning in which users choose and formulate a strategy, conduct trial by error, and finally obtain the information desired.
- Implicit level. Analysis at this level is to mine users' implicit psychological characteristics that control the explicit behaviors. These implicit characteristics form the reasons for users' behavior during their information searching, namely explain why a user chooses a strategy, why a user adjusts his/her previous strategy, and what kinds of psychological changes and correspondingly what kinds of behavior adjustments result in the success or failure of task performing. Two variables depict the implicit psychological characteristics, i.e. expectation and satisfaction. Expectation means the expected probability of success before a user adopts a strategy, while satisfaction is an evaluation that a user gives to the results obtained.
- Rule level. This level is a nexus between the above two levels. Investigating at this level is to discover the quantitative rules of user behavior adjustment which explain how the implicit psychological effects control the explicit behavioral presentation.

3.2. Behavioral Characteristics

We use four indicators, i.e. *initial strategy*, *stabilized strategy*, *stabilized turn*, *search time cost* to characterize the explicit reinforcement learning behavior of academic users during their information searching.

The *initial strategy* is the product (search function) which a user chooses as the strategy in the first turn of search task. To a great extent, this indicator is determined by users' habitual preferences. Analysis to this indicator can help us gain the insight of behavioral disposition and knowledge background of the users before their search tasks performing.

We introduce stability to describe a user who chooses a certain strategy in one turn of search task and never changes the strategy in the succeeding turns. And the

chosen and never changed strategy is referred to as *stabilized strategy*. This indicator can be examined to discover which search strategy is the most accepted by the users after learning and experiencing. Another related indicator, *stabilized turn*, is a number represents that after how many turns, a user never changes his/her strategy. To some extent, this indicator reflects the learning time cost or learning efficiency of a user from his/her original preferred strategy to the stabilized strategy.

The *search time cost* is the total time for a user to finish all the 10 search tasks. Also this indicator can be used to measure users' learning speed.

Table 1 and Fig 1 is comparisons of the above four indicators between freshmen and senior students. Table 2 is the results of statistical test to the search time cost in which the Wilcoxon rank sum test indicates that the learning time cost of freshmen and that of senior students are significant different with a power of 99.99%.

Table 1. The statistics of academic users' strategy selection

	Freshman	***Senior Students***
Percent of users with initial strategy as single textbox search	95%	82.4%
Percent of users with initial strategy as single textbox search	69%	91.2%
Average stabilized turn	5.61	3.94

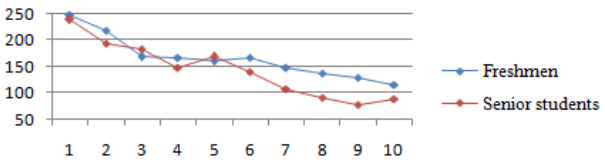


Figure 1. Search time cost comparison between freshmen and senior students

Several conclusions can be drawn from the above statistical data:

- The academic users' learning to use the search products are influenced greatly by their prior preference. For both the freshmen and the senior students, the simple search preference formed in their earlier Internet surfing experience affected them so deeply that when facing the new search environment, most users in deliberately employed the single textbox search strategy as they have done before. Although the percent of senior students who followed their earlier preference in the first turn is less than that of freshmen (see Table 1), it is not dispensable for senior students to learn academic search skills, despite the fact that they have already studied in the universities for almost four years.
- Driven by the search tasks, the majority of academic users could finally choose the most appropriate search strategy. This conforms to the contingency theory of preference [6][7]. For an individual, when placed in different task contexts, his/her preference is not fixed. Instead, this preference is constructed dynamically. When the task context or the behavioral utility changes, the preference changes consequently [8]. In our research, this behavioral utility means to what degree the results fed back by the database system can satisfy the requirement of a user after adopting a certain search strategy. That is to say, when a preferred strategy always leads to failures, the individual will naturally give up this preference and learn to seek another solution.
- Reinforcement learning is an adaptive process of trial and error, a process that users have to take time for. The statistics of indicator stabilized turn in Table 1 tell that after about 4 to 6 turns, users gave up their initially preferred strategy and learn to use multi-textbox search strategy. Obviously, the average learning time cost of freshmen is larger than that of senior students. From this point, it is important for the universities to impart the newly entered students knowledge and skills of academic literature searching. Besides, based on the statistics, it is still necessary to further train the senior students to more effectively retrieve academic literatures.

3.3. User Behavior Adjustment Rules

In this section, we will apply BM model, BS model and CR model to fit the experimental data and mine the rules by which the users adjust their behavior. We used the first 70% turns of the experimental data for model inference (parameter estimation), and the remaining 30% for model verification and validation.

Estimation of Model Parameters. $\boldsymbol{\alpha BM}$ and $\boldsymbol{\beta BM}$ are the parameters to be estimated for BM model, while $\boldsymbol{\beta BS}$ for BS model, and $\boldsymbol{\alpha CR}$ and $\boldsymbol{\beta CR}$ for CR model. We applied maximum likelihood estimation for model inference to determine these parameters. The likelihood function is defined as:

$$LL\ \Theta = Pnk\ t\ Tt{=}1\ Nn{=}1\ . \quad (9)$$

where Θ denotes the parameters, T he total task turns, N the total participants, and $Pnk\ t$ denotes the attraction of strategy k adopted by the nth user in period t.

The results of parameter estimation are shown in Table 2.

Table 2. Results of parameter estimation

	BM model	BS model	CR model
Freshmen	αBM=0.1; βBM=1	β^{BS}=0.28	α^{CR}=0.1; β^{CR}=0.1
Senior students	αBM=0.50846; βBM=0.775	β^{BS}=0.5	α^{CR}=0.1; β^{CR}=0.1
Freshmen	αBM=0.1; βBM=1	β^{BS}=0.28	α^{CR}=0.1; β^{CR}=0.1

In BM model, the parameters αBM and βBM reflect

978-1-935068-76-1 © 2011 SciRes.

the mechanism by which the searching strategy adopted in a certain turn updates the attraction of this strategy for a user in the subsequent turns, in which αBM means the weight assigned by the user to the positive payoff obtained in the latest period. For freshmen, αBM=0.1, while for senior students αBM=0.50846. This means comparatively, the freshmen concern more on the results obtained in all the finished periods (not only the latest one), while the senior students are more subjected to the outcome of the previous period. Besides, freshmen are more sensitive to previous failures than senior students. This also verifies why the correlation between confidence and satisfaction of freshmen is stronger than that of senior students as discussed in section 3.3.

In BS model, the parameter βBS takes the roles of αBM and βBM in BM model. βBS=0.28 for freshmen users means that they assign a weight of 0.28 to the current search results, while 0.72 to all the previous experience. Contrastively, for senior students, βBS=0.5. More intuitively, freshmen are more strongly disposed to act according to the past experience, and for them, overcoming the attraction of prior strategies and switching to a new one may take a long time, even though the efficiency of the habitual behavior is low.

The parameters αCR and βCR in CR model for the two groups of users all equal 1.

That means the influence of the search result fed back in the previous period on the strategy selection behavior in the subsequent periods is not so remarkable.

Model Verification and Validation. Replacing the update rules of strategy attraction with the estimated parameter values, we can obtain the models. Using the learned models, we can compute the strategy choosing probability for the remaining 30% experimental turns and predict the strategies the users may choose and adopt in those turns. Then, by computing the difference between the mean of the simulated strategies and that of the actual strategies as well as the difference between the standard derivation of the simulated strategies and that of the actual strategies for each model and each user group, we can verify the validity of those models in explaining the reinforcement learning behavior. The smaller the differences, the better the model.

Table 3 gives the results of model verification and validation, which indicate that BS model can best predict the learning behavior of freshmen users, while CR model best predicts the behavior of senior students. On the whole, taking all the users into consideration, CR model may be the best.

Table 3. Results of model verification and validation

	Differences	BM model	BS model	CR model
Freshmen	Mean	0.040004	0.0039	0.005919
	Standard derivation	0.002137662	0.004219317	0.11219055
Senior students	Mean	0.010969	0.024872	0.00782
	Standard derivation	0.000420205	0.100978	0.0020750

From the research of this section, we can conclude:

- As shown in Table 6, the differences between the mean and standard derivation of the simulated strategies and those of the actual strategies are all very small, which means that academic users manifest significant behavioral characteristics of reinforcement learning during their learning of information search. In other word, users can adjust their search strategy selection behavior according to the search results in an independent learning environment.
- BS model can best fit the learning behavior of freshmen users, which implies that freshmen always adjust their behavior based on their prior knowledge and experience, their habitual preference halters their later behavior and therefore they may need more time to learn new search knowledge. This consists with the conclusions in section 3.2 and 3.3. As novices, they have little academic background or poor academic literature retrieval skills, so they show a strong viscosity and inertia to the preference of single textbox search formed in their earlier Internet surfing. Besides, BS model also implies that freshmen always adjust their search strategy according to the agreement between their evaluation to the feedback and their expectation, as the analysis in section 3.3. A serious divergence between the confidence and satisfaction certainly leads to the strategy adjustment.
- CR model best explains the learning behavior of senior students, which suggests that not the search result fed back in the previous period but the accumulated utilities from all the finished periods influence the senior students' strategy selection behavior in the subsequent periods.

References

[1] Drew Fudenberg, David K.Levine: The Theory of Learning in Games. Cambridge: The MIT Press, MA, USA(1998)
[2] Thomas Brenner: Agent Learning Representation: Advice in Modeling Economic Learning. Tesfatsion L., Judd K. L. (ed.): Handbook of Computational Economics, Elsevier, vol. 2, pp. 895-947(2006).
[3] Robert R. Bush, Frederick Mosteller: A stochastic model with applications to learning. The Annals of Mathematical Statistics, vol. 24, no.4, pp. 559--585(1953)
[4] Borgers T., Sarin, R.: Naive Reinforcement Learning with Endogenous Aspirations. International Economic Review, vol. 41, no.4, pp. 921--950(2000)
[5] Cross John G A.: Stochastic Learning Model of Economic Behavior. Quarterly Journal of Economics, vol,87, pp. 239~266(1973)
[6] Yanwu Cui, Qin Su, Zhao Li: Consumptive Preference in E-commerce. Soft Science, vol. 21, no.6, pp.19-23(2007).
[7] Yuanlai Xiong: A Comment on Preference Reversal and Its

Theoretical Explanation. Journal of Zhejiang Shuren University, vol. 8, no. 5, pp.68--72(2008).

[8] Lihua Zhu: Survey of Preference Logic Study. Journal Of Tianjin University of Commerce, vol.28, no.6, pp.27-31(2008)

978-1-935068-76-1

The Humanity Hypothesis and Its Validation of Staff in Academic Institutions during the Construction of Institutional Repository

Daling LI[1,2], Yanhong YU[3], Yan WANG[3]

[1]*Center for Resource Sharing Promotion Institute of Scientific and Technical Information of China, Beijing, China*

[2]*Business College Tianjin University of Commerce, Tianjin, China*

[3]*Library Tianjin University of Commerce Tianjin, China*

Email: lidl@istic.ac.cn

Abstract: The definitude of the humanity hypothesis of academic institutions is the prerequisites to propose proper motivation measures. Based on the review on four often discussed humanity hypothesis and Maslow's Hierarchy of Need, Questionnaire analysis explores that the need of staff in academic institutions does accord with Maslow's Hierarchy of Need. The Complex Man Hypothesis of staff in academic institution is proved to be correct by theoretic analysis and finding of questionnaire.

Keywords: Institutional Repositories; open access; humanity hypothesis; Complex Man Hypothesis

1 Introduction

With the application of Institutional Repository we find that different institutes have various utility effects, for example some scholars are reluctant to upload their academic output if not forced. Theoretically, scientific analysis of stable behavior characteristics of staff in academic institute is very important for the success of Institutional Repository. Humanity hypothesis is the important theory of Modern Behavioral Science and Management Psychology on stable behavior characteristics of human. In order to learn the nature of researcher' behaviors to stimulate them to use institutional repository, a survey is implement to learn and verify the Humanity hypothesis of staff in academic institutes.

2 Main Important Humanity Hypothesis Theories

Edgar H. Schein summarized four humanity hypothesis theories include Hypothesis of Economic Man, Hypothesis of Social Man, Hypothesis of Self- Actualizing Man, Hypothesis of Complex man[1].

It is notable that Maslow's "hierarchy of needs theory" are recognized and accepted by many management experts and psychologists. Practice has proved that: the above four humanities theories did not really find a very perfect expression consistent with the theory of human characteristics. Hypothesis of economic man ignored the emotional needs of human which is proved by Hawthorne experiment. Hypothesis of Social Man proposed by Mayo and others is too much emphasis on the role of informal organization and has trend of less emphasis on the formal organization though they realized that the relationship between people for motivational, mobilize the enthusiasm of staff is more important than material rewards, but it is a dependent hypothesis of human nature and lack deep research on human positive initiative and motivation. The hypothesis of self-realization, it holds that people's self-realization is a natural development process, but in fact, human development is the results of social impact, especially social relationship instead of being a natural maturing process, Therefore, many scholars are skeptical of the theory basis. As to the Hypothesis of Complex man, Schein agree that different hypothesis have advantages and disadvantages, and he pointed out that people have different needs and have different abilities, so they will react differently to different management styles.

Consequently, there is not a generally effective management methods which suitable for any age, any organization and any individuals. Hypothesis of Complex man realized the complexity of human, but because of the irregular completely into theory, it is not desirable.

3 Maslow's Hierarchy of Needs Theory and Demands of Academic Staff

3.1 Maslow's Hierarchy of Needs Theory

As the famous American psychologist, Maslow's hierarchy of needs theory has been widely applied to the management area, as the fundamental theoretical basis for incentives, the survey results of this paper also verify Maslow demand theory. The following will give a simple description and analysis of Maslow hierarchy of human needs

According to Maslow view, the needs of people composed of five levels (Figure 1 [2].)

1. Humanities and Social Sciences project of the Ministry of Education of China (09YJC870019)
2. Youth Fund projects of Tianjin University of Commence

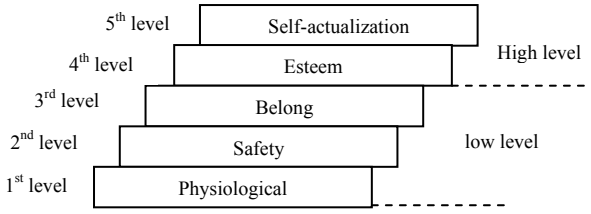


Figure 1. Maslow's hierarchy of needs

To maintain their own physical survival needs including hunger, thirst, clothing, housing, sexual requirements are the most basic requirements of human. If these needs are not met, people's survival became a problem. The people's physical needs is the most powerful driving force in this sense. Only the most basic needs are met to the extent which is necessary to survive, the other needs will become the new incentives.

Security needs is needs people asking for the protection of their own safety, the avoidance of unemployment, the loss of property and the invasion of occupational disease. Human's organs and intelligence and other energy are mainly security seeking tools. Even view of science and life can meet the security needs. Of course, when this need once relatively satisfied, they will no longer be a motivating factor.

Belongs needs also known as emotional needs, include two aspects, the need for love and need for belong, the need for love means that we all need a harmonious relationship between partners, colleagues to maintain friendship and loyalty; everyone hope to love and be loved. The need for belongs is that people want to belonging to a group, to become a member of the group. It is relate to a person's physiological characteristics, experience, education, religion.

Esteem need means that everyone wants to have a stable social status and wish their individual's ability and achievements been recognized by society. The needs can be divided into internal and external respect. In short, the internal self-respect is self-esteem. External one refers to a person's wish to have the rank, prestige and been respected, trusted, highly praised by others. If self-esteem needs are met, people can have confidence in themselves and serve for the community with enthusiasm and will experience their usefulness and value of living.

Self-actualization need is the highest level of need which refers to any needs about the achievement of personal goals, ambitions, the play of their individual capacity to the maximum extent. In other words, people should do what is suitable for them which will bring the happiest feeling.

3.2. The Investigation and Requirements Analysis of Academic Staff

In order to verify the needs hierarchy of academic staff and to explore the effective incentive measures, a survey of "incentive model of institutional repositories based on knowledge management" is pursued.

Considering the characteristics of institutional repositories, and the main participants of institutional repositories are composed of library and relating research institutions, respondents are divided into two levels, the first level includes librarians in university and academic institutions (including academic institutions which have institutional repositories,) and librarians in public library; the second level includes staff in university and academic institutions. The former understand relatively the concept of institutional repositories, especially those have participant in the construction of institutional repositories. The latter as the will-be users will give their requirements of institutional repositories which are critical to the success of institutional repositories.

200 questionnaires were set out to librarian and researcher. (table 1).

Table 1. Table of survey and return statistic data on incentives of institutional repositories

Respondents		Released	Returned	esponse rat	Valid	Valid rate
uni-	librarians	70	40	57.14%	40	57.14%
versity	researchers	130	80	61.54%	71	54.62%
insti-	librarians	30	15	50.00%	15	50.00%
tute	researcher	50	35	70.00%	30	60.00%
Public libraries		10	10	100%	10	100.00%

Factors which influence the staff in academic organization to upload academic output to Institutional Repositories are surveyed. Factors are matched to staff need during the design of questionnaire. Analysis results are shown in table 2.

Table 2. Order list of factors influecing the upload behaviors of staff

order	Factors	Absolutely important (A1)(%)	Strongly Important (A2) (%)	A1+A2(%)	Equally important (A3)	A1+A2+A3	ORDER
1	files Safety	95.18					1
2	Academic status authorized by Organization	55.42					2
3	Support from organization	52.41					3
5	Copy right Assurance	34.34	59.64	93.98			4
10	Research accepted by industry and community	25.90	67.47	93.37			5
4	Available of opportunity	42.77	37.95	80.72			6
8	Required by job evaluation	28.31	48.19	76.50			7
14	Cost of upload	12.65	62.05	74.70			8
13	Convenience of upload	20.48	53.01	73.49			9
7	Research output broadcasting	30.12	40.36	70.48			10
9	Better your job assessment	28.31	40.36	68.67			11
12	More profit than cost when upload	22.89	40.36	63.25			12
11	Economical	24.70	24.10	48.80	31.33	80.13	13

978-1-935068-76-1 © 2011 SciRes.

	incentive						
6	Others' upload behavior	30.72	13.86	44.58	21.08	65.66	14

Questionnaire analysis show: the important order from top to down are Self-actualization need, esteem need, social communication need, performance need, material benefit need (table 3).

Table 3. Need analysis of factors influcing upload behaviors

order	Influcing factors	Need
1	files Safety	Self-actualization need
2	Authorization of academic status by Organization	
3	Support from organization	
4	Copy right Assurance	esteem need
5	Research accepted by industry and community	
6	Available of opportunity	social communication need
7	Research output broadcasting	
8	Required by job evaluation	performance need
9	Cost of upload	
10	Better your job assessment	
11	Convenience of upload	material benefit need
12	More profit than cost when upload	
13	Economical incentive	
14	Others' upload behavior	

The upper findings show that Maslow's hierarchy of needs is suitable for the need of staff in academic organization.

4. Hypotheses of Human Nature of Staff in Academic Organization

In order to verify the hypothesis of human nature of staff in academic organization, the questionnaires based on the hypotheses of human released question about it, the survey question is "which hypotheses do you agree about hypotheses of human nature?" the choices are shown in table 4.

Table 4. Suvey results about hypothesis of human nature

hypothesis of human nature	Agree	disagree
Hypothesis of Economic Man	133	33
Hypothesis of Social Man	95	71
Hypothesis of Complex Man	152	14
Hypothesis of Self-realization Man	133	33

The survey found that most respondents choose more than one option mostly, from 2 to 4 options. Data in table 4 is the cumulative of options ,that is, when a researcher choose options of Hypothesis of Economic Man, Hypothesis of Social Man, Hypothesis of Complex Man, every agree option add 1, if they choose "disagree" the disagree option add 1.

The above phenomena of multi-choice shows that researcher in academic organization think that their self-requirements are multifaceted, multi-level. The nature of human is not simply economic, social or self-realization, but changed according to different circumstances. It is right one symbol of "complex human nature.

At the same time, from the perspective of the characteristics of academic institutions, we can see that staffs in academic institutions are different from the human in economy. Staffs in universities and institutes in China are generally belongs to institution system, their work are more stable, the change of economy has less influence on them. The work and salary can meet their survival need. With the promotion of position and the improvement of academic status, the economic income of researcher increased gradually. All above proved that the physical needs and security needs are met.

Researcher all over the world have generally received highly education, they are respected by people from all class and have higher social status which raises higher need level. Therefore, their aims to engaged in academic research is more than to gain economic benefits, their egoism motive for academic research will bring value and wealth for the universities and community.

At the same time, we must accept the reality than universities are no longer a paradise far away from the competition. We should not ignore the facts that the ranking of universities and the difference of research which makes more heat competence of research organizations for research fund and projects. All these requires researchers to improve academic levels and academic influence. A cooperative competition relationship appeared among researcher in academic organization. As all know than scientific research can't always be completed by one researcher, the great scientific achievement is a team work standing on the shoulders of giants. This brings the ownership needs of academic researchers to organization and research team. To those professors, their research goal to scientific research is to achieve their life and career value.

We found from the survey that after getting appropriate economic return, staffs in academic institutions have various goals when they engage in research. For young researchers they need more economic benefits to improve their living conditions, to support their further education and to improve their own research because they need financial support to participate in academic conference and to engage in research. They expect that their research output will provide opportunities when academic title promoted. Then they wish to enhance their academic status and realize self-worth and personal ambition. To those professors, the reason why they engage in research are not maximize the economic benefits but benefit the society, achieve self-worth and get pleasures during the research.

In addition, the hypothesis of complex man theory inherited the all the advantages of all other hypothesis of human nature though it denies the existing of universal management model and emphasis on flexible and contingent incentive. Moreover, the deny of universal man-

agement model does not mean the hypothesis of complex human nature will oppose the implementation of certain management model on certain community.

5 Conclusion

Survey results show that the need architecture theory of hypothesis of self-realization man meet the demand architecture of researchers in academic institutions, the social relationship need consistent with the hypothesis of social man theory, the diversity and motivation in demand meet the hypothesis of complex man theory. We can make conclusion that only hypothesis of complex man can explain the survey result about the hypothesis of human nature of staff in academic organization in this paper. We must take this hypothesis into account when we build institutional repository and design proper incentive method to encourage the researchers to use institutional repository and to guide the researcher to explicit their implicit knowledge and make their knowledge open access. All these will promote the transformation of knowledge from private to organization and improve the efficiency of academic exchanges. As a result improve the academic status of researcher and universities and institutes and make institutional repositories run more efficiently.

References

[1] Dong Ke yong., Ye xiangfeng. Introduction to Human Resource Management. Beijing: China Renmin University Press. 2006: 104-107

[2] http://www.ruralhealth.utas.edu.au/comm-lead/leadership/maslow-diagram.htm

978-1-935068-76-1

Application of Autocorrelation on Requests for Full-text of NSTL

Changqing YAO[1], Chunyun HAO[1], Lianjun ZHANG[2]
[1]*Institute of Science & Technical Information of China*
[2]*Key Laboratory of 3D Information of Capital Normal University Beijing, China*

Abstract: In this paper, studied the request information of full-text of NSTL network service system, applied correlation analysis method, analyzed the provinces' requests and judged the correlation exist or not, and support the help for making the tactics of purchasing resource and serving of NSTL in the future.

Keywords: Request of full-text; National Science and Technology Library; Correlation Analysis

1 Forward

NSTL is a virtual science/technology documentary information service organization that was approved by State Council to be founded on June 12 of 2000, consisting members of Documentation and Information Center of Chinese Academy of Sciences, National Engineering Technology Library (ISTIC, Mechanical Industry Research Institute of Information Industry, China Metallurgical Information & Standardization Institute, CNCIC), Library of the Chinese Academy of Agricultural Science, Library of Chinese Academy of Medical Sciences and etc. Now it is the largest virtual science/technology information service provider in China and its centric network service system has evolved to be one of largest science/ technology documentary information resource service websites. Such system is specially designed to serve all Chinese users working in science and technology fields. NSTL has set many service columns while viewed Document Retrieval and Full-text Service as its earliest and most important service item. [1]

SPSS is a statistical analysis tool that is well known in investigation statistics, marketing research, medical statistics, government-enterprise-level data analysis application and also the earliest statistical analysis software in the world, which was successfully developed by three graduates of Stanford University (USA) at the end of 1960s. It is mainly used in various fields of natural science, technology science and social science and now is mostly used by the world as special statistical software. This article, viewing full-text request information of NSTL network service system as the research target, selects documents in Chinese journals to undertake various correlation analysis activities by SPSS on the basis of 22 key categories [2] of books of Chinese Library Classification. We aim to gain correlation level of all categories and analyze research results for the purpose of providing decision support for future resource acquisition strategy and service strategy implemented by NSTL[3].

2 Data Source and Processing

2.1 Data Source

Data used in this research mainly includes three categories of Log DB of Ordering Information and Journal List.

(1) Log DB of Ordering Information. This data is used to record information of document ordered by users registered in this website, including user name, Full-text request ID, request time, delivery mode of document requested by user, collection of library where document is requested, and etc. Journal's title, name and author are also included. Initial DB is stored in TRIP DB and will be outputted as text document in particular format.

(2) Journal List includes name and type of Chinese journals. Journal's type is classified by Chinese Library Classification and includes 22 key categories, which will be replaced by different capital letters. Mapping table for above 22 key categories is shown as follows.

Table 1. 22 Classification Table

A	Marxism and Leninism, Mao Zedong Thought, Deng Xiaoping Theory	G	Culture, Science, Education, Sports	O	Mathematics, Physics and Chemistry	U	Transportation
B	Philosophy, Religion	H	Language, Letters	P	Astronomy, Geo-science	V	Aviation, Space Flight
C	Social Science -- General	I	Literature	Q	Bioscience	X	Environment Science, Safety Science
D	Politics, Laws	J	Arts	R	Medicine, Sanitation	Z	Comprehensive Books
E	Military	K	History, Geography	S	Agriculture Science		
F	Economy	N	Natural Science -- General	T	Industrial Technology		

2.2 See below Figure for Data Process Flow

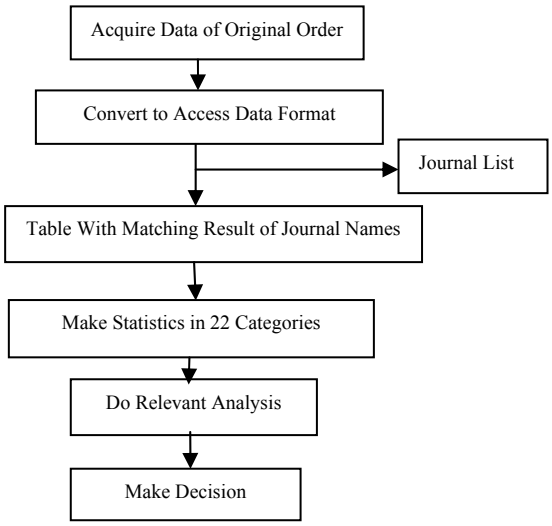


Figure 1. Data process flow chart

Date	A	B	C	D	E	F	G	H	I	J	K
2003	0	25	39	18	2	993	319	1	0	0	1
2004	0	33	32	28	13	987	453	1	0	0	0
2005	0	50	91	103	78	664	489	2	0	10	2
2006	0	17	26	106	350	360	285	0	0	0	0
2007		237	1140	392	337	3050	1264	20	1	3	15
N	**O**	**P**	**Q**	**R**	**S**	**T**	**U**	**V**	**X**	**Z**	**Total**
4208	4386	368	3132	18317	3185	33947	244	88	517	0	69790
3843	4918	309	3831	20421	3083	27091	235	93	418	2	65791
4243	6447	848	4490	22532	4647	31114	467	207	722	0	77206
3502	7981	1012	5382	23500	2917	30079	854	431	687	0	77489
6296	16926	2203	12221	65157	5353	50233	1099	807	1382	11	168147

Figure 2. List of number of 22 full-text requests

3 Analysis Method

Relevant analysis is designed to reveal correlation level of all elements of research target under study and test of such correlation level will be completed by calculation and inspection of relevant system. This article will use relevant analysis method to analyze all categories for gaining their correlation level [3-5].

3.1 Correlation of Two Sequences

Suppose there are two sequences:

X: $x_1, x_2, \ldots, x_n$ Y: $y_1, y_2, \ldots, y_n$

Then according to statistics theory, correlation coefficient between such two sequences will be:

$$r = \frac{\sigma_{xy}^2}{\sigma_x \sigma_y}$$

In which,

$$\sigma_{xy}^2 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n}$$

$$\sigma_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} \qquad \sigma_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n}}$$

$\bar{x}$ and $\bar{y}$ are respectively average values of such two sequences, n is the sample number of such two sequences.

As fundamental formula of correlation coefficient is difficult to be calculated, we may get below computation expression from such formula.

$$r = \frac{\sum x_i y_i - \frac{\sum x_i \sum y_i}{n}}{\sqrt{\left(\sum x_i^2 - \frac{(\sum x_i)^2}{n}\right)\left(\sum y_i^2 - \frac{(\sum y_i)^2}{n}\right)}}$$

Value range of correlation coefficient (r) is $1 \geq r \geq -1$ and its values have the following meanings:

r=1, Sequences of X and Y are linear positive correlation;

r=-1, Sequences of X and Y are linear negative correlation;

r=0, Sequences of X and Y have not any correlation;

1>r>0, Sequences of X and Y are positive correlation and correlation level will be more extensive along higher of r value;

0>r>-1, Sequences of X and Y are negative correlation and correlation level will be higher if r is closer to -1.

3.2 Relationship Correlation Coefficient (r) and Sample Number (n)

Label correlation coefficient calculated by n samples as r_n. In theory, r_∞ gained will not really represent correlation level of two sequences only if sample number (n) tends to infinity. When sample number is finite, value of r_n will oscillate hereabout and deviation level of r_n and r_∞ will be smaller if n is bigger. As analysis formula of r_n distribution derived from arithmetic formula is very complex, we will use Monte Carlo Method to study distribution rule of r_n (two sequences will be used to study distribution of their correlation coefficients for the purpose of simplification).

Two random sequences without any relation is then generated by computer (they will be labeled as one group).

X: $x_1, x_2, \ldots, x_n$

Y: $y_1, y_2, \ldots, y_n$

As such two sequences have no relation, their correlation coefficient (r_n) will be *0* in theory. However, when sample number (n) is an finite value, r_n will be not 0 surely but a random value around 0. M groups of random sequences will be generated by computer to gain M r_n and its distribution may be shown in Fig. 1 (M=50000):

In Fig. 3, when n=100, probability of $|r_n| < 0.1655$ will be 90%, probability of $|r_n| < 0.1964$ will be 95% and $\sigma^2 = 0.0101$;

978-1-935068-76-1 © 2011 SciRes.

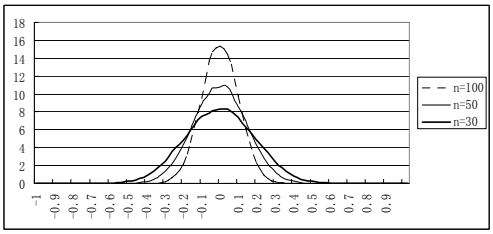


Figure 3. Distribution of Correlation Coefficient of 50000 Groups of Irrelevant Random Sequences

When n=50, probability of $|r_n|$ <0.2362 will be 90%, probability of $|r_n|$ <0.2811 will be 95% and σ^2 =0.0206;

When n=30, probability of $|r_n|$ <0.3064 will be 90%, probability of $|r_n|$ <0.3608 will be 95% and σ^2 =0.0347;

From above analysis result, we may conclude that distribution of r_n is basically positive trend and its variance (σ^2) is inversely proportional to n and it's about equal to $1/n$. therefore we may conclude that we could not easily get conclusion that two sequences have relation when we know correlation coefficient of such two sequences is not 0, and vice versa when correlation coefficient of such two sequences is 0, we can not get that two sequences have no relation. Such two result just help us get probability of relation of no relation between such two sequences and such certainty will be higher in case n is bigger.

4 Application

We get Table 2 from analysis on correlation of full-text request data by above method. We get the following conclusion upon rules of correlation in a result:

|r|>0.95 has significant correlation;

|r|≥0.8 high-level correlation;

0.5≤ |r|<0.8 medium-level correlation;

0.3≤ |r|<0.5 low-level correlation;

|r|<0.3 such relation is very poor and suggest there is no correlation.

Table 2 lists data result processed by SPSS and it has no correlation as number of A journals is O. As most of correlation systems in this Table are positive numbers and almost above 0.8, it means there is high-level correlation. We can get from this Table that correlation coefficient of B and C journals is 0.996 and this means significant correlation between B and C. There is very significant positive correlation between B journals and other journals of D, F, G, H, K, O, T, X and Z level. However correlation coefficient of B and E journals is just 0.523 and this means there is medium-level correlation between them. Correlation coefficient of E and G journals is 0.473, Correlation coefficient of E and N journals is 0.383, Correlation coefficient of E and S journals is 0.317 and this means there are low or medium-level correlations between E and most journals. Correlation coefficient of J and B journals is 0.168, Correlation coefficient of J and other journals of C, D, E and F is less than 0.2 and this means there are very small correlation coefficient between J and other journals. They have poor relations and we may get there is no correlation. E represents military journals and J represents art journals.

Table 2. List of Correlation Coefficient of Request Number of 22 Journals

	A	B	C	D	E	F	G	H	I	J	K	N	O	P	Q	R	S	T	U	V	X	Z
A	.(a)	.(a)	.(a)	.(a)	.(a)	.(a)	.(a)	.(a)	.(a)	.(a)	.(a)	.(a)	.(a)	.(a)	.(a)	.(a)	.(a)	.(a)	.(a)	.(a)	.(a)	.(a)
B	.(a)	1	.996	.961	.523	.968	.994	.998	.991	.168	.996	.978	.947	.909	.962	.987	.844	.952	.718	.857	.945	.977
C	.(a)	.996	1	.968	.570	.967	.983	.999	.999	.103	.997	.970	.961	.923	.973	.995	.806	.965	.751	.884	.952	.979
D	.(a)	.961	.968	1	.725	.872	.943	.959	.963	.198	.964	.903	.995	.988	.993	.984	.820	.923	.877	.959	.991	.923
E	.(a)	.523	.570	.725	1	.375	.473	.535	.583	-.079	.539	.383	.769	.807	.734	.645	.317	.536	.964	.887	.717	.514
F	.(a)	.968	.967	.872	.375	1	.964	.974	.969	.003	.963	.972	.865	.795	.891	.941	.740	.937	.573	.750	.848	.976
G	.(a)	.994	.983	.943	.473	.964	1	.988	.977	.197	.982	.968	.925	.881	.944	.973	.853	.917	.674	.822	.918	.978
H	.(a)	.998	.999	.959	.535	.974	.988	1	.997	.119	.997	.978	.949	.909	.964	.990	.818	.963	.724	.864	.944	.980
I	.(a)	.991	.999	.963	.583	.969	.977	.997	1	.052	.992	.961	.962	.918	.974	.995	.774	.963	.756	.889	.944	.983
J	.(a)	.168	.103	.198	-.079	.003	.197	.119	.052	1	.163	.205	.108	.209	.094	.090	.666	.085	.065	.050	.257	-.010
K	.(a)	.996	.997	.964	.539	.963	.982	.997	.992	.163	1	.983	.950	.919	.961	.986	.842	.974	.733	.867	.957	.963
N	.(a)	.978	.970	.903	.383	.972	.968	.978	.961	.205	.983	1	.877	.841	.894	.941	.861	.966	.604	.764	.904	.935
O	.(a)	.947	.961	.995	.769	.865	.925	.949	.962	.108	.950	.877	1	.987	.998	.983	.758	.913	.902	.977	.980	.925
P	.(a)	.909	.923	.988	.807	.795	.881	.909	.918	.209	.919	.841	.987	1	.976	.950	.784	.891	.934	.982	.989	.859
Q	.(a)	.962	.973	.993	.734	.891	.944	.964	.974	.094	.961	.894	.998	.976	1	.991	.763	.918	.875	.963	.973	.947
R	.(a)	.987	.995	.984	.645	.941	.973	.990	.995	.090	.986	.941	.983	.950	.991	1	.787	.948	.809	.924	.964	.974
S	.(a)	.844	.806	.820	.317	.740	.853	.818	.774	.666	.842	.861	.758	.784	.763	.787	1	.778	.549	.653	.845	.725
T	.(a)	.952	.965	.923	.536	.937	.917	.963	.963	.085	.974	.966	.913	.891	.918	.948	.778	1	.719	.846	.937	.908
U	.(a)	.718	.751	.877	.964	.573	.674	.724	.756	.065	.733	.604	.902	.934	.875	.809	.549	.719	1	.972	.876	.685
V	.(a)	.857	.884	.959	.887	.750	.822	.864	.889	.050	.867	.764	.977	.982	.963	.924	.653	.846	.972	1	.948	.835
X	.(a)	.945	.952	.991	.717	.848	.918	.944	.944	.257	.957	.904	.980	.989	.973	.964	.845	.937	.876	.948	1	.882
Z	.(a)	.977	.979	.923	.514	.976	.978	.980	.983	-.010	.963	.935	.925	.859	.947	.974	.725	.908	.685	.835	.882	1

5 Conclusion and Perspectives

By above study, we can get significance of correlation analysis on NSTL full-text request by statistical analysis method. Firstly data process may help decision-making level understand situation of journals in above 22 categories in recent years and their development trend. We may get change of request number of such journals in 22 categories.

Secondly, we may guide to reasonably collection and assign resources of information documents, use this research method to get number of request of such journals in 22 categories by NSTL users and relationship between categories and scale relation between categories and identify industrial structure of documents. This study may lead reasonable guiding role to collection of documentary resources and favor further optimization of configuration of resources.

However we still find some problems in this study, for example, we just give correlation analysis for request data of original-text journals. In fact we may get better result if selecting more methods for further analysis, such as regression analysis, cluster analysis and etc.

References

[1] Hao Chunyun, Several Research Methods on NSTL Users' Full-text Request Behavior. Researches in Library Science [J]. 2007, (1): 79-80.

[2] Chinese Library Classification Committee. Chinese Library Classification. Beijing: National Library of China Publishing House.1999.

[3] Zhu Jianping and others. Application of SPSS in Statistical Analysis. Tsinghua University Press. 2008

[4] Luo Jiyu, Xing Ying. Analysis Method and Forecast of Economic Statistics. Beijing: Tsinghua University Press. 1987

[5] Zhang Chongfu, Chen Shuyun, Hu Xiling. Statistical Analysis Method and Its Application. Chongqing: Chongqing University Press.1995.

Construction and Assessment of Governmental Decision-Making Oriented KMS

Hongliang HU

Institute of Scientific and Technical Information of China (ISTIC), Beijing, 100038

Information Management School Wuhan University, 430072

Abstract: This paper takes the transformation of traditional information service in science and technological information industry into governmental decision-making (especially strategic decision-making) oriented knowledge service and management system as an example, originally puts forward the content requirement and functional requirement of governmental decision-making knowledge management system (hereinafter referred to as KMS), probes into the assessment methods for the system so as to promote the application of KMS in governmental decision-making.

Keywords: Decision-making oriented; KMS; construction; assessment

In order to meet emerging needs and provide governmental strategic decision-making oriented knowledge service, an effective approach is to build a governmental decision-making oriented knowledge service system. Taking technological information industry as an example and decision support system construction at strategic level as a basis, this paper probes into methods for construction and evaluation of said system.

1 Concept of Governmental Decision-making Oriented KMS

Scientific decision-making, especially at the strategic level, is central to governmental management and service. From this point of view, governmental decision-making oriented KMS is an important part of general e-government system. However, at present, our country's e-government system construction pays much direct attention to traditional administrative affairs and operation system management such as service and business process, official document management, office automation and so on, wherein government functionary instead of decision-making advisors is the largest user of the e-government system. From this perspective, governmental decision-making support oriented KMS differs a lot from e-government system in a narrow sense.

Governmental decision-making support oriented KMS is defined as a knowledge and service-providing oriented management system frequently used by government policy makers or decision-making advisors for putting forward decision-making references for policies and regulations or macroscopic strategic formulation, which is a significant supplement to e-government business system as well as an important part of e-government.

2 Users of Governmental Decision-making Oriented KMS

Users of governmental decision-making oriented KMS not only include national civil servants but also a large group of government decision-making advisory staff. Meanwhile, it would be better for the system to be open to the public given that government decision-making puts more and more emphasis on listening to public opinions. A report from the Institute of Scientific and Technical Information of China (ISTIC) shows that the largest users of this system are researchers from governmental decision-making support departments, and then are civil servants and the public cared for the tendency of national science and technology. These three types of users make up of 85% in the total amount, forming a spindle-shaped distributed architecture as below:

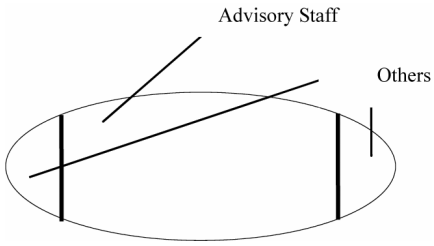


Figure 1. A Apindle-shaped Distributed Architecture

3 Content Requirement for Governmental Decision-making Oriented KMS

When government formulating national macroscopic strategic policy, what is most needed is the macroscopic, holistic, authoritative and perspective information analysis. Xianzhong NIU from Taiwan of China points out in his book "Strategic Research" that, as far as a strategist is concerned, knowledge required for strategy comprises three parts: the first is about all static knowledge of a country, reflecting existing status of the country; the second is about dynamic knowledge of a country, reflecting current capability of the country; the third is about knowledge on potential tendency of a country, re-

978-1-935068-76-1

flecting intention of the country[4]. Upon our many years' experience in providing decision-making support service for Ministry of Science and Technology, generally speaking, the knowledge required for micropolicy and strategic policy formulation can be classified as follows:

1) Basic scientific and technological materials mainly comprising assembly of policy and regulations, collection of control regulations, official documents, basic statistic data, science and technology management document and so on.

Specifically, these materials primarily include policies, laws and regulations instituted by the state for scientific and technological management, various documents, research findings, investigation and survey report interspersed among ministries and commissions and departments at local levels, as well as national basic statistic data and national bibliographicfiles. It is just because the higher requirement for accuracy and authenticity that these materials are taken as fundamental knowledge for scientific and technological decision-making.

2) Dynamic scientific and technological information characterized in rapid updating

These kinds of information, in particular, the dynamic information on scientific and technological progress at home and abroad and abrupt, prewarning, strategic and critical information in the field of scientific and technological decision-making are of great reference value for forward-looking governmental decision-making. Featured in rapid updating, suddenness, alteration, the information mainly comes from news channel and interactive forum system. For example, ISTIC has been following the tracks of foreign scientific and technological progress over the years and providing references for related national decision-making departments through special magazines like "Scientific and Technological Reference"; Chinainfo (www.chinainfo.gov.cn) also sets up special section "Foreign Headlines" to render a service for scientific and technological decision-making, which is very popular with users[5].

3) Subject assembly, argument library or case library directed at significant decision problem

An important characteristic of KMS is to provide solutions, which makes it different from traditional information service. Here is a vivid metaphor: traditional information consultancy service can be compared to "get medicine" while KMS "write a prescription"; in other words, KMS needs to provide solutions or argument support close to solutions.

In the US, the Congressional Research Service (CRS) affiliated to the Library of Congress is a service delivery institute specially established to offer public policy studies for US congress. CRS established a special database for collecting research results achieved by experts from various fields and assorts these data in sequence according to subject.

With regard to major decision problems, it is necessary to extensively collect positive and negative arguments and analyses of domain experts at home and abroad, even related detailed background information as a supplement for forming specific solution set or significant case library, thus providing the most direct support for governmental decision-making.

The aforesaid three types of information constitute the content framework of governmental decision-making oriented KMS, and the most desirable information source for advisory staff of governmental decision-making.

4 Fundamental Function Requirement for Governmental Decision-making Oriented KMS

From the whole constitution of KMS, it can be seen that content requirement and function requirement are inalienable as a whole. However, most system constructions in domestic practice end up with a failure for preferring to function rather than content. Therefore, great efforts should be made to avoid this kind of preference, and meanwhile to design function based on contents during the system construction process.

A basic knowledge process is shown as below:

Figure 2. A basic knowledge process

Upon such a process, we hereby put forward the functional requirements for governmental decision-making oriented KMS:

4.1 An Efficient Information Generation or Knowledge Formation Function

Information lays a foundation for knowledge generation and the generation of knowledge is a source and basic material for KMS. Efficient information generation or knowledge formation is an antecedent and basis for system construction and the function thereof comprises two major parts:

The first is rapid information collection capability. It is essential for system to possess functions like automatic

collection technique, autoabstract, automatic translation and automatic de-emphasis and so on, especially under the environment where network information is increasingly becoming the main source of knowledge.

The second is integration capability in data base or knowledge system. According to "Survey on 2004 Internet Information Source Quantity in China", scientific information database accounts for 7.7% based on the total amount of online databases existed in the nationwide websites, ranking the fifth place[7]. Among decision-making reference information, database concerning laws and regulations, policy documents, scientific and technological literature or statistics has taken shape initially, which should be integrated into or applied to KMS.

4.2 Knowledge Discovery Function

Knowledge discovery is a treatment process for extracting credible, novel, efficient and comprehensible pattern. Studies on knowledge discovery have aroused extensive concern worldwide and its application will have a bright future. Currently, focuses on knowledge discovery research in information resources field include standardization of data mining language, knowledge acquisition and organization and intelligent search under network environment, visualization method for data mining, new approaches for complicated data-typed mining, Web content mining and privacy protection and information security of data mining and so on.

Specifically, knowledge discovery in the field of scientific and technological decision making, is to analyze and process a large amount of scientific information and related data to find novel, efficient, useful and comprehensible pattern and reveal the inherent rules therein, and further to represent these rules by making full use of information technology tools like information analysis, scientific measurement, information extraction, data mining, knowledge network, semantic analysis in response to scientific and technological innovation and management. For instance, the existing scientific and technological literature resources can be utilized for researching and tracking evaluation indicator system for scientific and technological output represented by academic papers and invention patents and the leading indicators thereof from the three layers of country, region and academy to calculate the number of national and domestic scientific and technological papers outputted in our country and figure out the condition of paper citation, indicators like quotient, ranking, academic distribution, regional distribution and institution distribution around the whole world; intercomparison between the number of invention patent application and that of granted patents and the development tendency thereof; contribution of research and development expenditure input to citation frequency of papers and invention patents and the like so as to provide references for scientific and technological decision-making.

Techniques required for knowledge discovery mainly contain data mining, intelligent classification, clustering technology and so on.

4.3 Knowledge Sharing Function

An expert team is always necessary during the formulation process of scientific and technological decision-making. For example, thousands of experts from various fields are put to use to work out "National Planning for Medium and Long-term Development of Science and Technology"; Intelligence agency, more often than not, employs internal and external experts to provide intellectual support for decision-making consultancy. Peter F. Druck had once mentioned that "Knowledge organization can be viewed as a symphony." [8] In this case, knowledge sharing function plays a more important role in KMS in that it not only does good to the construction of a learning-oriented knowledge organization but also boosts the transformation of implicit knowledge into explicit knowledge, thus helping gathering the intellectual achievements of the public.

On the one hand, knowledge sharing should be convenient for users to release knowledge at any time and any place; on the other hand, an open cooperative working atmosphere is needed so as to facilitate the discussion and communication among users.

4.4 Knowledge Transmission Function

How to guarantee a smooth communication of knowledge and information between sectors at higher and lower levels is vital to a governmental decision-making KMS. The channel and efficiency of knowledge transmission plays an important part in the realization of knowledge sharing and the function thereof. Traditionally, reports presented to higher level departments are through daily consultation, official documents, thematic briefing and so on. However, information age requires knowledge to be passed to decision makers quickly through internet, so KMS needs to provide an interface to e-government. For instance, when ISTIC taking charge of constructing "Scientific and Technological Decision-making Foundation Database", an interface to business collaboration (Independent Operation) with Ministry of Science and Technology is set up with respect to conclusive knowledge.

In addition, governmental decision-making also needs the support of the public, particularly the support from beneficiary related to the decision-making, as well as their feedback and suggestion. Consequently, it is also necessary for KMS to have an interface to the public.

5 Model Structure of Governmental Decision-making KMS

From the foregoing, there forms a simple model structure of governmental decision-making KMS as bellow:

978-1-935068-76-1 © 2011 SciRes.

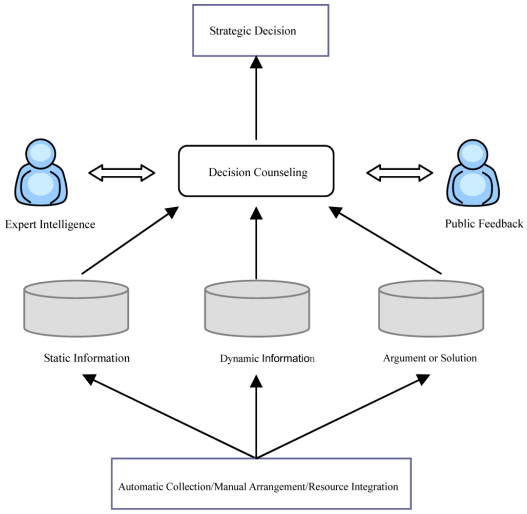


Figure 3. Model structure of governmental decision-making KMS

This model shows that content and function are two indispensable aspects for KMS, which constitutes the basic framework and requirements for KMS by taking content as a main line and function as a link.

6. Assessment of governmental decision-making KMS

For the sake of better understanding and constructing governmental decision-making KMS according to guidelines, it is necessary to discuss evaluation or assessment of this system. Assessment of an information system usually includes many contents, for example, assessment of construction, assessment for use value, assessment of customer satisfaction degree. This papers mainly talks about assessment of application and efficiency of this system and puts forward preliminary criteria based on content and function efficiency, wherein content efficiency lays particular stress on assessment of the degree of contribution made by knowledge content to decision support and function efficiency on assessment of functions in realizing of knowledge management and sharing.

5.1 Content Assessment of KMS

Content assessment mainly stresses assessment of knowledge quantity and knowledge quality possessed by KMS, the significance thereof is reflected on the degree of contribution by the system to decision support. The major criteria comprise:

Richness of knowledge: richness of knowledge relates to extent and depth of knowledge content provided by KMS, wherein knowledge extent requires KMS to cover basic information contents in fields concerned and knowledge depth requires the content quality of KMS to meet the standard of decision support. Specific assessment criteria for knowledge richness consist of gross information content, information value and so on.

Timeliness of knowledge: knowledge or information has evident timeliness, which will become valueless beyond a certain time limit. In an age with information continuously emerging and rapidly updating, close attention has been gradually paid to the information timeliness when people intend to acquire desirable information, among others. The updating rate of knowledge embodies the timeliness thereof. So, assessment of timeliness varies with knowledge content of different classes, for instance, it requires KMS to present latest documents published or promulgated with respect to information on laws and regulations; while as to news information, a higher standard for promptness is necessary.

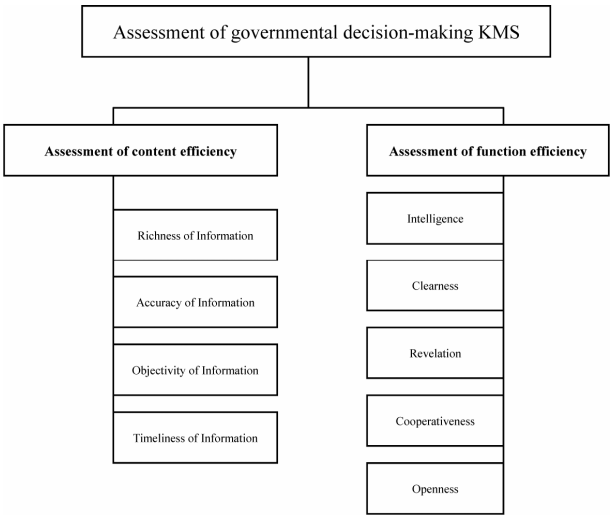


Figure 4. Assessment frame of governmental dicision-making KMS

Credibility of knowledge: it is also named as objectivity of content, indicating the credible degree that knowledge reflects objective things. In terms of governmental decision support, knowledge credibility embodies authority of knowledge. Owing to the openness of internet, any information should have clear source indication; meanwhile, authenticity and objectivity along with authority and popularity of information source indication in this field should also be identified and verified. For example, during the process of constructing "Scientific and Technological Decision-making Foundation Database" for Ministry of Science and Technology, ISTIC had meticulously investigated and researched the source of "Science and technology are the primary productive forces", a famous quotation of comrade Deng Xiaoping, till the accurate original source thereof was determined finally, which provides helpful reference for leaders' correct citation.

Availability of knowledge: it can be interpreted as relativity between knowledge and users. Concept of this criterion should reflect users' needs, that is to say, knowledge that is deemed to be available by users is valuable[9]. In order to ensure that the applied knowledge will meet the requirement of governmental decision-making, close contact with users at the stage of content design is necessary; furthermore, service goal

and object of this content should be defined clearly so as to satisfy the users to the utmost. This criterion is obtainable from utilization frequency of or users' evaluation on information knowledge.

Content organization attribute of knowledge: it comprises correlation and correspondence between information content and knowledge; in particular, some relationship between explicate knowledge and implicit knowledge should be revealed. Organizational attribute of content plays an important part on knowledge demonstration and discovery.

5.2 Function Assessment of KMS

A well-developed KMS, above all, is a traditional information management system. As a result, as far as function is concerned, KMS should comprise fundamental functions possessed by ordinary information systems, such as security, accessibility, stability and others. What's more important, KMS functions in knowledge process and management, for example, intelligence (in fields of intelligent search, intelligent mining, etc.) and openness (an interface between e-government and other systems). Function assessment stresses on investigation of the knowledge service capability of KMS, which can be expressed by the following features:

Intelligence: it mainly probes into KMS's reflection of the outside world including users and resources of information and knowledge, for example, evaluation on functions of intelligent collection, classification and updating of information, discrimination and treatment of valuable information and junk information, intelligent search, self-learning of users and the like.

Clearness: it mainly studies and evaluates function of content organization and demonstration. Primarily, KMS should be rich in information quantity; meanwhile, in aspects of knowledge navigation capability, knowledge search capability and knowledge correlation demonstration capability, KMS should be well developed for the purpose of searching out the most appropriate knowledge in the shortest time. Another technical development tendency is towards friendliness to interface and knowledge visualization.

Revelation: it mainly studies and evaluates knowledge discovery and knowledge exploration capability, comprising discovery of hot information, explicit or implicit relation of knowledge, demonstration of pre-warning information; besides, evacuation, records and sharing of personal interests and preferences should also be encompassed in this criterion for users of KMS usually have certain preference in some fields.

Cooperativity: it mainly studies and evaluates knowledge sharing function and cooperative work capability of knowledge system. As to KMS, sharing capability of knowledge is a very important criterion. The cooperative work of knowledge organization is only realized after knowledge is shared and passed among members.

Openness: it mainly studies and evaluates openness degree of KMS, including compatibility with other systems and expandability of interface.

The above-mentioned assessment model is no more than a preliminary outline and more in-depth criteria and more specific evaluation methods are necessary for constructing governmental decision-making KMS.

The construction of governmental decision-making KMS is a relatively complicated system engineering. In accordance with aforesaid train of thought, the author of this paper has taken charge of governmental decision support oriented information system construction for several times, whereby a few functions have been realized; moreover, scientificity and fast response of decision making are embodied when ISTIC provides decision-making consultancy for Ministry of Science and Technology and other government departments. Partial contents and functions also are used for offering decision support service for department at local levels and the public by means of "Chinainfo" system, achieving satisfactory influence and social benefits. Up to now, KMS has been a more mature system than other similar ones in our country.

In a word, although theoretical researches and practices on knowledge management application are very common, such researches and practices specially directing at governmental strategic decision support still strand at the stage of e-government. This paper aims to provide some thoughts and viewpoints for the development in governmental decision-making KMS on the basis of practical experience and research for many years so as to promote governmental decision-making to be more scientific, democratic and swift.

References

[1] http://www.chinainfo.gov.cn/document/1_30_1.html 2006.5.18
[2] http://www.loc.gov/crsinfo/whatscrs.html#goals 2006.5.10

978-1-935068-76-1

An Interpretation on the Interactive Relationship between Enterprise Information Environment and Knowledge Innovation via Information Ecological Views

Jie MA, Ruoxing ZHENG, Xiaole LIU
School of Management, Department of MIS, Jilin University Changchun, China
Email: mj@jlu.edu.cn

Abstract: Information Ecological Views with human, information and information environment is a newly advanced views regarding the information overflow and chaos in the information society. It consists of the views of system, balance, interaction, humanity and cycle. An enterprise has an information ecosystem within itself and its knowledge innovation has a complex relationship with the enterprise's information environment. This article will give an interpretation on the relationship between the enterprise information environment and its knowledge innovation.

Keywords: information environment; information ecology; enterprise; knowledge innovation

1 Introduction

In the year 1886, German biologist Haecke first put forward the definition of "ecology", saying that ecology is a science which studies organism and its surroundings—including non-biosphere and biosphere's interaction [1]. Human beings should conquer nature, transform nature on the premise of abiding by the law of the natural ecology, or human beings will bear the serious consequences of the ecological imbalance.

The definition "information ecology" was first used by some western scholars in the 1980s and 1990s. It was used to express the relationship between the sense of ecology and increasingly important and complex information environment[2]. To learn to use the concept of environment information to explain problems in the information society so as to make humans creative activities follow the laws and principles of information ecology is a new exploration of people in the new social conditions.

From the information ecology's point of view, each company's interior, the company and its industrial environment, the company and the country that it is located, even the whole globe form a unique information ecological system. The information ecological system which the company is in makes up the information environment of the company innovation and the enterprise relies on the information that received from the system. It is helpful that exploring the relationship between the enterprise information environment and the innovation, which enable the enterprise to discover the innovation rule and model, improve the innovation effect.

2 Information Ecological views

Information ecology is an artificial system with a certain self-adjustment ability which consists of information-human-information environment. In the information ecosystem, the information exchange between information producer, transmitter, decomposers, consumer and the external environment make up a information eco-cycle. Information ecology treat human, information and information environment as a unity. The survival of spirit (information society) needs a corresponding information ecology environment just as the material existence of man demand the nature ecological environment[3]. The information, information people and information environment objectively formed a symbiotic relationship between a chain of needing, providing, updating and feedback. And it also composes a balanced motion. What attracts most attention is the human behavior using technology. Therefore, we can arrive at the five concepts of information ecology, namely, systematic view, balanced view, interactive view, humanistic view and cycle view[4].

2.1 Systematic View

Information ecology is a unitary relationship that involves human, information resources and other important factors and resources of the information environment. There are strong inter-linkages and interdependence between the various components of the information ecology. They all change systematically and once a factor has changed, it would affect the whole information ecology system. So a systematic view should be adopted in weighing and sizing up a information ecological problem

2.2 Balanced View

Like the natural ecological system, the static shift of balance and imbalance is the important part of the information ecological system in which the proportions of all the factors, their operating models, functional structures, allocation of resources and energy exchanges are in rela-

tively stable state. The relative balance results from the internal self-restoration and the regulation of external forces. However, the system has limited restoring capability by means of internal control and self-purification. Therefore, balance is relative and imbalance is inevitable.

2.3 Interactive View

Information ecosystem is an open ecosystem, the factors of which have a interrelated, interactive, intermagnetic and interdependent relation with other systems and the social economic environment. All people who own information in different fields at different levels, using different tools and vector to carry out their duties cooperatively and act on each other. Interactive view requires the elements of the system to be closely integrated and coordinate in development.

2.4 Humanistic View

Creatures make up the main part of natural eco-system while man is the main part of information eco-system. In the development of social science and technology and economy, the role of technology once was stressed excessively, leaving human being ignored. Man, which is the producer and consumer of information, is the most active factor of information eco-system. Information eco-system emphasizes the people-oriented concept, which means the final purpose of keep the balance of information eco-system is to maximize the benefit of people in the system.

2.5 Cycle view

For the information ecosystem to survive and develop, they should rely on the virtuous cycle of resources in each ecosystem and keep the resource flowing steadily and regularly both within the system and between the system and the outside, which keep the function and frame of the system. Or information eco-system will lose its balance, degrade and even collapse.

3 Enterprise Information Environment

The enterprise can constitute an information system and the information environment that it locates in is an important part of the ecosystem. When the information environment is propitious for knowledge innovation, we name it as the enterprise's innovation environment. One important purpose of the information management is to convert the simplex information environment to an innovation environment which can inspire staff's creativity. This has been reached a consensus in the field of knowledge innovation. Innovative environment is essential for the success or failure of innovation, Karzt's study in 1997 shows the failure of innovation is no longer for a simple technical reason, and more out of the complicated interaction between the different professionals and incentives of the technical personnel and inter-departmental staff . In practice, Gang Delin (2000) made a study of the 3M company with the world most innovative spirit for more than a decade and then summed up: 3M Company has been able to fully stimulate and improve the innovative energy of the whole staff, which the most prominent "Talent" is to have built a overall work environment in which a variety of factors promote each other's growth.

From the information ecological point of view, the enterprise itself, constitutes a unique ecosystem of information, with the three elements, namely its employees, enterprise information and enterprise information environment.

Business innovation occurs in the information ecosystem. This article believes that a broad enterprise information environment is the environment for innovation and enterprise. Employees work in the enterprise, in addition to accepting the narrow sense of corporate environmental information -- information management system to pass on his products, technology, sales such as areas' data, but also the humanities passed by the managers in the business, enterprise system, business workflow and so on. Such humanistic information and data information gives employees the psychological feelings, which constitute the broad information environment of the staff.

For example, the information of a respected business leader and a bureaucratic corporate leader to staff is completely different, which determines the employee's information environment, that is, the innovation environment is good or bad. In the final analysis a person's environment is a psychological environment, which is the psychological environment by the people themselves to the outside world of information consisting of subjective feelings. That is why some people are unhappy in the mansion, and some people take it easy in the poor house. In view of the ecological information, enterprise information environment is the business environment for innovation.

All elements of the enterprise information, which comes from the Enterprise strategic, the enterprise culture, the enterprise management, the enterprise organization structure information, various data of the knowledge management system and the demand of the outside market passed special information onto employees. These information promoting or hindering the employees' activities of knowledge transformation, that is, the employees' activities of innovation. All of them structure the enterprise innovation environment. To interpret enterprise innovation mechanism with information ecology concept is propitious to accelerate enterprise's work of building a information ecological environment which can support all-involvement innovation.

4 Enterprise Knowledge Innovation Mechanism

The innovation concept is first proposed by Schumpeter. He thought that innovation is a kind of new production functions, a new combination of production ele-

978-1-935068-76-1 © 2011 SciRes.

ment and the production function(Schumpeter,1934). Presently, the domestic scholars generally consider that innovation is a creative idea that comes true.

4.1 Enterprise Innovation Type

The classification of innovation types is the most controversial issue in basic research on innovation. At present, the two widely used classifications are: first, their innovative features and impact, innovations can be divided into "do-better innovation" and "do-different innovation"[5]. These two types of innovations are quite different in their subjects, innovation process, forms taken and the paradigms they use. Second, by subjects taken, they can be divided into strategic innovation, technological innovation, management innovation, system innovation, service innovation, product innovation, and so on.

4.2 Interpretation of the Enterprise Innovation Mechanism

The principal part of business innovation is people. To a certain enterprise information environment, people receive all kinds of information stimulus, have thinking activities, create innovative consciousness, carry out innovative behavior, and finally develop new knowledge. The means that the new knowledge generates in the brains of the staff is the core of business innovation mechanism.

The SECI model of Ikujiro Nonaka, to a certain extent, has revealed how the new knowledge generates in our brains. He thinks the knowledge creation is realized through the four courses that the explicit knowledge and tacit knowledge transformed, namely socialization, extemalisation, combination, and internalization, and that transformation occurs in the brains of the staff[6].

The accomplish of the extemalisation process that tacit knowledge transforms into explicit knowledge has a decisive meaning, some of this extemalisation take place after the innovation conscience generates, and the others must be finished after the innovative behavior being carried out.

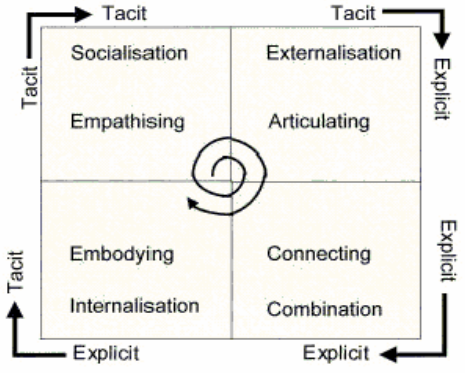


Figure 1. SECI model of knowledge transformation

Data source: Ikujiro Nonaka, Ryoko Toyama，Noboru Konno. SECI, Ba and Leadership: a Unified Model of Dynamic Knowledge Creation .Long Range Planning 2000 (33).

5 The Relationship between the Enterprise Information Environment and Knowledge Innovation

Employees work in a positive thinking and minds of continuously carried out in explicit and implicit knowledge into the needs of enterprise environment for innovation, which is broad support for the enterprise information environment. The enterprise's strategy, the culture, the management system, the information of its organization structure, various data of its knowledge management system and the demand of the outside market consist the entire information environment. Their relationship with the knowledge conversion—the core of the knowledge innovation is revealed in Figure 2.

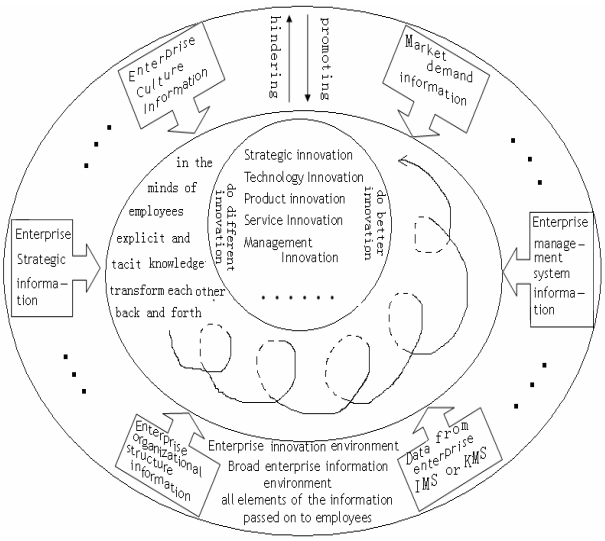


Figure 2. Enterprise innovation machanism

5.1 Information Environment is an Important Factor of the Knowledge Innovation System—Systematic View

It is helpful for the enterprise to consist an information ecology environment to promote knowledge innovation by using information ecology concept to interpret the relationship between enterprise information environment and knowledge innovation.

Enterprise innovation mechanism shows a systematic relationship. Employees of a enterprise receives information from all components of the enterprise information environment and carry out the innovative activity which is conducted in the information ecosystem of the enterprise, not operated on its own. Enterprise strategies, organization, administration, culture and information and

knowledge administration form a unitary system in which any change in one of them may invoke change of all. The staff, information and information environment are also interrelated and interdependent. Information environment is an important factor of the entire knowledge innovation system. We should admit it and take the advantage of this relationship, follow the new mechanism and improve the knowledge innovation.

Innovation can be promoted by accepting and employing the systematic relationship and acclimating innovation mechanism. For instance, whether the innovation is implemented by entrepreneurs, or experts of different specialties, or the strategy of the staff will have an effect on all the factors of the innovation environment and the staff's enthusiasm in innovation and the quality of the innovation.

5.2 An Imbalanced State of the Innovation System Caused by Information Environment—Balanced View

The core of the knowledge innovation is the staff's active thinking of knowledge conversion ,which needs motivation. And source of the motivation is the various imbalanced state in the enterprise ecosystem. For example, when the demand in the market does not tally with the number of the products manufactured, the enterprise need eliminate the imbalance, even meet and lead the market to reach a new balance by means of technical, product and service innovation. By the function of the enterprise information environment, the interaction of all factors and imbalance of all components in the information ecosystem occur all the time and become a driving force that keeps the Enterprise Innovation Mechanism alive. The self-adjustment in the enterprise when facing the imbalance guarantees its innovative capability.

5.3 The Interact Movement of the Enterprise Information Environment and Knowledge Innovation—Interactive View

Creation is a process that is obviously dynamic. Employees who want to make knowledge transformed must interact with the information he has accepted, and the elements in the information environment should also interact between each other. The styles of many kinds of behavior interaction have many differences, they followed their own disciplines and rules. Therefore, we must know about the creation mechanism on the base of dynamic process. The factors in internal creation environment of the enterprise should transmit the information fluently, resolve the contradictions, interact actively in order to provide protection to creation of company.

One of the basic characters of company information ecosystem is openness. As the business and political economic environments the company in make up different levels of information ecosystem, the company should be able to receive the information of the information ecosystem and distinguish it, at the same time, they output their own information to external information environment to exert the force of influence, interact actively to direct the creation of enterprise.

5.4 Staff is the Main Body that Using the Information to Carry out Knowledge Innovation—Humanistic View

In the present researches of company creation, they implement the view of humanism successfully. Through analyzing the mechanism of company creation, it is not hard to find that the key to the creation is the knowledge transmission which occurs in the mind of the employees, creation is the result of creative thinkings of human beings. The knowledge and skills that creation needs distribute widely in each employee. Therefore, by using information environment to deliver useful information, the enterprise can arouse every employee's participation, carry out effective cooperation and promote the entire staff's creativity. In the company creation, they must value everyone's effect but not the corporeal material conditions such as technology and equipment.

5.5 The Operation of the Knowledge Lifecycle is Pushed by Information Environment—Cycle View

Information ecosystem develops by the circulation of information and knowledge in its information environment and the alternation from one balance state to another. The development by means of cycle is a main force of the knowledge innovation. The new knowledge the staff generate by transforming their knowledge and innovative activities can be "new" for ever, rather it evolves with the cycle of knowledge generation, shift, application, aging and regeneration. The innovation is endless, the staff need to promote innovation by continuously transforming their recessive and dominant knowledge.

6 Conclusions

Innovation should be supported by data and knowledge generated from the workflow in the enterprise as well as the external environment. The information and knowledge come from the knowledge or information management system, which constitute a narrow enterprise information environment.

Enterprise innovation also needs incentives such as strategies from the business, management, enterprise culture, which transmit the information to the employees directly and have an effect on the staff mental environment and give employees subjective feelings about the business environment. These constitute a broad enterprise information environment.

At present, enterprises with innovation as its lifeline for the survival and development, broad enterprise information environment becomes the innovative business

978-1-935068-76-1 © 2011 SciRes.

environment.

From the ecological point of view, the broader business environment, business information and employees, together constitute the enterprise information ecosystem.

The core of the knowledge innovation is that by the inspiration of the information environment, the staff can make full use of all resources provided by the various elements in the information environment, carry out frequent transformation of explicit and tacit knowledge for the creation of new knowledge.

Interpret the relationship between the enterprise information environment and knowledge innovation with a systematic, balanced, interactive, humanistic and cycle view is helpful for the enterprise to explore new rules, build a harmonious enterprise information ecosystem and promote enterprise innovation under a complex information environment..

References

[1] Secondary source from Lu Jinrong, Guo Dongqiang. Progress in Study of Information Ecology Theory [J]. Journal of Information. 2007(3): 84-86.

[2] Xiao Feng. Philosophy Dimensionality of Information ecology[J]. Hebei Academic Journal, 2005 (1):49-54.

[3] Zhang Xinming, Wang Zhen, ZhangHong yan. Researches on the Construction of the Human-oriented Information Ecology System [J]. Information Studies:Theory & Application，2007(4):531-533.

[4] Cheng Peng,Xu Qian. Breakthrough Technology Development Dilemma by Using Information Ecological Concept [J], Pioneering with Science & Technology Monthly, 2008(3):8-10.

[5] Secondary source from Xie Zhangshu. The Mechanism and Management Mode of All-involvement Innovation from TIM Perspective[D].doctoral dissertation of Zhejiang University,2006.4:58.

[6] Ikujiro Nonaka, Ryoko Toyama, Noboru Konno. SECI, Ba and Leadership: a Unified Model of Dynamic Knowledge Creation [J]. Long Range Planning 2000, (33) 5±34.

Deep Analysis of the Copyright of Content Resources in Institutional Repository

Huan QIAO[1], Ying JIANG[2], Hefeng TONG[3]

[1]*Department of Information Management, School of Management Beijing Normal University Beijing, China, 100875*

[2]*Department of Information Management, School of Management Beijing Normal University Beijing, China, 100875*

[3]*Research Center for Strategic Science and Technology Issues Institute of Scientific and Technical Information of China, Beijing, China, 100038*

Email: qiaohuan1955@163.com, jy16881688@yahoo.com.cn, thf2003@istic.ac.cn

Abstract: The copyright of the content resources which the institutional repository uses is very complex. It becomes one of the critical factors for restricting the development of institutional repository. The content recourses which the institutional repository uses include four types: the copyright of the content belonged to institution, the copyright belonged to periodical office, the copyright belonged to author and the copyright belonged to the publisher. The review of the current situation of the copyright in the content construction of institutional repository can help researchers keep thinking in depth, pay attention to formulating customized schedule measures and workflow for different kinds of content. Thus, it will be helpful for speeding up the construction of the institutional repository.

Keywords: copyright; content resources; institutional repository

1 Introduction

For the past few years, Institutional repository (IR) has proved to be the leading role in the Open Access movement. More and more university and research institution have taken part in the troops of construction and development of institutional repository. According to the statistics of the Directory of Open Access Repositories (OpenDOAR): up to Mar.1, 2011, the number of the institutional repository the OpenDOAR contains is 1873, and 10 in China. Compared with the number which the writer gets in Feb.1, 2011, the number has increased of 200 in just one month. In the meantime the number of institutional repository increased at high speed, some troubles also become surfaced, as the copyright of the content resources in the institutional repository is one of them.

The writer has tracked many IR website, at the same time, consulted to amount of relevant document of the institutional repository in recent years, In order to deep analyze the copyright of the content resources of institutional repository, respecting to give some reference for the constructors of institutional repository and the follow-up study of the copyright of content resources.

2 The Copyright of Four Types of Content Resources in Institutional Repository

According to the statistics of the Directory of Open Access Repositories (OpenDOAR) in the types of content resources (Fig. 1): the content resources which the 1873 institutional repositories contain include journal articles, theses and dissertations, unpublished reports and working papers, books, chapters and sections, multimedia and audio-visual materials, bibliographic references, learning objects, datasets, patents, software and other special item types.

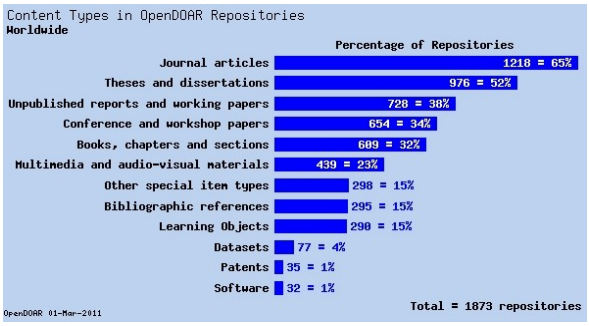


Figure 1. The statistics of OpenDOAR in the types of content resources

The writer divides the above content resources into four types according to the copyright they belonged to, the result is as follows:

2.1 The Copyright of Some Kind of Content Resources Belonged to the Institution

Unpublished reports, working papers, conference and workshop papers, and so on, which are the achievements of the institutions they belonged to, the copyright of them belonged to the relevant institutions. In line with the meaning of the institutional repository, the function of institutional repository is: keep deposit the whole scientific research and teaching achievements of the institution in order to give expression to the comprehensive scien-

978-1-935068-76-1

tific research ability of the whole institution. For this reason, these content resources ought to be deposited into the institutional repository of the relevant institution. By means of this kind of behavior, there are fundamentally no troubles in the copyright of this kind. Of course, in recent years, there are some institutions have worked out some relevant agreements for the reason of insurance of the copyright. For instance, the Chinese Academy of Sciences(CSIST) has worked out a agreement that between the Chinese Academy of Sciences and the authors who attend the meeting which holed by the Chinese Academy of Sciences about the electronic version of the authors' works, Chinese Academy of Sciences want to seek for the agreement of the authors to deposit their electronic version of their meeting works. The agreement says[1]:the certigier(means the authors who attend the meeting) can refer to the 3.0 version of Creative Commons(CC), can choose any compound mode from signature, freely available for noncommercial use, No Derivative Works and same in use to give authorization to the relevant institution of the pattern of their electronic version of the works that submit to the meeting.

2.2 The Copyright Some Kind of Content Resources Belonged to the Periodical Office which Geared to the Institution

Some universities or institutions in and outside the world have their own periodical office; the periodical office belonged to the universities or institutions. But they are independent unit, their output belongs to themselves. Such as the Library and Information Service which is founded by the National Science Library of Chinese Academy of Sciences, the Academic Journal of Beijing Normal University which is founded by Beijing Normal University. So, the copyright of one of the output: the electronic version of the journal belongs to the periodical office. According to the deposit require of the institutional repository, the journal which the periodical office compiled also is the research output of the relevant institution, it ought to be deposited into the institutional repository. But as the above analysis, the copyright of the journal does not belong to the institution but the journal periodical office. So, the institutional repositories must get the agreement of the journal periodical office if they want to deposit this kind of the content resources. For example, in allusion to this problem, The Chinese Academy of Sciences (CAS) announced an agreement with respect to the depositing and releasing the electronic version of the journal which is published by its research institution in the relevant institutional repository. The agreement also says [1]:the periodical office must upload the electronic version in the form of what the two parties agreed with in one month after the final formal publish of the journal, and at the same time, for the institutional repository, it must set up specific, public and lawful information to stick up for the authors and the journal according to the rule.

2.3 The Copyright of Some Kind of Content Resources Belonged to the Authors themselves

The copyright of the content resources as the dissertation, unpublished or the not transferring the possession thesis, teaching notes, original manuscript and courseware and so on belong to the authors themselves. Among these content resources, the dissertations refer to the most of the fruit of some research and experiment that take advantage of the materials, technical factors and some other conditions inside the relevant institution (especially the university or college) in line with the tradition in and outside the world. The university or college keeps the right of take precedence to make use of these dissertations for the purpose of collecting, preserving and making use of them in the procession of education and scientific research. For this reason, the majority of the universities have announced a rule that the students must submit the printing-out section and electronic section in the full text of their dissertations to the library after their thesis defense. At the same time, there is an addition, that is, there must be "the certificate of authorization for the use of the copyright of the dissertation of graduate degree" which is cosigned by the student writers and the mentor in the title page of the dissertation [2]. The certificate of authorization says that the university is entitled to retain the copies and the electronic version of the students' dissertation and at the same time, they also have the right to indicts all the content of the dissertation to the relevant database for the purpose of retrieving. It is means that in the meantime the university also has the right of retaining the dissertation in the institutional repository of the university. The Law of Copyright in China stipulates that the ownership of property can be totally or partly transferred by its author. So, the authors are entitled to deposit the materials such as the resources: the pre-print of electronic version, the dissertation which is unpublished or the copyright be not transferred, the teaching notes, the original manuscript, the courseware and so on to the relevant institutional repository and endow the most of web users the right such as copy and transmission in the web and so on .In this condition, the constructor of the institutional repository all that is need to conclude and sign an agreement of storage with the authors if they need this kind of materials deposited. On the basis of the recommendation of the AHDS, the content of the agreement of storage could consist of the right for the authors to legally deposit these materials to institutional repository, the right for the depositor of the institutional repository to own the content of what is deposited, the right for the institutional repository itself to maintain the content of what is deposited and at what kind of condition the institutional repository can delete some kind of materials, and so on[3] .

2.4 The Copyright of Most Official Publications Belonged to the Publishers or Some Third Party

As for the influence of the tradition mode of the publish industry, most of the official publications have assigned their copyright in the process of the publication. The project RoMEO have made a survey between the publishers of the academic journal at the point of the copyright agreement, and the results showed that[4], 90 percent of the publishers demand the assignment of the copyright. But the reality is that majority of the authors do not care the copyright of their works, they have no sense for whom the copyright should be owned by, that is amazing. Or, some authors have the pressure of the scientific achievements or some kind of competition in the institution, they have signed the protocol for the assignment of the copyright to assign the copyright of their works, so, the result is most authors of the works do not own the copyright of the published works, they can not spread their own research output as they like. Also, for this reason, they can not deposit their works of this kind to the institutional repository. In this condition, the feasible way is as follows if the constructor or the maintainer of the institutional repository wants to realize the deposit of this kind of content resources in the opinion of myself, that is, adjusting the agreement for the authors to assign the copyright of their works and concluding and signing an agreement for the deposit and releasing of the works of the members in the relevant institution with the relevant publishers.

By the means of the investigate and survey the content condition of the institutional repository at home and abroad, we find out that the most of the official publication which is deposited in the institutional repository is journal papers that contain research papers that are formally published by the learned periodical, the academic report, the letters for academic discussion and some other academic contents. Journal papers are the direct demonstrate and the direct confirm of the ability of the scientific research of the authors. The length of the journal papers is shorter than monograph, and they also occupy little room in the database, the number of the journal papers that is created by one author is very large. As for these advantages, the journal papers have become the chief form for deposit the content resources of the institutional repository of course. So, how the copyright assigned journal papers be deposited in the institutional repository, we will set forth now as follows.

1) Adjusting the agreement for the authors to assign the copyright of their works

What is the meaning of the title" Adjusting the agreement for the authors to assign the copyright of their works", that is, when the authors sign the agreement of the assignment of the copyright of their works , they will not assign all the copyright of their works, they choose to continue to have some kind of their rights to the works. For example, the right that the author can deposit the works in the institutional repository as he like, the right that the author make the works published and free use for most users who want to use the works as he like, and also the right for the author do not agree to sign the agreement of the copyright assignment with the publishers.

In order to help the authors to maintain their copyright of their own works from the publishers, many institutions have put forward some kind of the appendix of the agreement of the assignment of copyright for the authors. The meaning of the appendix is that it must enclosed herewith the agreement of claim of right for the author when the author signs the assignment of the copyright with the publisher[5]. The famous agreement appendixes include SPARC Author Addendum which is drawn up by the alliance of academic publish resources, the three kind of Author Addenda which is drawn up by Science Commons and the MIT Amendment to Publication Agreement which is drawn up by Massachusetts Institute of Technology[6].The institutional repository of QiJi in China also put forward a propose that the author should ask for the right of release their own works freely in the network from the publisher[7].In the wake of the vigorously spark plug of the institutional repository and the keeping development of the institutional repository, more and more authors have adopted the flexible authorization means such as keep all the copyright of their works, keep all the copyright of the works and at the same time share part of rights with the users, assign part of the copyright to the publishers and so on .In addition ,the author has the right of discretion to the printed copy of their works which includes depositing the works to institutional repository and some others. Because the publishers have got enough investment income from the issue of the printed journal papers, they should not capture every right of the works from the author exclusively. The agreement appendix mode which is drawn up by SPARC is a very representative embodies [8].The mode allows the author of the journal papers add their right to the copyright agreement with the publishers which includes the right for the authors to deposit their enunciable journal papers in the institutional repository and so on .The good news is that some publishers have accepted this mode already.

2) Concluding and signing an agreement for the deposit and releasing of the works of the members in the relevant institution with the relevant publishers.

More and more publishers have relaxed restrictions of the copyright of the results of self-achieving. For example, the Elsevier has permitted the authors to deposit the last print of their journal papers in their own or their institute's websites at the premise that the deposit must include the home page of the journal paper which is published and at the same time it is not for business and non-profit. Nature Publishing Group has indicated brightly that they do not need the copyright of the works

978-1-935068-76-1 © 2011 SciRes.

from the author any more, and allowed the author to release their journal papers in the relevant scientific research institutions in the condition of non-profit. While the appendix is the author must sign a agreement of copyright that the authors must offer the hyperlink of the website of the journal in the institutional repository after 6 months of the print of the journal papers[9]. As the statistical result of project SHERPA/RoMEO shows[10] (figure 2): date to Feb. 17, 2011, there is 62% publishers who support the self-achieving of some form in the number of all the 918 registered publishers in it.

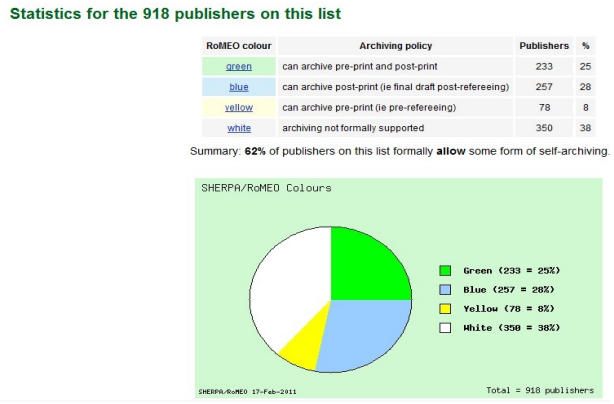


Figure 2. The statistical result of project SHERPA/RoMEO

Each publisher's entry is coded according to one of these color categories. The entry for each publisher also lists conditions or restrictions imposed by the publisher which govern archiving rights or activities. Conditions are taken as terms which can be easily accommodated and which do not hinder an author in archiving their work. A typical condition is to acknowledge the publisher's copyright in the work. Restrictions are more prohibitive, typically requiring some additional action on behalf of the author. Where a Restriction effectively blocks access to the eprint, such as in the case of an embargo on its public release, or requiring password- controlled access, then the partial archiving right is noted but the full color categorization does not apply. Sometimes open access discussions talk about "gold" publishers. This is a later development independent of RoMEO categories, and is used to describe publishers of open access journals. For the purposes of archiving, all open access journals allow archiving and can be taken as RoMEO "green". Some of the larger publishers have different archiving rights for different journals. This is particularly the case where they publish learned society journals on behalf of the society. A learned society might insist on a more liberal, or more restrictive archiving policy than the general publisher's copyright agreement allows. The RoMEO colour coding relates to the overall permissions given by a publisher. For example, a publisher has to apply the "green" archiving rights across all of their journals for their code to be "green"[10].

Although the majority of publisher and journals allow authors to archive their work under certain conditions, other publishers are more restrictive. Typically, when an article is published, the author assigns copyright, or gives a copyright license to the publisher. Depending on the particular agreement that is signed, the author retains more or less rights to use the article. Some agreements forbid the author from photocopying the article, using it in teaching, or mounting it on-line. Other agreements are more liberal and allow the author to retain rights to use the article as they wish.

Conclude and sign deposit agreement with the publishers is the most convenient and fast method to deposit journal papers into the institutional repository at present. Although the copyright of the journal papers has owned by the publishers, that is not mean that the journal papers are not the scientific research results of the institution where the authors are. The institutions where the authors are have the responsibilities and obligations to promote and guarantee the research results their numbers created using the public capital can be used widely. The agreement between the institution and publishers can clearly and definitely raise the right for their numbers to use, release and recreate the journal papers in the process of educating or scientific researching. At the same time, they should also agree some clauses that would safeguard the profit of the publishers, for example, the agreement about strictly prohibiting using the institutional repository to do business as the agreements with the Elsevier or the Nature say. It may include some other clauses, for example, setting up some public summons to maintain the legal right between the author and the learned periodical; putting down in writing the issue information of the journal paper in the institutional repository and make the learned periodical as the first publish resource of the journal paper; supervisory controlling the process of encroach on the author and the learned periodical by the three of the institutional repository, the author and the learned periodical. at the same time ,punishing the tort according to the law; including the source and the link of the journal paper when the institutional repository provide the metadata of the content to other retrieving institutions; asking for the institutional repository ,its members and users to comply with the creative commons licensing scheme., and so on. This licensing scheme should be encouraged by the policy as an option. In the nutshell, Creative Commons seeks to move from ‘all rights reserved’ as the default in the copyright system to a system where authors have some rights reserved (or no rights reserved at all). ‘All rights reserved’ means that the author or creator of IP assumes exclusive rights (save for the exceptions and limitations). The CC licensing scheme has different licenses which stipulate the kind and extent of rights reserved by the author. The CC Attribution-ShareAlike (BY-SA) requires that work distributed, under that license can be shared and adapted, but

the originator must be acknowledged. Any subsequent work drawing from that work must be distributed under the same license. The other CC license of interest is the CC Non-commercial (NC), which requires that work can be shared but only for non-commercial purposes. Any commercial activities (such as adaptations) require permission from the originator or rights-holder. Wikipedia5 has an excellent matrix of possible license options (common one) under cc scheme:

1. Attribution alone (by)
2. Attribution t Non-commercial (by-nc)
3. Attribution t NoDerivs (by-nd)
4. Attribution t ShareAlike (by-sa)
5. Attribution t Non-commercial t NoDerivs (by-nc-nd)
6. Attribution t Non-commercial t ShareAlike (by-nc-sa)

For example, some institutions such as the Harvard University, the Massachusetts Institute of Technology, the Univ Goettingen Germany, the Max Planck Society and so on have conclude and sign the framework agreement of open access with part of the publishers. The Chinese Academy of Sciences as the first institution in Asian-Pacific region has conclude and sign the framework agreement of open access with the Spinger Group in Oct.2010[11].The agreement clearly and definitely allows the authors of Chinese Academy of Sciences who have publish journal papers in the Springer can deposit their final text of their journal paper in the relevant institutional repository of the Chinese Academy of Sciences, and can be obtained in 12 months after the journal papers delivered. It also ruled that if there is any disaccord with the reality of the transfer of the journal papers, the framework agreement has the right of way.

3 Conclusion

In allusion to the deep analysis of the copyright of the four kind content resources, as the great variety of the content resource and the complex copyright problem, we should keep thinking in depth, pay attention to formulating customized schedule measures and workflow for different kinds of content. Guarantee that the quantity and the quality of the content resources in the institutional repository in the legal condition. Thus, it will be helpful for speeding up the construction of the institutional repository.

References

[1] CAS IR Grid[EB/OL].[2011-02-26]. http://210.77.64.222/.

[2] Wang Yafeng.Deep analysis of the laws in institutional repository of academic library[J]. Journal of Academic Library and Information Science, 2010, 28(5): 9-11.

[3] Feng Xiaohong. The practice and enlightenments of dealing with copyright problems of institutional repositories in foreign countries[J]. Journal of Library Construction, 2009(12): 4-8.

[4] ReMEO: Directory of Publishers Who Have Given Their Green Light to Self- Archiving [EB/OL]. [2011-01-07]. http://romeo.eprints.org/.

[5] Romeo: How Academics Want to Protect The Open-Access Research Paper [EB/OL]. [2011-01-17]. http://wwwlboroacuk/departments/romeo/

[6] Hu Fang,Zhong Yongheng. Research on the Copyright of Institutional Repository [J]. Library and Information service, 2007(7): 50-53.

[7] Miracle Library[EB/OL]. [2011-03-01] . http://www.qiji.cn/.

[8] SPARC[EB/OL]. [2011-02-13]. http://www.arl.org/sparc/.

[9] Li Chenying,Yang Guodong, Zang Lin.The settlement of the journals'copyright in the institutional repository[J]. Journal of Academic Libraries, 2010(1): 74-79.

[10] SHERPA/RoMEO [EB/OL].[2011-01-17]. http://www.sherpa.ac.uk/romeo/.

[11] Chinese Academy of Sciences signs the agreement of open access with Springer Group [EB/OL]. [2011-02-13]. http://www.cas.cn/xw/yxdt/201011/t20101102_3001705.shtml.

978-1-935068-76-1

Analysis of Copyright Issue about Open Access Resource for Long-term Preservation

—Take the Library as an Example

Chen WEI

Department of Management, Beijing Normal University, Beijing, 100875

Email: bnuweichen@126.com

Abstract: Long-term preservation of digital resources is the key to the success of open access, while copyright issue has restricted the development of long-term preservation. This paper takes the library for example to analyze the possible copyright issues about open access resources for long-term preservation, with reference to the related domestic and foreign copyright laws and regulations, so as to make some suggestions to help library carry out the work of long-term preservation.

Keywords: long-term preservation; open access; copyright

开放存取期刊长期保存的版权问题分析

——以图书馆为例

魏晨

北京师范大学管理学院，北京，100875

Email: bnuweichen@126.com

摘　要：数字资源的长期保存是开放存取发展成败的关键，而版权问题则是一直制约着长期保存发展的瓶颈。本文结合国内外有关长期保存中版权问题的法律法规，以图书馆为例，分析开放存取期刊在长期保存中可能遇到的版权问题，对图书馆开展长期保存工作提出建设性意见。

关键词：长期保存；开放存取；版权

1 引言

1.1 开放存取

2001 年，开放协会研究所（Open Society Institute，OSI）在布达佩斯召开了开放存取国际研讨会，起草和发表了《布达佩斯开放存取计划》（Budapest Open Access Initiative，BOAI），这标志着“开放存取运动”正式诞生。《布达佩斯开放存取计划》（BOAI）中对开放存取的定义是指对网络公共领域里文献的免费获取，它允许用户对全文阅读、下载、复制、传递、打印、检索及链接。作为一种高效、公平的学术信息交流新机制，开放存取从一定程度上解决了期刊价格上涨与图书馆经费紧张之间的矛盾，受到学术界的广泛关注，学术影响力日益提高。[1]

1.2 开放存取的长期保存

目前一些较为著名的外文数据库（如 Elsevier）所收录的期刊有一大半都可以开放获取，这说明了开放存取资源的数量巨大，也间接证明了开放存取资源的高质量。国际图书馆协会和机构联合会（The International Federation of Library Associations and Institutions，简称国际图联，IFLA）与国际出版者协会（International Publishers Association，IPA）在《永久保存世界记忆：关于保存数字化信息的联合声明》中也有如下表述：“日益增长的仅以数字形式出版的信息如同传统印刷型出版物一样重要，同样具有长期留存的文化价值和历史意义。”“尽管长期保存这类信息的费用是昂贵的，但不保存则将造成灾难性的损失。”[2]开放存取当然是这类信息中很重要的一部分。由于未加以长期保存或保存不当，目前已经有一些开放存取数字资源消失，不能再获取。

1.3 开放存取的长期保存与图书馆

图书馆作为一种公益性质的文化机构，其保存职能就是保存历史和文化，使文明得以延续。作为开放存取在实施过程中的主要问题之一，数字资源的长期保存面

临着许多问题。国际图联与国际出版者协会的联合声明指出“出版机构应该担负短期保存的责任，长期保存的责任由图书馆承担”。如果仅能访问但并不拥有文献资源将会危及图书馆作为文化遗产保管者的角色；而如果仅进行保存不提供访问和使用，那又失去了对数字资源进行长期保存的意义。据统计，目前全球15%的开放存取资源已经消失[3]，因此，我们有必要就图书馆与开放存取资源长期保存的问题进行探讨。而开放存取资源的长期保存是一个动态过程，涉及的版权问题和如何解决这些问题与整个保存过程密切相关。

2 国内外有关长期保存版权问题的法律法规

《中华人民共和国著作权法》规定著作权人，即版权所有者，享有发表权、署名权等十七项权利[4]；世界知识产权组织在《版权及相关权利基本概念》中，将版权划分为复制权、公开表演及传播权、翻译和改编权、道德权四项[5]。根据图书馆对开放期刊进行长期保存的整个过程，以下将主要从资源的复制、存档和传播三个角度对国内外有关法律法规进行梳理。

2.1 国内情况

我国与长期保存版权问题有关的法律主要集中于《中华人民共和国著作权法》（2001修订，以下简称《著作权法》）、《中华人民共和国著作权法实施条例》（2002，以下简称《实施条例》）和《信息网络传播权保护条例》（2006）。

资源的复制方面，《著作权法》第二十二条第八款指出，图书馆、档案馆、纪念馆、博物馆、美术馆等为陈列或者保存版本的需要，复制本馆收藏的作品，可以不经著作权人许可，不向其支付报酬，但应当指明作者姓名、作品名称，并且不得侵犯著作权人依照本法享有的其他权利。[4]

资源的存档方面，《实施条例》第二十三条规定，使用他人作品应当同著作权人订立许可使用合同，许可使用的权利是专有使用权的，应当采取书面形式，即为订立合约[6]。但目前尚无明确的法律条款对资源存档的具体操作进行规定和约束。

资源的传播方面，《信息网络传播权保护条例》第七条规定，图书馆、档案馆、纪念馆、博物馆、美术馆等可以不经著作权人许可，通过信息网络向本馆馆舍内服务对象提供本馆收藏的合法出版的数字作品和依法为陈列或者保存版本的需要以数字化形式复制的作品，不向其支付报酬，但不得直接或者间接获得经济利益。规定的为陈列或者保存版本需要以数字化形式复制的作品，应当是已经损毁或者濒临损毁、丢失或者失窃，或者其存储格式已经过时，并且在市场上无法购买或者只能以明显高于标定的价格购买的作品。[7]

此外，由最高人民法院于2004年1月发布并生效的《关于审理涉及计算机网络著作权纠纷案件适用法律若干问题的解释》（2006年修改）明确指出，《著作权法》中规定的各类作品的数字化形式同样应受到著作权法的保护。已在报刊上刊登或者网络上传播的作品，除著作权人声明或者报刊、期刊社、网络服务提供者受著作权人委托声明不得转载、摘编的以外，在网络进行转载、摘编并按有关规定支付报酬、注明出处的，不构成侵权。但转载、摘编作品超过有关报刊转载作品范围的，应当认定为侵权。[8]

2.2 国外情况

相比于国内，国外的有关立法则开展得较早。早在1886年英国、法国、德国、意大利等10国就在瑞士伯尔尼签署了《保护文学和艺术作品伯尔尼公约》（Berne Convention for the Protection of Literary and Artistic Works，简称《伯尔尼公约》）。《伯尔尼公约》第九条第二款“本联盟成员国法律有权允许在某些特殊情况下复制上述作品，只要这种复制不致损害作品的正常使用也不致无故危害作者的合法利益。”[9]这是有关资源复制的最早的法规条款，《世界知识产权组织版权条约》第十条[10]，《世界知识产权组织表演和录音制品条约》第十六条中也有相似规定[11]。

欧洲议会和理事会2001/29/EC关于版权和相关权利问题的指令（Directive 2001/29/EC of the European Parliament and of the Council）第五款则规定了对复制权限制的“例外情况”，对公众开放的图书馆、教育机构以及博物馆和档案馆等非商业行为，以及用于学术交流而进行的传播等。[12]

资源的存档方面，英国为确保文献资源得以长期保存，建立了缴送制度。缴送制度是以国家法律形式规定信息资源生产者向国家指定图书馆免费提交资源的一种制度。随着数字资源的发展，英国2003年新制定的《法定缴存图书馆法》（Legal Deposit Libraries Act），对1911年版权法中有关印刷出版物缴送的条款进行了修改，增加了数字资源呈缴的相关内容，要求出版商向英国和爱尔兰的6个图书馆提供数字资源存储。其中第六条详细规定了有缴存义务的单位和个人所需呈缴数字资源的内容，包括相应的副本、电脑

978-1-935068-76-1 © 2011 SciRes.

程序、介绍信息等，如果该数字资源拥有多个版本及格式，则应将其一并呈缴。[13]

资源的传播方面，《伯尔尼公约》第十条对可以免费使用作品的几种行为进行了规定，用于教学与科研的引用及说明，用于新闻广播等大众传媒。虽然规定了使用者必须标明作品的来源及作者，并且规定只要是在为达到正当目的所需要的范围内使用，并且符合正当习惯，即可由本法律联盟成员国以及成员国之间现已签订或将要签订的协议加以规定，违反规定的会受到各国法律的制裁，但是对“正当目的”没有做出明确定义，对于符合正当习惯的行为也没有做出明确说明。[9]

美国 1998 年颁布的《数字千年版权法案》（Digital Millennium Copyright Act）为网上作品著作权的保护提供了法律依据。规定没有版权所有者的允许，对其作品进行复制将属于违法行为；同时规定删除或者绕过版权所有者为防止复制所进行的技术控制也属于违法行为。虽然这种技术保护措施能够有效防止未经授权使用作品，但却不可避免地影响了版权法下版权所有者与公众之间的权利平衡，一些版权法规定公众所享有的权利，如合理使用、图书馆特权等，受到了侵害。最新通过修订的美国著作权法（2007）也已经包括了其相关内容。[14]

国际图联在其 2000 年关于数字环境下的版权声明中（IFLA Position on Copyright in the Digital Environment）针对版权法中涉及数字格式作品的条款做了相关规定，但没有涉及具体的法律规定。对于数字形式的作品，不必付费或寻求授权，图书馆所有用户都享有以下权利：公开浏览可获取的版权作品；在馆内或以远程登陆方式私人阅读、聆听或观看市场上公开销售的版权作品；为了个人教育或研究需要，复制或通过图书馆和信息工作人员复制合理数量的数字作品。版权法应该涵盖电子媒介的法定存储问题；版权法应以平衡版权所有者和用户的利益为目标，可以通过技术手段保护版权所有者的利益，而对合法的、无侵权目的的用户，则可规避这些技术措施；版权法应确切地阐明，在版权法不可能实际或合理实施的情况下，第三方应负的责任的限定。[15]

作为国际图联的会员之一，中国图书馆学会也于 2005 年正式公布了《中国图书馆学会关于网络环境下著作权问题的声明》。除了支持国际图联的立场之外，还指出我国著作权保护所应特别关注的三个问题：著作权立法应充分行使国内立法权；实现信息网络传播权保护与限制的平衡；设计和制定简便易行的许可使用协议制度。[16]

3 图书馆开放存取期刊长期保存中的版权问题与解决办法

图书馆作为长期保存的重要机构，其提供服务的方式主要有以下两种：图书馆作为保存责任者向用户提供服务；图书馆与保存责任者合作向用户提供服务。前者是指图书馆自建数据库对开放存取期刊进行长期保存，还包括一些开放存取的出版商将资源交给图书馆代为长期保存，如生物医学期刊出版中心 BioMed Central 为了使自己拥有的 170 多种生物医学杂志能够得以长期保存而在美国国家医学图书馆等 4 家机构做了镜像保存。后者是指图书馆通过签订协议的方式从出版机构接收和获取资源，自身拥有数据库的开放存取出版商以及受出版商委托的第三方存储机构才是担任保存的责任者，如荷兰国家图书馆与数字资源长期保存服务商 Portico 签订协议，由 Portico 向国家图书馆提供其收藏的 600 多万篇文章的离线拷贝。

3.1 图书馆作为保存责任者的版权问题与解决办法

首先，图书馆自建数据库对开放存取期刊进行长期保存，可以在许可期限已满以后继续保持对期刊的长期保存和永久访问。大英图书馆自 1999 年开始，成立专门团队对自身馆藏资源进行数字化，对一切数字资源进行保存，并为读者提供数字化服务。[17]但是，这会涉及到版权问题中的存档权、复制权、以及保护作品完整权。除图书馆自身拥有或出版的期刊以外，对外采的期刊进行长期保存，需要图书馆方面获得对作品的存档权以及对存档数据的处理权。在对开放存取期刊进行存储的过程中需要利用备份、迁移等技术，这又涉及到作品的复制权以及保护作品完整权。此外，新西兰、英国、丹麦等国家也已制定出台了相关法律和缴送制度，确保图书馆有收藏和保存数字资源的权限。加利福尼亚大学图书馆的数字保存计划也在着手开发由数字保存仓储（Digital Preservation Repository, DPR）和其他工具组成的基于公认标准的基础结构体系，以支持学术信息的识别、获取、描述、组织和持久管理等。[18]

其次，对于一些开放存取出版商交由长期保存的资源，图书馆也需要通过与之签订协议，在获得出版商及版权所有者的许可之后方能将其所保存的资源向

授权用户提供服务。2002 年，荷兰国家图书馆同 Elsevier Science 签署了里程碑性的归档协议，将 Elsevier 出版的电子期刊保存在 e-Depot 数字存档系统以确保其长期使用，并且荷兰国家图书馆也可以将其所保存的 Elsevier Science 数据向到馆读者提供服务。[19]

3.2 图书馆与保存责任者合作的版权问题与解决办法

首先，图书馆直接从出版商自身的数据库获取开放存取期刊的数据。多数情况下，图书馆都是通过与出版商签订采购合同来获得其数字资源的使用权。一旦合同终止，或者出版商出现破产、合并等情况，或者所采购的期刊被移交给另外的出版商时，都可能使图书馆丧失对这些数字资源的持续使用权。此外，对于合同签订之前出版的数字资源图书馆是否也有权使用。对这些问题的解决，只能依赖于更加具体和详细的立法规定，以及对权利义务、服务范围都更加明确和完善的协议规定。

其次，图书馆获取的数字资源来自于受出版商委托的第三方存储机构。这与图书馆直接从出版商自身的数据库获取资源类似，只是出版商出于技术费用等考虑委托第三方存储机构对其开放存取期刊的数据进行存储和维护。在出版商与存储机构已经存在合法协议的基础之上，图书馆可以直接与存储机构签订协议或合同，对其所存储的数字资源进行有效利用。这其中最具代表性的是 Portico 项目，该项目由 JSTOR 过刊数据库和美国国会图书馆（The Library of Congress）共同主持。通过直接与出版商、图书馆签订保存许可协议，Portico 一方面从出版商那里获取源文件，把不同的文件转换成标准的、可长期使用的存储格式；另一方面 Portico 为图书馆保存其所采购的数字资源并在突发条件下为其提供数字资源的访问权。[20]

无论图书馆采取哪一种服务方式，其对长期保存的开放存取资源都不能不加限制地提供服务。若不加限制，不但可能侵犯版权所有者的复制权、信息网络传播权等，还可能影响到版权所有者的其他经济利益和信息服务市场的正常运转，进而阻碍数字资源的广泛传播和公众的知识获取。图书馆利用长期保存的开放存取资源提供服务，应在法律许可或协议许可的范围内，向拥有授权或者协议规定的用户范围提供服务，并且必须保护数字资源的版权，主要有整体署名权和作品完整权等，对用户合理使用数据尽到合理的管理责任，在保护公共利益的同时也尽力维护版权所有者的正当利益。

4 结语

原则上，图书馆及其用户都需要确保没有任何经济或地理障碍地获取电子期刊，但事实并非如此。为了克服成本、时间和低效率的障碍，获取电子期刊的新形式开放存取已经形成并越来越流行，而开放存取的流行也引起了人们对数字资源长期保存的关注。图书馆作为保存历史和文化的机构，进行开放存取资源的长期保存并提供服务也是必然的。但是这不仅需要有国家政策的保障，需要制定保存的国家标准，还需要完善相应的法律法规，建立相关的合作机制。国外图书馆在这方面的许多尝试和实践则可以为我们提供参考借鉴，如荷兰国家图书馆在数字资源缴存方面所做的实践表明，通过协议方式实现国内数字资源的缴存和国外数字化学术文献的采集保存是一种现实和可行的选择，其实践对我国建立国家数字资源长期保存和读取制度亦具重要参考价值；大英图书馆在其 2006 年制定的《数字保存策略》中提出了一个实现长期保存的十年计划，并对可能存在的风险进行了分析[21]，对于国内图书馆也有一定的参考意义。只有解决了开放存取资源在长期保存中所出现的版权问题，才能使开放存取资源得以长期妥善的保存和利用。

References (参考文献)

[1] http://www.soros.org/openaccess/read.shtml [2010-10]

[2] Preserving the Memory of the World in Perpetuity: a joint statement on the archiving and preserving of digital information.
永久保存世界记忆：关于保存数字化信息的联合声明. http://archive.ifla.org/V/press/ifla-ipa02-cn.pdf [2011-01]

[3] Feng Xiaowei. Influences of the Long-term Preservation of the OA(Open Access) Resources on the Development of Library[J]. *SCI-TECH INFORMATION DEVELOPMENT&ECONOMY*, 2008, 18(25), P26-27.
冯效卫. 开放存取资源长期保存对图书馆发展的影响[J]. 科技情报开发与经济，2008, 18(25), P26-27.

[4] http://www.gov.cn/banshi/2005-08/21/content_25098.htm [2010-10]

[5] Basic Notions of Copyright and Related Rights. http://www.wipo.int/copyright/en/activities/pdf/basic_notions.pdf [2011-01]

[6] http://www.people.com.cn/GB/14677/40759/41276/3022343.html [2010-11]

[7] http://www.gov.cn/zwgk/2006-05/29/content_294000.htm [2010-10]

[8] http://www.court.gov.cn/qwfb/sfjs/201006/t20100604_5810.htm [2010-11]

[9] Berne Convention for the Protection of Literary and Artistic Works. http://www.wipo.int/export/sites/www/treaties/en/ip/berne/pdf/trtdocs_wo001.pdf [2010-11]

[10] WIPO Copyright Treaty.

978-1-935068-76-1 © 2011 SciRes.

http://www.wipo.int/treaties/en/ip/wct/pdf/trtdocs_wo033.pdf [2010-11]
[11] WIPO Performances and Phonograms Treaty. http://www.wipo.int/export/sites/www/treaties/en/ip/wppt/pdf/trtdocs_wo034.pdf [2010-11]
[12] Directive 2001/29/EC of the European Parliament and of the Council. http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2001:167:0010:0019:EN:PDF [2010-11]
[13] Legal Deposit Libraries Act 2003. http://www.legislation.gov.uk/ukpga/2003/28/data.pdf [2010-11]
[14] Copyright Law of the United States. http://www.copyright.gov/title17/circ92.pdf [2010-11]
[15] http://archive.ifla.org/V/press/copydig.htm [2011-01]
[16] http://www.lsc.org.cn/CN/News/2006-04/EnableSite_ReadNews13633071143993600.html [2010-11]
[17] http://www.bl.uk/ [2011-01]
[18] http://www.cdlib.org/services/uc3/dpr.html [2011-01]
[19] http://www.elsevier.com/wps/find/authored_newsitem.cws_home/companynews05_00020 [2011-01]
[20] http://www.portico.org/ [2011-01]
[21] British Library Digital Preservation Strategy. http://www.bl.uk/aboutus/stratpolprog/ccare/introduction/digital/digpresstrat.pdf [2010-11]

Analysis to the Sharing Obstacles and Effective Ways of the Implicit Knowledge under the Network Environment

Take the Business Synergetic Platform of Science & Technology Information Research of Shandong Province as an Example

Shunmei YUAN, Jian WANG, Changmei JIANG, Fei TANG

Center for Innovation and Strategy Research, the Institute of Science and Technology Information of Shandong Province, Jinan, China, 250101

Email:happystar04@126.com

Abstract: Sharing the Implicit Knowledge is an important guarantee of improving the quality of knowledge service. This paper analyzes the sharing obstacles and effective ways of the Implicit Knowledge under the network environment, taking the Business Synergetic Platform of science & technology information research of Shandong province as an example.

Keywords: Implicit Knowledge; sharing obstacle; effective way; business synergetic platform

1 Introduction

Knowledge service is such a service, which is driven by user's goals, through sorting, matching and refining various explicit and implicit information resources which can be utilized. It is based on the collection, organization, analysis, recombination of the information and the knowledge, and facing the contents and the solutions. Knowledge can be divided into the Implicit Knowledge and the Explicit Knowledge. Among them, the Explicit Knowledge is easy to be communicated, shared, imitated and copied, can be obtained by ordinary learning and accumulation, and be internalized to Implicit Knowledge by digestion and absorption. While the Implicit Knowledge is difficult to be obtained by the ordinary way and to be communicated, shared, imitated and copied, and it plays a vital role in realizing the individual value objectives and improving the organizational core competitiveness. Sharing the Implicit Knowledge can contribute to the generation of the new knowledge and the new viewpoint. It is the source of knowledge's innovation, having great significance for good operation of the projects and lifting the qulity of the knowledge service. But there are plenty of obstacles which block the road to sharing the Implicit Knowledge, especially in the network environment, where people's face-to-face communication opportunity decreases. Here, the authors take the Business Synergetic Platform of science & technology information research of Shandong province as an example, analyze the sharing obstacles and effective ways of the Implicit Knowledge under the network environment, in order to speed up the flowing of the knowledge and the information, accelerate the sharing and recreation of the knowledge, and improve the quality of the knowledge service.

2 Overview of Explicit Knowledge and Implicit Knowledge

Knowledge is the real dominant resource and decisive factor of production at present, and it is the most important resource in modern enterprises and social economic development. It can be divided into Implicit Knowledge and Explicit Knowledge, which was put forward by Polanyi in 1958 in "personal Knowledge".

2.1 Definition and Features of Explicit Knowledge and Implicit Knowledge

Explicit Knowledge is such a knowledge, which can be expressed by formal and systematic language and can be shared in the shape of data, scientific formula, specification and operation. It can be coded, easily copied and quickly disseminated.

Implicit Knowledge is on the contrary,which is highly personalized, deeply rooted in the process, behavior, communication and interaction, practice, responsibility, vision and feeling. It is under the influence of the personal experience and knowledge, and it is difficult to be copied, imitated, formulated, clearly expressed and formed a unified format.

Table 1. The different features of explicit knowledge and Implicit Knowledge①

Explicit Knowledge	Implicit Knowledge
easy to be coded and formated	difficult to be coded and formated
theoretical, may be not gained by practice	practical and path-dependence
expression in the form of character, language, chart and also	expression in the form of inspiration, experience, knack and also
non-monopolistic	monopolistic
Easy to be communicated, shared and spread	Difficult to be communicated, shared and spread

2.2 Role of the Sharing of the Implicit Knowledge

Implicit Knowledge often has more creative value than Explicit Knowledge in the practical work, and it is the underlying intelligence cost and the key of the knowl-

978-1-935068-76-1

edge innovation. Sharing Implicit Knowledge can increase individual and institution's quantity of achievements' output, improve its quality, shorten its time and reduce its cost. Sharing the Implicit Knowledge of members in the organization is the magic weapon of enhancing the core competitive ability and making the organization impregnable. In the knowledge society, and it is also the important guarantee of stimulating the sustainable development of the organization.

3 Introduction of the Business Synergetic Platform of Science & Technology Information Research of Shandong Province

3.1 Background of the Business Synergetic Platform

The development history of the science & technology information research have been more than 50 years. During this period, the whole province's research institutes at various levels come out from scratch, and constantly improve the functions on the technology information's collection and sorting, storage and transmission, analysis and research and so on. The researchers' team is stronger and stronger, the means of information service have been varied, the above promote the development of the science & technology information research powerfully, and become the backbone in local department's work of science & technology management.

But at the same time, because of the restriction of region and industry etc, the fragmentation phenomenon of the information research is very serious, the "big information" situation hasn't formed between the science & technology information institutions, which restricts the further development of the information research. Today, the knowledge service is becoming highly valued. To realize the fast knowledge innovation, the effective way is to organize the persons who may have different knowledge background and may come from different fields together to use interactive and cooperative innovation for common tasks[②]. To promote information research to deep development and to make information research keep up with the situation and reach a higher level so as to serve the society better, changing "fragmentation" phenomenon to "cooperation" phenomenon, realizing the collaborative research is carried on important agenda. The booming development of the network and the openness of the network environment provides the possibility and opportunities for breaking the "fragmentation" phenomenon and realizing the cooperation phenomenon. In this context, we prepare to establish the Business Synergetic Platform of science & technology information research of Shandong province.

3.2 Logical Structure of the Business Synergetic Platform

The synergetic work function of the Business Synergetic Platform is realized mainly by four work modules, they are "post management", "human resource management", "assessment statistics" and "project management". Figure 1 shows the mutual relationship.

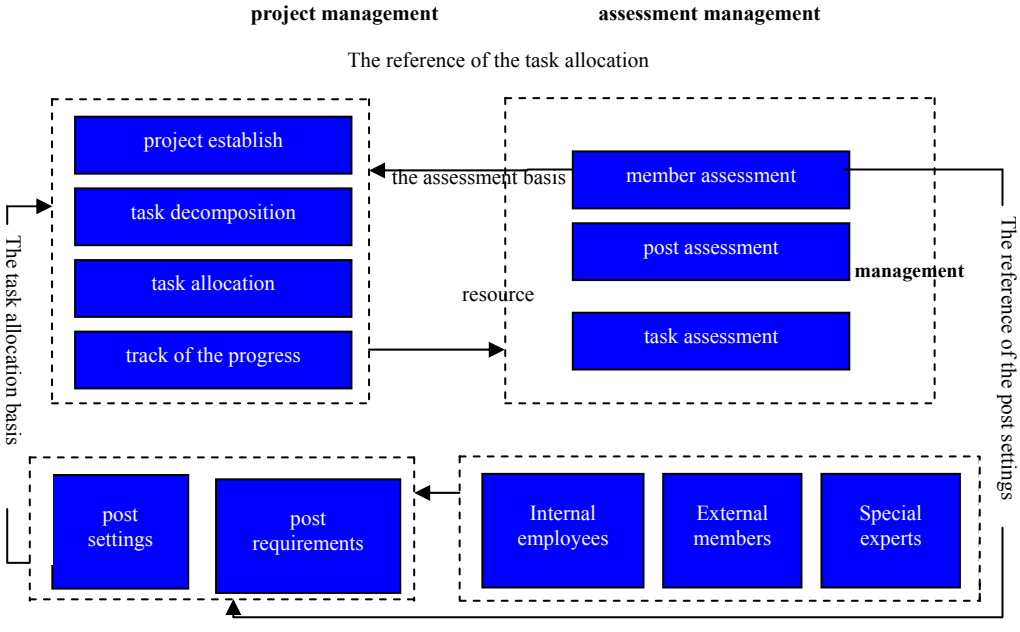


Figure 1. The Logical Structure of the Business Synergetic Platform

3.3 Source of the Implicit Knowledge on the Business Synergetic Platform

The Implicit Knowledge on the Business Synergetic Platform mainly comes from the individuals in the "human resources management" module, which is composed by "internal employees", "external members" and "special experts". Among them, the "internal employees" unit is according to the specific requirements of the system construction and manage the persons who come from the interior of the information institute, participate this system and have the post qualifications. The "external members" unit is according to the specific requirements of the system construction and manage the persons who come from the exterior of the information institute, participate this system and have the post qualifications, such as the researchers from the other cities and the other institutes. The "special experts" unit is according to the specific requirements of the system construction and manage the experts and scholars who are academic leaders in some domestic fields or who are capable of managers of the research topics. The human resources in these three units divide and cooperate. They play individual research expertise and form the project team, providing service together for the science & technology decision of the Shandong province.

No matter who they are, internal employees or external members or special experts, they all accumulate plenty of the Explicit Knowledge and the Implicit Knowledge in the course of actual work and study. The Implicit Knowledge mainly includes belief, insight, inspiration, intuition, way of thinking and research feelings which can't be or still not be opened. The American scholar named Roy Lubit divided the Implicit Knowledge into four classes, they are the skills which are difficult to express, the mental model, the ways that handle the issues and the organizational routines[③]. The first three can be regarded as the Implicit Knowledge which are stored in the individual minds. According to this, we can divide the Implicit Knowledge of the human resources on the Business Synergetic Platform into three classes (see Figure 2).

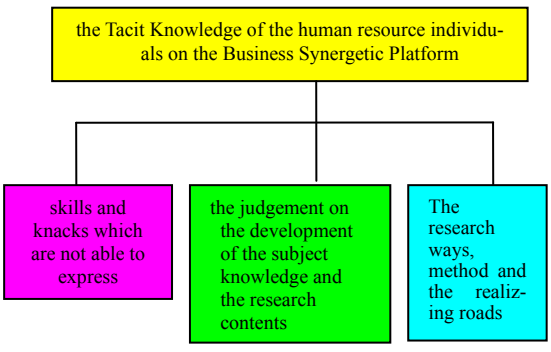


Figure 2. The Implicit Knowledge of the human resource individuals on the Business Synergetic

4 Main Obstacles of the Sharing of the Implicit Knowledge on the Business Synergetic Platform

4.1 Objective Obstacles

4.1.1 Character of the Implicit Knowledge that can't be Expressed Exactly

One of the objective obstacles of the Implicit Knowledge explicating is caused by the character of the Implicit Knowledge itself. It is not easy to be expressed exactly, and it is not standard, structured, systematic and coded, besides, it is difficult to flow. The researchers on the Business Synergetic Platform have plenty of the Implicit Knowledge, but they can't be expressed exactly, which hinder the sharing of the Implicit Knowledge between the researchers on the Business Synergetic Platform.

4.1.2 Sharing of the Implicit Knowledge Relies More on Face-to-face Contact

The sharing of the Implicit Knowledge relies mainly on the direct face-to-face contact between people, it is a connected transmission in the process of the life, work and study[④]. But the character of the network's virtualization makes the face-to-face interaction opportunity greatly reduced, which affects the sharing of the Implicit Knowledge. On the Business Synergetic Platform, the opportunity of the face-to-face contact is not very much, so the spread of the Implicit Knowledge encounters obstacles to some extent.

4.1.3 Different Members Have Different Professional Backgrounds

The Implicit Knowledge of the Individuals mainly includes technological elements, cognitive elements, experience elements, emotional elements and beliefs elements[⑤]. The researchers on the Business Synergetic Platform may have different background knowledge and cultural literacy. For example, some researchers may possess the professional knowledge in economics, while some researchers may have chemical expertise, the different professional backgrounds will affect the depth of the information acquisition. When the Implicit Knowledge spreads among the researchers, it may be more difficult to be understand because of the different majors.

4.2 The Subjective Obstacles

4.2.1 The Opportunity Cost Factor of the Sharing of the Implicit Knowledge

Opportunity cost means gain the greatest value by giving up something in order to get another thing. Any use of the scarce resources always forms the opportunity cost, no matter whether be paid during the process of the actual use. It's necessary for a person who want to get other's Implicit Knowledge to expend certain time even material and financial resources. As the provider of the Implicit Knowledge, its sharing also inevitable takes his

978-1-935068-76-1 © 2011 SciRes.

time and energy. So the consideration based on the opportunity cost, will bring the obstacles of the sharing of the Implicit Knowledge.

4.2.2 Mutual Trust Degree of the Implicit Knowledge's Sharing Individuals

Trust is the foundation of the knowledge sharing. If the mutual trust degree of the Implicit Knowledge's sharing individuals is very below, it will inevitably affect the cooperation level, including affect the sharing behavior of the Implicit Knowledge. Only under the precondition of the mutual trust, can the Implicit Knowledge be well spread and shared. On the Business Synergetic Platform, the researchers may come from the same or different institutes. If some researchers come from the same institutes, they may feel familiar and the trust degree may be higher. To the contrary, if some researchers come from the different institutes, they may feel strange and can't trust each other very much. This kind of situation especially prone to appear during the startup period of the Business Synergetic Platform.

4.2.3 Incentive Mechanism of the Business Synergetic Platform

Based on the consider of the egocentric factor, the Implicit knowledge owners will have the tendency of protecting the Implicit Knowledge. Effective incentive mechanism can push researchers produce the inner motivation and improve their enthusiasm and initiative of sharing the Implicit Knowledge. If the incentive mechanism lacks, some of the owners of the Implicit Knowledge will not contribute their Implicit Knowledge voluntarily, thus the sharing of the Implicit Knowledge will be hindered.

5 Effective Ways to Share the Implicit Knowledge on the Business Synergetic Platform

5.1 Exploit the Researchers' Communication Channels

5.1.1 Use Brain Storm, and Excavate the Implicit Knowledge

The characteristic of the brain storm is that the participants can open their minds, and all kinds of minds arouse creative brain storm in mutual collisions. Every project team of the Business Synergetic Platform can make full use of the favorable opportunity such as the topic selecting meetings, project bidding meetings, project argument meetings, and the task evaluation and report conferences, ect, use brain storm to let researchers answer the questions and mutually edify, and reveal the Implicit Knowledge in the process of talk and discussion.

5.1.2 Use Network Technology and Construct the Sharing Atmosphere

Under the network environment, the Implicit Knowledge's face-to-face spreading chance is less and less. But it can be fulfilled by making full use of the network technology. On the Business Synergetic Platform, the internal employees, external members and special experts can release their Explicit Knowledge(such as researcher's research achievement) by establishing academic blog, besides, they can share their feeling and experience, thinking and judgment etc in the research processes. The frequent interaction can also promote the mutual trust among the researchers, so that the sharing effect can be well raised. In addition, they can exchange and communicate their feeling and experience with others by establish BBS or install an instant messaging tools, and they can explicit the Implicit Knowledge by the visual tools such as VISIO, Project and Conceptual Map and so on. Besides, as the multimedia data such as video data lack rich semantic content annotations, we can use some knowledge management framework for conference video-recording retrieval that aims to bridge the gap between the Implicit Knowledge and the explicit Knowledge[⑥].

5.2 Establish Effective Long-term Incentive Mechanism

On the Business Synergetic Platform, one of the measures to encourage researchers share the Implicit Knowledge is to establish effective long-term incentive mechanism. Here are two main methods, one is assessment and rewards on the researchers for their behavior of sharing the Implicit Knowledge, arousing their sharing motive, mobilizing their sharing enthusiasm and initiative, so as to promote the communication and sharing of the Implicit Knowledge. The other is assessment on the whole job performance of the project team, which is linking up with the payment of every researcher, so as to link up every researcher's personal benefit with the whole benefit and promote the sharing of the Implicit Knowledge[⑦].

5.3 Construct the Cognitive Map and Establish the Database of the Implicit Knowledge

The cognitive map is a sharing tool of the Implicit Knowledge, which forms the sharing mental model through combining personal knowledge map, and thereby realize the sharing of the Implicit Knowledge[⑧]. The Business Synergetic Platform should construct every researcher's cognitive map at first, then combine all of theirs, and on this basis construct the whole cognitive map of the Business Synergetic Platform.

Establish the database of the Implicit Knowledge base is also an effective way to realize the sharing of the Implicit Knowledge on the Business Synergetic Platform. The relevant person can collect enough researchers' achievement reports and experience and so on, then summarize, generalize and establish the database of the Implicit Knowledge. The following is to manage the Implicit

Knowledge and realize its search function, so as to better share the Implicit Knowledge.

6 Conclusion

The spread of Implicit Knowledge meets many blocks of objective and subjective factors. We must make the best use of the circumstance, make full use of the modern information technology, combine with the scientific and effective management measures, accelerate the sharing process of the Implicit Knowledge and ensure its sharing effect. Only in this way can we improve the core competitive ability of the organization and the level of information service. The preparation of the Business Synergetic Platform of science & technology information research of Shandong province provides a possibility of theory and practice to change the real situation of research activities' fragmentation which is caused by the restriction of industries and regions, and the research on the full sharing of the Implicit Knowledge of the researchers on the platform can really raise the level of project team's scientific research, therefore we can provide knowledge service of higher quality to government and society. In Web2.0 era, to meet the need of the society perfectly, the researchers on the platform must master and use new techniques expertly, manage personal Implicit Knowledge effectively, pay attention to the exchange and cooperation with others, enrich and perfect personal knowledge structure.

References

[1] Huiling Wang, Zhuzhu Han, "the analysis of the ways to make the Implicit Knowledge explicit in knowledge management", Science & Technology Management Research, 2009(1), pp. 212-214.

[2] Xinmiao Li, "the research of the knowledge service of new products' development team in the network environment", Lirbary Research, 2009(7), pp. 70-72.

[3] Roy Lubit, "Implicit Knowledge and knowledge management: keys to sustainable competitive advantage", Organizational Dynamics, 2001(4), pp. 164-178.

[4] Kaiping Geng，Yu Xu, "the research of the spread of Implicit Knowledge in the complex organization Existing members flow", Journal of Information, 2009(5), pp. 115-117.

[5] Ping Wang, Qiang Wang, Lixia Zhou, "the explicit of the Implicit Knowledge of the subject librarian in the academic library", Information Science, 2010(8).

[6] Sokhn,M. Mugelini, E.and Khaled, O.A. "Knowledge management framework for conference video-recording retrieval" Proceedings of the 21st International Conference on Software Engineering and Knowledge Engineering, SEKE 2009 , pp. 709-714.

[7] Yafang Zhong, "the discussion about the sharing of the Implicit Knowledge in the project team", Enterprise Management, 2010(3), pp. 70-71.

[8] Hongyan Zhu, "the management of Implicit Knowledge in the library, based on cognitive map", Journal of Library and Information Science in Agriculture, 2010(10), pp.284-287.

978-1-935068-76-1

Clean Energy Technology and Evolution of Intellectual Property System

Ying FENG
Institute of Scientific and Technical Information of China (ISTIC) Beijing, 100038

Abstract: This paper takes the transformation of clean energy technology and its great impact on traditional Intellectual Property System. Then it analyzes the relationship between public departments and clean energy technology and reaches to the existing intellectual property system hampering the application and flow of clean energy technology to some extent. In conclusion, clean energy technology putting forward new demands for future intellectual property system.

Keywords: Clean Energy Technology; Intellectual Property System; renewable energy

The recent years have witnessed the significant adjustment in global energy structure, and meanwhile the development of clean energy has been unprecedentedly lifted to a strategic level that is closely related to the future development of many countries. On January 27, 2010, Obama stressed in the first State of the Union that "the nation that leads the clean energy economy will be the nation that leads the global economy. And America must be that nation" [1]. On April 27, 2009, Obama pointed out again that "the nation that leads the world in 21st century clean energy will be the nation that leads in the 21st century global economy" [2], when he delivered a speech at the American Academy of Science.

1 Clean Energy Technology Posing Challenges to The Existing Intellectual Property System

As the second largest county of energy consumption and greenhouse gas emission in the world, clean energy has turned to be one important topic of discussion in politics, economy, foreign affairs and science and technology of China. Clean energy, climate change and the like are generalized as "new public problems" in this paper. The substantive progress in developing clean energy and combating climate change hinges on whether the impact brought about by such kind new problems will shake some orders and thought patterns of the modern society and further promote more profound changes. These "new public problems" are of public nature, not belonging to the individual category any more, thus requiring a holistic thinking, in other words, a system thinking capable of taking each individual into consideration.

Under the traditional economic development pattern, intellectual property aims to stir up individual's innovation initiative and to voluntarily increase the benefit of the whole society on the condition that personal benefit is obtained at the same time. In this sense, the stimulation of individual's innovation initiative is bound to be a fundamental feature of the intellectual property system under the traditional economic development pattern. However, when the subject of economic development becomes clean energy development directing at dealing with global climate change, individual's innovation has changed to a means of realizing both social and personal benefits. Compared with the traditional material wealth growth pattern, this pattern is not a process of creating "private goods", but a process of creating both "private goods"(material wealth) and "public goods"(ecological wealth). Owing to having different properties, the increase of private goods and public goods is in a direct proportion within the load range of public goods while an inverse proportion beyond that range. The ecological wealth, for its nature of public goods, should be produced on state-oriented basis instead of individual's initiative. This is because that the intellectual property that takes "dematerialization" as an objective should not be transformed into individual's economic benefit through individualized contract, and the effect generated by "dematerialization" technology should be an ecological effect having a nature of public goods. The public goods effect renders a failure in the transformation of benefit through market contract. That is to say, the technological innovation of dematerialization is hard to find individualized market subject in request of this technology through market. Therefore, personal benefit under the intellectual property system aiming to dealing with global climate change is better to be realized on the premise that the whole ecological benefit is realized first through participation in the supply of state-oriented environmental public goods, but the contract and agreement under market mechanism.

As a result, the conversion process of individual economic value under individual intellectual property system handling global climate change must be a process where the environmental benefit of the whole society is realized at the same time. This requires the intellectual property system, on the one hand, should maintain the individual nature of intellectual property rights (IPR) so as to stimulate the individual to develop innovations that are capable of providing technology support for the

state's effort in overcoming climate change; on the other hand, the intellectual property system is still responsible for the value goal of the common realization of private benefit and environmental benefit of the whole society.

In fact, the current intellectual property practice has embodied such a trend and requirement. For example, "order-based" government procurement of technology that is directed at coping with climate change and the scope and condition of enforced implementation of intellectual property become more expansive and universal; overseas technology transfer and exclusive license must be submitted to related departments and be approved; a large number of state-owned specialized technological innovation corporations are emerging as the main part of intellectual property. However, under the economic growth development pattern aiming to coping with global climate change, a single market is incapable of playing a part in benefit adjustment for the ecological wealth is of a nature of public goods, thus requiring the state to create innovation needs on behalf of the whole society. This creation process is a process of public selection but market selection. The state expresses the overall needs for suitable climate in the name of the whole society is not only an inherent requirement for combating climate change but also a new type of national duty in modern society, wherein the most concentrative and representative reflection is reduction goal or target of greenhouse emission at different levels and the plan targets of state's developing and using clean energy.

2 Relationship between Public Departments and Clean Energy Technology

2.1 Aid from Public Department Playing a Significant Role in the Development and Research of Clean Energy Development

Based on the Derwent database of Thomson Reuters, the paper conducted a search on the patents published from 1990 to 2009. The searching results are analyzed and recorded according to the classification of clean technology in "Global Gaps in Clean Energy Research, Development, and Demonstration" published in December of 2009 by International Energy Agency (IEA), which involves more than one hundred thousands of patents, wherein patent documents in solar energy, advanced vehicles, construction, industrial energy saving and other related technological fields account for 76% of the total amount.

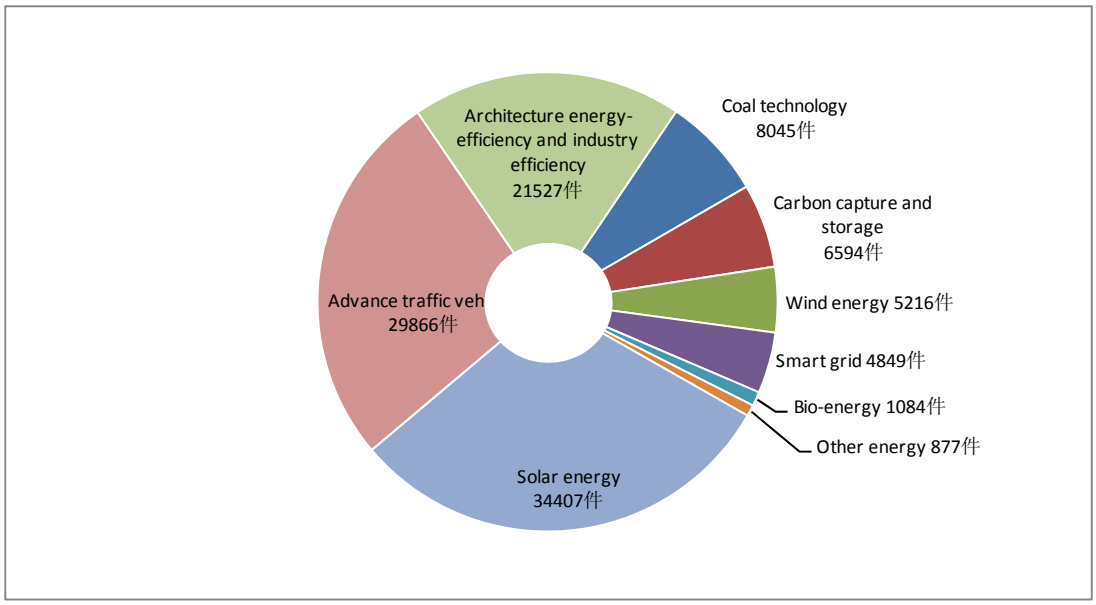


Figure 1. Distribution of Low-carbon Patent Technology

The analysis shows that some renewable energy technologies, particularly photovoltaic technology, are still comparatively expensive in usual application. Consequently, enterprises will be irresolute when voluntarily carrying out important researches on this field, thereby shifting their attention to such fields where a large amount of subsidy is provided by the government for sake of cost saving, for instance, the popular ethanol industry in America. As a major funding agency for clean energy technology in the US, the Department of Energy offered at least 356 million dollars in the budget proposal of 2008 for three kinds of clean energy technologies: solar photovoltaic, wind energy and biomass energy. In European Union, the researches conducted by public departments take up more than 50% of the overall researches and in 2002 the investment from public depart-

978-1-935068-76-1 © 2011 SciRes.

ments was 349.3 million euros while from private sectors was 340 million euros.

2.2 The Existing Intellectual Property System Hampering the Application and Flow of Clean Energy Technology to Some Extent

Clean energy offers a hope for solving the future energy problem of human beings as well as a significant means of coping with climate change. United Nations Framework Convention on Climate Change (UNFCCC) calls on developed countries to speed up related technology transfer to developing countries. However, a problem arises when developing countries acquire these technologies from developed countries, that is, the problem of intellectual property. In view of this, International Centre for Trade and Sustainable Development (ICTSD) had entrusted the ex-chairman of Committee on Intellectual Property Rights (CIPR), Dr. John H. Barton, an emeritus professor at Stanford Law School, to conduct special research on this problem. And soon after a research report titled "Intellectual Property and Access to Clean Energy Technologies in Developing Countries: An Analysis of Solar Photovoltaic, Biofuel and Wind Technologies" was published, which primarily probed into the problem that whether the IPR is an obstacle for technology transfer in fields of solar photovoltaic, wind energy and biomass energy; meanwhile, it also investigated the industrial structure of these three technological fields. China, India and Brazil and other developing countries where technology is relatively advanced became the major focuses.

Table 1. Intellectual Property Implications: PV, Biofuel and Wind (By John H. Barton)

TECHNOLOGY	PV	BIOFEUL	WIND
IP access limitations on current market for energy (For reducing emissions or participating in CDM)	Few concerns over IP	Essentially no concerns over IP	Possible concerns over IP, but likely to involve at most a small royalty
Major developing country concerns in future market for energy	Possible difficulties in obtaining advanced IP-protected technologies	Possible barriers or delays in obtaining cellulosic technologies	Possible risk of anticompetitive Behavior given concentration of industry
IP access limitations on entering the industry as a producer of key components or products	Possible barriers or delays in obtaining or creating the highest quality production systems	Possible concerns over access to new enzymes and conversion organisms – but at most a royalty issue	Possible difficulty in obtaining most advanced technologies
Most important overall concerns in area	Access to government-funded technologies, standards	Global trade barriers in the sugar/ethanol/fuel context. Access to government-funded technologies, standards	Access to government-funded technologies, Plausible anticompetitive behaviors, Standards

As seen in the table 1, the relation between IPR and environmental protection will be tense if IPR is overused in one country or by a patentee, for intellectual property imparts patentee the exclusive right of using patented technologies within a specified time.

Here goes a living example: in the case of Canon Inc. vs Recycle Assist Co., Ltd, the plaintiff, Canon Corp., is the owner of a patent right over a technology on printer toner cartridge. A company, after recycled the used toner cartridges in China, cleaned the cartridges by punching holes thereon and then refill toner therein for reutilization. The defendant, Recycle Assist Ltd., bought these refilled toner cartridges and imported them into Japan for selling. When the case was heard, the defendant especially emphasized that selling these refilled toner cartridges was a recycling of products, which did good to resources conservation and environmental protection. The Supreme Court of Japan, however, still adjudicated the behavior of the defendant infringed the patent right of the plaintiff, although it is true that environmental protection is in favor of the civilized society development and the welfare improvement of the whole human. Because the plaintiff did not provide consumers with services of reusing the toner cartridges, these cartridges had to be abandoned after use; consequently, it is inevitable to result in a waste in intellectual property protection and environment pollution.

From the forgoing, it can be seen that not only the developing countries but also developed countries are suffering from the existing intellectual property system.

3 Clean Energy Technology Putting forward New Demands for Future Intellectual Property System

The charm of knowledge and idea lies in, among others, the non-competitiveness and non-exclusiveness, which makes them distinguishing from most materials. However, the existing intellectual property system takes knowledge as the competitive and exclusive resource, for example, if I apply for patent right for one idea, others are banned to adopt the idea unless they can afford the monopoly price of this patent. In fact, there are better methods for stimulating invention and creation, so it is unnecessary to commercialize knowledge and monopolize it like this. A successful example that can be listed here is the open source movement in IT industry and free

software. The open source movement has millions of followers who voluntarily contribute their valuable time but ask for nothing in return. Technologies created in this movement, including Linux and Open Office, are so exciting that they offer a low-cost or even zero-cost consumption options for the worldwide users as well as feasible alternative products for those who are unwilling or unable to afford the expensive software sold by some monopoly enterprises like Microsoft. Free software movement abides by the principle of "Copy Left" that runs counter to the traditional copyright and intellectual property. This indicates that anybody enjoys the freedom of using, researching, duplicating and sharing related achievements, publishing and sharing the amended, or derivative achievements with others; of course, all these derivative achievements should be published and shared under the principle which is identical or corresponding with that of "Copy Left".

The open source movement in IT industry is advancing step by step. When clean energy sounding a new bugle around the world, the technological fields in energy saving and sustainable development also cry for an open source movement. However, it is a pity that the progress is very slow at present. And, the technology transfer system under UNFCCC has not urged developed countries to transfer at least one technology to developing countries up to now. Some transnational corporations established a World Business Council for Sustainable Development (WBCSD) including about 200 member corporations, which is specialized in handling problems in commerce and sustainable development. In the first half year of 2008, the council starts a patent share project on environment, appealing transnational corporations to donate environment-friendly patents to open to the public. Any company can participate in this project and obtain favorable reputation so long as it donates one patent voluntarily. So far, there has had seven members including IBM, Nokia, Bosch, Xerox, Dupont, Pitney Bowes and Sony. However, the patents contributed by these companies make no difference in important technological breakthrough or expansion of potential selling markets. In addition, during the period of Poznan negotiation held in December of 2008, a representative from the council stated that "there is no way for industrial circles" to accept any compulsory article on patent expansion contained in the UN-led climate agreement. From this, it can be seen that what the council really want is the technology transfer only happens under the framework where those transnational companies enjoy the membership and occupy the dominating place. All the behaviors let people cannot help but wonder the original intention of such paten share project on environment as set up by the council is not meant to boost sustainable development but to publicize those transnational corporations with boastful words and impractical behaviors. What is worse, those transnational corporations are defending themselves by attacking first and with preemptive strategies with a view to preventing the implementation of compulsory license over patents.

Actually, new trials on green patent technology licensing have been put into practice. In January of 2008, IBM, collaborated with WBCSD, founded "Eco-Patent Commons" together with many other companies, and for the first time opened dozens of innovative environmental protection patents for public use. "Eco-Patent Commons" is directed at promoting the application, implementation and subsequent development of environmental protection technologies, which provides a platform convenient for technology sharing and encouraging cooperative utilization and development of environmental technological proposals. In a way, "Eco-Patent Commons" carries forward the idea of free transmission of "Creative Commons (CC)". "Creative Commons" is aimed at increasing the circulating accessibility of creative works, offering a basis for others' creation and sharing while "Eco-Patent Commons" supplies a significant and special green patent technology licensing and operating model. That is, "Eco-Patent Commons" supports the sustainable development with innovative technologies and solutions sharing so that the enterprises obtain distinctive leadership, and meanwhile offers an opportunity for enterprises and other entities to conclude common interests and establish new partnership in further developing patent technologies and other fields.

How to deal with the conflict between the clean energy technology and intellectual property system? Is it possible for the open source movement in clean energy field to come true? All these problems are challenging everyone who takes part in combating the environmental problem; no matter they are from the governmental department, business circle or non-governmental organizations. We are in the hope of a leader like Linus Torvalds (the founder of Linux) who could lead the open source movement, will also appear in the climate field one day and more environmental technologies to be open to the whole world. Clean energy development is both the opportunity and challenge in face of everyone, so joint efforts and cooperative creation of human beings should be made regarding the development in this field.

References

[1] President Obama's 2011 State of the Union Address.
[2] Obama's remarks at the national academy of sciences annual meeting. 2009. 4.27.

978-1-935068-76-1 © 2011 SciRes.

Construction Personalized Information Environment for Scientific Research Team

Ruixue ZHAO

Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing, China, 100081

Email: zhaorx@mail.caas.net.cn

Abstract: Scientific Research team is the most dynamic scientific and academic organization in the modern research institutions. The paper analyzed the development of information technology impact on scientific research activities and the new features of team building and management in the new era. Some general ideas and solutions for building a personalized information environment for scientific research team were proposed at last.

Keywords: Scientific research team; personalized information service; information environment

面向科研团队的个性化信息环境建设

赵瑞雪

中国农业科学院农业信息研究所，北京，中国，100081

Email: zhaorx@mail.caas.net.cn

摘　要：科研团队是目前科研机构中最具活力的科研学术组织。文章分析了信息技术的发展对科研活动的影响以及赋予科研团队的新特征，针对新时期科研团队建设与管理中的问题，提出了构建面向科研团队的个性化信息环境的总体思路和经验体会。

关键词：科研团队；个性化服务；信息环境

1 引言

科研团队一般是由学术带头人组织带领的针对具体的学科方向、由一些技能互补、愿意为共同目标承担责任的科研人员组成的群体，是目前科研机构中最具活力的科研学术组织[1]。

随着现代科学技术的快速发展，家庭作坊式的个人创造活动开始被群体创造活动所取代，团队合作越来越受到重视，并逐渐演变为普遍的科研活动形式，在科技创新中发挥越来越重要的作用。科研团队的建设情况以及团队的科研创新能力已经成为影响科研机构整体科研实力和科研成果产出的重要因素。

信息技术的迅速发展，深刻改变了传统的科研方式和科研组织活动，呈现出数字化、网络化、协同化的趋势。这些变化引发了科研团队的研究环境、交流方式和交流手段等方面的变化和对新型信息环境的需求。科研人员的科研活动更加依赖于网络化的数字资源，交互、协同的数字化网络化科研平台逐渐成为他们的工作场所。

面向科研团队的个性化、知识化科研信息环境建设正是基于上述需求，目的是为科研团队的科研创新活动和团队建设提供有效的支撑，也是专业图书馆和研究型图书馆深化服务、走入科研一线的重要内容。

2 网络环境下科研团队建设与管理需求

为了充分了解和把握网络环境下科研团队建设与管理的需求，掌握科研团队对新型科研信息环境的要求，作者针对中国农科院的重点学科团队进行了相关调研，对普遍的需求进行了总结：

（1）“一站式”网络化获取科研所需的分布式资源与服务仍然是普遍的需求。科研人员希望能在自己熟悉的科研环境中高效、低成本获取相关资源和服务，能够随时了解国内外相关领域动态、研究进展和同行信息，能够获取支持科研活动全过程的专业化服务。

（2）科研团队承担着艰巨复杂的科研任务，知识复杂性日益加深，迫切需要加强知识资产的管理，包括长期积累的文献资料以及科研产生的科研报告、科学数据、论文、项目资料等。这些知识资源作为科研团队最重要的智力资本，直接影响着团队的发展壮大和竞争能力。加强团队知识管理，不仅便于团队成员

之间的知识共享，更有利于个体隐性知识到显性知识的外化以及隐性知识的“群化”共享，促进知识的创新与知识应用。同时，可以避免因团队成员流动导致的资源流失，有助于团队知识的持续积累和整体实力的增强[1]。

（3）团队领导希望制定合理的知识评价制度及配套的评价系统，以便随时考核成员在知识管理中的贡献，有针对性地实施激励措施，鼓励团队成员知识交流，约束成员的不合理行为，实现个体目标和团队目标的协调发展，为科研团队积累财富。

（4）希望建立有效的科研交流与协作机制，提供适当的工具和环境。科研创新活动需要经过多个环节和过程，通畅的信息流通与反馈对于促进团队成员间的知识共享、加速科研成果的产出至关重要。

（5）团队更希望有展示自己科研能力的空间，便于扩大影响，吸引各方关注，增加广泛合作的机会。

总之，信息技术在科研活动各个环节的渗透正在引发科学研究的新革命，科研团队的建设与管理也面临许多新的特点，迫切需要一个新的、基于网络的数字信息环境支撑，将资源、组织、人员、服务等要素有机地联结成一个整体，提供集成化的服务。

3 国内外科研信息环境建设现状

目前，国内外对科研信息环境尚并没有一个统一的概念和定义，各国采用了一些不同的术语来开展科研信息环境的研究与实践。如 20 世纪 90 年代美国开始研究建设面向学科的协作研究体（Collaboratory），重点研究分布式计算与数据资源的获取、学科化的分析工具、共享的工作空间和合作交流机制。2003 年，美国国家科学基金会启动“先进信息化基础设施计划”(Advance Cyberinfrastructure Program，ACP)，明确提出要建设为科研与教育提供新的知识环境的整合基础设施。美国康奈尔大学开发的开源程序 VIVO，可以实现包括科研专家（People）、学术活动（Education and Training）、研究工具（Research Tools）、仪器设备（Facilities）、学术机构（Academic Units）等领域知识聚合和导航服务，并提供领域的知识内容创建与管理、知识关联检索、知识发现等功能。英国在 2004 年启动了虚拟科研环境（VRE，Virtual Research Environment）项目，VRE 旨在集成科研团队涉及的各方面科研信息，并发现领域内外支持科研活动的各种需求。澳大利亚 2005 年启动了 e-Research 建设项目等。在国内，北京邮电大学、香港科技大学、华中师范大学等也尝试基于开源虚拟学习软件 Sakai 探索虚拟科研环境的搭建，目前还主要集中在课程管理和兴趣小组间的知识共享。中国科学院国家科学图书馆基于康奈尔大学的领域科研信息环境工具 VIVO 构建了专业领域知识环境，基于 Portal 和 CMS 内容管理软件建立了学科组个性化信息环境，面向领域内外的科研人员提供知识导航与研究合作支持。中科院网络中心协同工作环境研究中心研发了一套支持 e-Science 的虚拟科研平台（Duckling），为团队协作提供综合性资源共享和协同工作环境。

综合各国有关科研信息环境研究与建设实践，总结出科研信息环境建设的核心要素：

（1）需求和应用环境

用户的需求和应用环境是科研信息环境建设的前提。主要包括：专题信息获取，相关机构查询，用户认证和对资源的授权访问，使用恰当的协议下载资料，采用合适的工具进行交流互动，开展研究工作，撰写论文以及内容发布等。

（2）通用服务

服务是为用户提供多种信息获取的途径和方式。可以分为如下几类：e-Collaboration 支持服务：如日程、在线交流等；e-Research 应用服务：如成果发布、任务管理等；e-Learning 支持服务：资源列表、学术资源管理等；数字化信息服务：内容管理、目录导航、信息资源发现等；一般服务：如用户管理、用户认证等。

（3）资源

资源是科研信息环境建设的信息来源，是实现服务价值的前提。科研信息环境的建设要根据不同用户的研究方向和信息需求特点组织可用的资源，分析这些资源的分布规律、获取要求、使用限制和技术接口标准等。

（4）架构/用户接口

架构是服务实现的技术框架，支持内容和服务的综合集成，如基于 SOA 的体系架构。标准化和互操作性是架构设计的关键因素。用户接口用以实现界面及功能的嵌入，建议采用 JSR-168（JAVA 的插件式的 UI 组件标准）和 web 服务（WSRP, WSDL, WS-I），架构可以实现对这些用户接口的“即插即用”。

（5）协议与标准

协议和标准用以实现分布式系统的开放整合，以便充分利用各种资源和服务。如：支持各个系统之间链接调度的协议 OpenURL、DOI、CrossRef 等；支持

978-1-935068-76-1 © 2011 SciRes.

数据交换和数据收割的协议和标准 Z39.50、METS、OAI-PMH、Dublin Core 等；支持信息聚合的标准协议 RSS 、ATOM 等；支持资源和服务发现的协议和标准 SRW/ SRU、WS-Discovery、SDLIP 等；支持资源和服务注册、描述和传输的协议 UDDI、SOAP、WSDL 等[2]。

4 面向科研团队的个性化信息环境建设思路

“一流的科研成果需要一流的信息服务”。中国农科院国家农业图书馆为了做好全院科研创新的信息保障工作，开展了面向全院科技创新基地、专业研究所、科研创新团队的个性化、学科化知识服务工作。“面向科研团队的个性化信息环境建设”是该项工作的重要内容，目的是为中国农科院重点科研创新团队创建一个能够发布、管理、保存团队内外学术资源、实现团队内部资源共享与交流、对外宣传与展示的集成信息环境。科研团队个性化信息环境将充分整合相关学科领域的资源和服务，为具体科研课题的开展提供各个阶段全程的信息服务，强调知识的关联性和隐性知识的挖掘和保存。建设框架如图 1 所示：

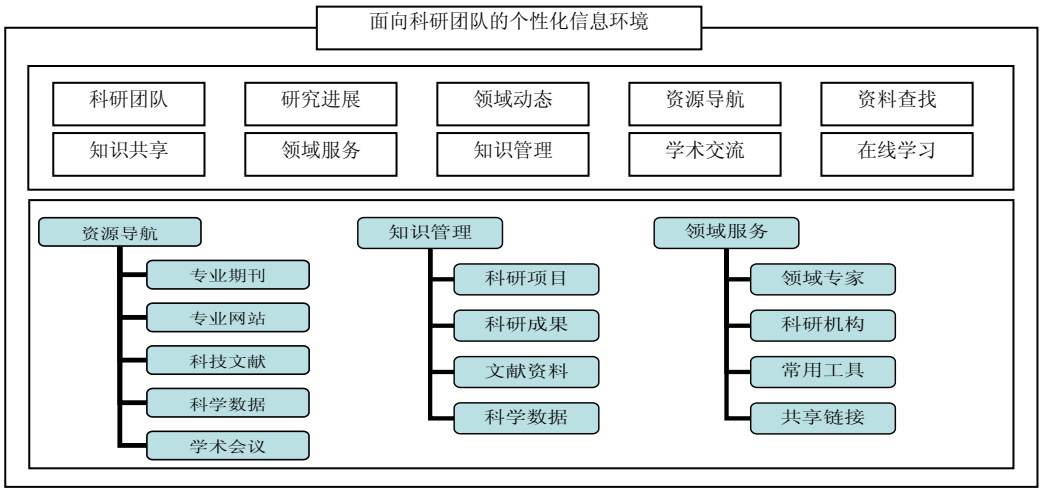


Figure 1. Construction frame of research environment for Scientific Research Team

图 1. 科研团队科研信息环境建设框架

（1）科研团队：主要提供团队成员的介绍和宣传。

（2）研究进展：用以宣传展示团队自身研究项目、研究思想、阶段成果等信息。

（3）领域动态：基于学科馆员的知识，结合网络科技监测和信息采集技术提供相关学科国内外最新动态信息，便于科研人员了解科研趋势，把握科研方向。

（4）资源导航：面向用户实现相关学科资源（包括专业期刊、专业网站、OA 资源、专业数据库、专业会议以及项目、机构、人员、设备等）及集成揭示和导航服务。

（5）资料查找：提供“一站式”文献信息和科学数据的检索与获取服务。

（6）知识共享：实现团队知识资产（包括科技资料、科研项目、学术活动、获奖成果、鉴定成果、取得专利、论文与专著、研究报告信息）的关联检索与共享联服务，支持开放获取服务。

（7）领域服务：提供学科领域内的知识链接共享服务，包括领域专家、科研机构、常用工具以及共享链接。

（8）知识管理：知识管理是团队建设的核心，用以收集、组织、存储、管理团队内的科学数据、研究成果和其他资料，实现科研团队知识资产的保存和管理、访问等应用。

（9）学术交流：提供常规文档的在线阅读与编辑，实现团队项目文档的传阅与修改；支持团队成员就某一专题发起离线或实时交流，并支持交流内容的编辑与归档。支持团队组织网络会议和开放论坛。

（10）在线学习：支持学习资源的创建、组织、发布和管理，支持学习资源的交流与共享以及团队成员的在线培训。

5 建设体会

为了顺利完成面向科研团队的个性化信息环境的搭建和应用，确保这种服务模式的实践效果，国家农

业图书馆在实践中总结出如下几条关键成功因素。

5.1 用户的参与和配合是前提

满足用户的需求是科研团队个性化信息环境建设的直接目的。这就需要建设工作首先要得到用户的认同和积极配合，让用户对信息环境建设的目标、功能、结构和运行有较深入的了解，建设人员也需要充分了解用户的需求特点、使用习惯和使用环境，激发和引导用户的潜在需求，协调好用户与计算机系统之间的关系。科研信息环境的建设不同于一般的工程项目，不能成为一项“交钥匙工程”，用户必须以信息环境主要建设者的身份直接参与全过程，从需求的确定、信息的收集处理、信息环境的设计、系统开发测试、到后续的运行维护等。无论多么先进的服务理念和系统，如果脱离了用户的配合和用户自己的科研环境，将注定是失败的结局。由于图书馆服务的公益性特点，在寻求用户配合和参与方面往往会遇到很多困难，因此，必须做好充分的准备，用扎实的专业知识、先进的服务理念和务实的工作态度去赢得用户。

5.2 个性化资源的挖掘整理是核心

资源是服务的基础与核心。在知识爆炸的信息时代，在互联网高速发达的今天，图书馆要想博取用户的青睐，必须能够建设和汇集充足的信息资源，否则，服务就成了无米之炊。但是，由于经费的限制，任何图书馆都不能无限制购买昂贵的电子资源，必须充分利用各种技术手段，通过不断地发现、挖掘和积累，链接、整合、共享全球一切可以利用的资源与服务，按照特定的学科领域、科研项目、科研活动等不同视觉建立相应的知识关联体系。该项工作需要一个动态长期的积累过程，需要图书馆建设人员充分了解用户的科研活动特性，掌握各种资源获取知识和技能，了解用户的资源取向和相关学科领域资源分布状况，保证整合资源的适用性和完备性。

5.3 严格的“分析—设计—实施”是关键

科研团队信息环境建设是一项复杂的信息系统工程，应遵循信息系统建设的规律性。在进行技术设计和实施之前，首先要进行科学的规划和充分的调查、分析、论证，识别系统建设的关键成功因素，弄清楚要为用户解决那些问题，满足用户哪些具体的信息需求，同时还要弄清楚用户的哪些需求不能解决，因为计算机不是万能的，即先解决“做什么”的问题，然后才能进入设计和实施阶段，解决“怎么做”的问题。纵观很多信息系统建设项目，由于建设人员急于进入技术实现阶段，分析工作粗糙不彻底、随意性大，导致用户需求得不到全面、准确的反映，系统建成后遗留问题多，从而使项目返工、周期延长，造成人力、物力甚至资金的浪费，更严重的是会失去用户的信任，不利于今后服务工作的开展。因此，图书馆用户服务系统的建设要严格遵循系统建设的客观规律，有步骤、有方法地进行。

5.4 流程化的业务协同是保障

科研信息环境的建设不是单纯的技术问题，高效运转的平台除了技术保障外必须有丰富的资源、专业的服务和密集的劳动。是集用户的参与、领导的支持、资源的保障、服务的引领和技术的支撑等一体的综合成果，是多方协同工作的结晶。尤其是在现代网络信息环境下，传统以工作任务划分职能部门的图书馆组织形式已经不能适应现代信息服务的要求，任何单部门作战都难以快速响应用户深度、综合的服务需求，必须建立以“用户为中心”的一体化服务体系及流程化的协同作业机制，完成从用户需求分析、到资源建设、技术平台搭建、直至服务实施的全程工作。明确流程中各环节之间的关系，协调好局部与整体的关系，树立整体最优、高效的全局理念。

总之，面向科研团队搭建数字化信息环境是一项深化图书馆个性化、学科化服务的有益尝试，是密切图书馆与一线科研人员联系、提升图书馆在数字信息环境下的服务能力、存在价值和发展空间的有效途径。

References (参考文献)

[1] Wu hongfa, Yang Chunxia,Zhanglinying. Construction of University Research Team in the Network Environment. Science and Technology Management Research[J], 2010(10):P73-75
伍宏发，杨春霞，张林英. 网络环境下高效科研团队建设的思考[J]. 科技管理研究，2010(10):73-75

[2] Zhang zhixiong, Linying, Guo shaoyou, Jiang qi. Towards the New Framework of the Institutional Information Environment. Digital Library[J], 2006(3):P2-5
张智雄. 新型机构信息环境的建设思路及框架[J]. 现代图书情报技术，2006(3):2-5

[3] Bai guangzu, Wu junsheng, Wuxinsheng. Preliminary Study on Research Information Environment. Information Science[J], 2009,27(4):P502-506
白光祖，吕俊生，吴新年. 科研个性化信息环境初探[J].情报科学，2009, 27(4):P502-506

[4] Wu yuewei, Song yibing, Li yinjie. Construction Mode and process on Personalized Information Environment for Research Users. Library and Information Service Online[J], 2010(1):P1-4
吴跃伟，宋亦兵，李印结. 面向科研用户的个性化信息环境建设模式与流程[J]. 图书情报工作网刊，2010(1):1-4

978-1-935068-76-1 © 2011 SciRes.

On Competitive Intelligence Functions Developing from S&T Archive Resources for a S&T Innovation-based Enterprise

Feng CHEN

Research Center for Information Science Methodology, Institute of Scientific and Technical Information of China, Beijing, China, 100038

Email: chenf@istic.ac.cn

Abstract: Base on several cases, the competitive intelligence functions developing from scientific and technical (S&T) archive resources for a S&T innovation- based enterprise are studied.Three major competitive intelligence functions developing from S&T archive resources are concluded: one is that it can help a S&T innovation-based enterprise knowing itself well, the second is that it can help a S&T innovation-based enterprise knowing other players especially its competitors well, the third is that it can help a S&T innovation-based enterprise knowing its competitive environment especially technology development trend in its feild well. The viewpoints are stressed out in the end that a S&T innovation-based enterprise must pay much attention to utilize outside S&T archive resources except inside and to utilize other information resources together except S&T archive resources only.

Keywords: competitive intelligence; scientific and technical archives; function; case; scientific and technical innovation; enterprise; information resources

1 Introduction

Scientific and technical (S&T) innovation- based enterprises are respectful. It is this group that act as major pioneers and propellers, that changed the endless magic power of science and technology into rich and colorful commodity to the world and helped the human being to realize their dreams. Among the respectful group, some enterprises became the examples of S&T innovation-based enterprises lasting more than several dozen years, example for GE, Siemens, Sony etc., some enterprises took off in recent years, example for Apple inc. But, glories are not the whole story for the group of S&T innovation- based enterprises. For any S&T innovation- based enterprise, they will meet stupendous difficulties and fierce competitions in all process of their development.

What are the functions of S&T archive resources for a S&T innovation- based enterprise? S&T archive resources have important function that it can conserve knowledge memory of S&T activities on the one hand. Because they are always related to the confidential intellectual assets of almost any enterprise, so S&T archive resources have competitive intelligence function obviously on the other hand.

In the academic research, some typical scholars in archive research field, example for Li Jufang [1], Liang Aizhen etc., have done some research work in relationship between S&T archive resources and S&T innovation for S&T innovation- based enterprises, but all of them only concluded that S&T archive resources have obvious promoting function for enterprises for S&T innovation and do nothing in further study on the competitive intelligence function about S&T archive resources for S&T innovation in enterprises. Some leading scholars in competitive intelligence field, example for Shen Guchao [2], Miao Qihao etc., have studied the methods that how to collect competitive information from reverse engineering, but do little in study on the competitive intelligence function about S&T archive resources for S&T innovation in enterprises. The study about the functions of S&T archive resources for S&T innovation- based enterprise is ignored obviously.

With three kinds of methods including cases study, documents research and interview with some competitive intelligence professionals in S&T innovation-based enterprises, the competitive intelligence functions about S&T archive resources for S&T innovation-based enterprises can be studied in more deeply.

Supported by National Nature Science Foundation of China (No.70973119)

2 S&T Archives Resources Can Help a S&T Innovation-Based Enterprise Knowing Itself Well

S &T assets are crucial for any S &T innovation-based enterprise. Strength of any S &T innovation- based enterprise results from a continuous progress that S &T assets are innovated, lineaged and accumulated by several offspring all the time. It is S&T archives resources that recorded the almost S&T activities of a enterprise and can provide complete first hand information to the successor. If a later people in any S &T innovation-based enterprise want to do any innovation work, a basic pre-

condition is that he must can answer well almost relevant questions, example for: what is the current situation of S&T activities in my enterprise? What are flaws for them tc., For answering the questions well, S&T archives resources is most valuable information resource. In many cases, developing S&T archives resources is indispensable. In other words, any people want to do any innovation work in any S &T innovation- based enterprise, he must know itself well. S&T archives resources are the irreplaceable assets for he can do further effective innovation stood on the shoulders of former and can avoid doing useless S&T activities. A story about a turbine manufacturing company illustrated the function that S&T archives resources can help a S&T innovation- based enterprise knowing itself well.

A shocking earthquake that destroyed nearly everything happened in 12 May, 2008 in Wenchuan area in Sichuan province of China. A rescue team with 451 people from Public Security Fire Department of Zhejiang Province arrived in earthquake-stricken area at the first time. As soon as the rescue team arrived in Mianzu City where the headquarter of Dongfang Turbine Co., Ltd (abbreviation is Dongqi in Chinese) locates, they received a urgent cry for help from Dongqi: rescue our S&T archives! The rescue team rescued all S&T archives after going through innumerable trials and tribulations and the leaders of Dongqi expressed their a thousand gratitudes shed tears.The leaders of Dongqi told the extremely importance of S&T archives for Dongqi. "You rescued our Dongqi! The S&T archives is our lifeblood, is our family estate accumulated by several offspring also.We can recovered our strength depended on S&T archives in short time. If we lost the S&T archives, Dongqi and even the whole technical competence in turbine manufacture industry of China will go backwards two years at least…" [3]. S&T archives resources can help a S&T innovation- based enterprise knowing itself well, this case gave a strong prove.

3 S&T Archives Resources Can Help a S&T Innovation-Based Enterprise Knowing Other Players Especially Its Competitors Well

Any people in any S &T innovation- based enterprise want to do any innovation work, it is unenough to know his enterprise itself well. Another basic precondition is that he must know other players especially his competitors well. S&T archives can become a resultful information resource for knowing much information about target players, example for the current situation of S&T activities, technical data, key technologies, know-hows, technical competence, R&D direction etc., then they can do effective innovation work that surpasses other players especially its competitors.

The S&T archives resources that can be helpful in knowing other players especially its competitors include outside public S&T archives especially S&T archives resources in governmental departments. Many scholars in competitive intelligence field, example for Zeng Zhonglu [4], Cai Jianwen [5], Ian Gordon etc., disclosed many cases that some S&T innovation- based enterprises defeat their competitors by getting key technical intelligence though public S&T archives resources in governmental departments.

A case about a lighter manufacturer illustrated the function rightly that S&T archives resources can help a S&T innovation- based enterprise knowing other players especially his competitors well.

Company A that headquarter locates in Wenzhou City of Zhejiang Province in China and founded in 1979, is a cigarette lighter manufacturer. From a little handcraft workshop growing to be one of the largest manufacturer of cigarette lighter in the world. There are many key success factors that promoted Company A creating business miracles continually since 1979 naturally, but being good at developing S&T archives resources about lighter products is an indispensable key success factor. From the start to become a top lighter manufacturer in the world, Company A has being done a work long that construct S&T archives resources about lighter products in the word. They collected almost all lighter product samples about other lighter manufacturers around the world, then got all almost technical intelligence, established technical files to join their S&T archives libraries. As soon as they see a new lighter product in the market, the will collect the lighter product sample and get all technical intelligence at the first time, then they will develop a more competitive new lighter product in short time, so Company A can get competitive advantage comparing to other players especially his competitors from 1979 to mow. This case illustrated the competitive intelligence function rightly that S&T archives resources can help a S&T innovation- based enterprise knowing other players especially his competitors well.

4 S&T Archives Resources Can Help a S&T Innovation-Based Enterprise Knowing Competitive Environment Especially Technology Development Trend Well

Change fulnesses outside an enterprise have strong impact on effectiveness for its S&T innovation. Lots of crossroads, risks, pitfalls and so on existing in the process of innovation will frustrate all efforts in S&T innovation for a enterprise. Any S &T innovation- based enterprise that wants to do any effective innovation work must know outside competitive environment especially technology development trend well, example for technology development direction, emerging technologies especially disruptive technologies in the future, changes of population structure, industrial characteristics, user need preference, important strategic plan by government

978-1-935068-76-1 © 2011 SciRes.

etc., For knowing outside competitive environment and technology development trend, S&T archives resources that include from both inside and outside especially governmental institutions can be provided with this function. A case about a S&T innovation- based enterprise in new energy industry illustrated the function well.

Company B, that headquarter locates in Shenzhen Special Economic Zone of China and founded in 1995, is a S&T innovation-based enterprise major in rechargeable batteries. The founder of Company B is Mr. W who had done R&D work in a institute of metallurgy industry major in new energy several years. Under the leadership of Mr. W, Company B has being created business miracles continually since 1995. Just like the key success factors in case of Company A, being good at developing S&T archives resources about new energy industry is an indispensable one. Mr. W had paid much attention to utilize S&T archives resources that include them both from inside and outside in his company, especially to utilize S&T archives resources from governmental management institutions and public R&D institutions at home and abroad. Company B has being done a work long that construct S&T archives resource about new energy industry especially in rechargeable batteries field.

Except utilizing S&T archives resource, Mr.W has being crazy about collecting competitive information from almost resources can be available, example for viewing exhibition and product shows, interviewing governmental officials and scientists, reading newspapers etc., Mr W possesses powerful competitive intelligence competence so he can know the changes of competitive environment especially technology development trend in the future well, and possesses insight for identifying opportunities in the future. With the powerful insight competence, Mr. W can make right strategic decisions also. Among the right strategic decisions, there are three typical decisions that demonstrated the relationship between competitive intelligence function of developing S&T archives resource and the right strategic decisions. The first was that he running a huge risk resigned his public service and founded the Company B. The second was that he invested a huge sum of money in taking the lead in developing Li-ion battery new product. The third was that he invested a huge sum of money in taking the lead in developing electric vehicle new product. The three right strategic decisions promoted competitive position of his enterprise greatly.Base on powerful competitive intelligence and uncommon insight, Mr.W can took the lead of other players identified the opportunities behind changes about the competitive environment especially technology development trend. This case ilustrated that S&T archives resource have the competitive intelligence function that can help a S&T innovation- based enterprise knowing competitive environment especially technology development trend well rightly.

5 Conclusions

"Knowing yourself well, knowing your competitors well, and knowing the changes of your competitive environment well"- this is just the essentials of competitive intelligence function and S&T archive resources can provide the functions for S&T innovation- based enterprises exactly.

When a S&T innovation- based enterprise do its innovation work by S&T archive resources, it is unenough to utilize its inside S&T archive resources, the enterprise must pay much attention to utilize outside S&T archives resources example for from governmental management institutions and public R&D institutions at home and abroad.

Furthermore, it is unenough to utilize S&T archive resources only, the enterprise must utilize other competitive intelligence resources at the same time in most cases.

References

[1] Li Jufang.On roles of S&T archive for enterprises in S&T innovation work. Archive About Mechanical, Electronic, Weaponry and Shipping Industry. 2009.2:12~13.

[2] Shen Guchao. Theory and Practice of Competitive Intelligence. Science Press. Beijing, China, 2008:251.

[3] Shen Weiguang,Li Xingxiang,Chenshuyuan and Zhouying. On-the- Spot Report about Rescuing Valuable Archives of Dongfang Turbine Co., Ltd by the rescue team wth 451people from Public Security Fire Department of Zhejiang Province. http://www.zgdazxw.com.cn/NewsView.asp?ID=2675. [2011-1-23].

[4] Zeng Zhonglu. Management of competitive intelligence in enterprises: the secret weapon to triumph over your competitors. Jinan University Press.Guangzhou,China,2004:60~66

[5] Cai Jianwen.To be King by Intelligence: Offense and Defense Ways in Business Intelligence.Golden City Press.Beijing, China, 2008: 25-28.

Scientific Research

Integration of the Subject Information Resources and the Management Mechanism Construction by an Academic Library

A Practice of the Library of Beijing Institute of Graphic Communication

Junling PENG, Xiongang ZHANG

Library of Beijing Institute of Graphic Communication, Beijing, China, 102600

Email: pengjunling@bigc.edu.cn, zhangxionggang@bigc.edu.cn

Abstract: This article introduces a practice that explored by the library of Beijing Institute of Graphic Communication to integrate the various discipline digitalized information resources and to construct the corresponding management mechanism for the academic institute. The technology implement mechanism and the work flow management mechanism are discussed. The work is summarized as investigating and archiving the various networking resources, and classification and integrating according their content, and navigating the emphasized disciplines digital resources, and the mechanism of co-construction and share of library with different departments is established.

Keywords: disciplines digital information resources; integration and management mechanism; emphasized disciplines navigating; library of Beijing Institute of Graphic Communication

1 Introduction

Along with the development of the Chinese higher education informationization, the campus networking construction is developing rapidly. The campus infrastructure with kilo mega- trunk networking and table terminal has come into being worthiness scale. At the present time, through the cable networking laying and the wireless networking layout, the campus networking can span the distance of space or time and extend the information across classroom or lab or library to every corner of the campus. Teachers and students can use the institute discipline information resources anytime and anywhere. However, on the other hand, with the different constructers and managers and other reasons, many characteristic and useful discipline and subject resources cannot be used very well. Therefore, the great task of the academic library nowadays is to apply the advanced digital library platform and integrate and manage the academic documentation resources of the institute, and improve the cognition and utilization effect of the discipline information resources

2 Research Compositions

By the support of the higher education research task of our institute, we library undertake the integration and management of the educational and researching digital resources of the institute. The concrete research composition includes: according to the various undertakers and managers of campus networking resources construction, we investigated and archived the corresponding discipline resources which come from library and teaching departments and teaching management units, and classified the resources according their contents , then constructed databases, institute characteristic resources, center/lab websites and study platforms, which sums up 4 big kinds and 6 small kinds and more than 300 resources aggregates.

Sponsored by the higher education research task item of Beijing Institute of Graphic Communication

3 Classification of the Institute Discipline Digital Resources

According to the different classification standards, the institute discipline digital resources can be classified with different types.

3.1 Classification by the Existence Forms of Institute Discipline Digital Resources

Academic digital resources can be divided into network subject digital resources and offline resources.

- Network subject digital resources—They are also called virtual subject resources. They are recorded in digital forms, expressed in multimedia formats, stored on a network computer magnetic media, optical media and a variety of communications media, and rely on computer networks for delivery of the information content of a collection of various disciplines. The institute network subjects resources are most extensive, most favorited by many people. Such resources are as network databases, coursewares, teaching websites, and their common features include: 1) stored digitally. 2) diversified forms, in addition to text, you can image other forms such as audio, video, software, databases and

978-1-935068-76-1

other means. 3) web media, reflecting the social network resources and sharing. 4) transmission modes are dynamic and real-time.

- Offline disciplines resources—as opposed to subjects network resources, they are independent of the network application environment, and these resources come from various subjects, which are not or have not been passed by means of computer networks, such as stand-alone versions of the coursewares and so on.

3.2 Classification ACCORDING to the function of Institute Subjects Digital Resources

According to the function of digital resources by academic division, the subject digital resources can be divided into such several types:

- Teaching Materials: teachers or professionals would screen through the materials that are suitable for teaching and research, such as images, sound, video, animation and other supporting materials, classify them and record them in digital form. These marerials include pictures, video, animation, audio, etc.
- Teaching Resources: Their main service functions are for teachers preparing lessons, teaching and research, and online communication. They are the accumulation of teaching resources, and physical existence forms. They are the formation of a school, an institue, and a discipline and the characteristics of individual teachers. They are the important components of teaching, including electronic lesson teaching plans, subject web pages.
- Learning Resources: network classroom and electronic resources databases and network libraries are included.

3.3 Classification According to the Content of Institute Subjects Digital Resources

Divided by institute subject content of the resources, they can be divided into such types: various types of commercial databases introduced from elsewhere, self-built special resources packages, laboratory and teaching sites, web courseware, as well as thematic resources.

4 Subjects Resources Metadata Indexing and Database Construction

The so-called metadata, refers to the provision of information resources or data on a structured data. It is the structured description of information resources . By describing the characteristics and attributes of the data itself, the organization digital information get provision. Metadata indexing is the foundation of re-foundation of information.

This research task aims to explore the integration and management of subjects resources of our institute, in fact, it is a sort of metadata indexing on the basis of investigating our institute's subjects resources. Through the certain digital library system platform, constructe the discipline GRID platform architecture underlying structure, complete data input, through the index reconstruction, providing a single condition, the combination of conditions and the data within the database to retrieve full-text search and many other ways, and provides search results on-demand sort out.

- Metadata design

Network meta-data of academic resources, the main factors includes: name, creator, manager, and other responsible persons, themes and key words, disciplines, description, URL, resource evaluation, etc.

No -network subject resources metadata, including the main factors are: name, managers and their subordinate departments and contact information, and other responsible persons and their subordinate departments and contact information, subject and keywords, subject expertise, resources evaluation etc.

- Database construction and publishing

With our library existing digital library management platform, we construct disciplines database resources by ourself, and then integrate and publish them. Readers can access the database through a browser, search through a variety of ways, directly or indirectly, access to relevant resources.

- Database Maintenance

By artificially checking back the database regularly to ensure the accuracy of the network subjects resources collected , and to add new resources in time.

5 Implementation of Key Disciplines Navigation

In the principle of simple and utility, we restruct the digital resources for 6 disciplins that own the right to grant master degrees. At the same time, we bring together the representative and influential professional sites information on the Internet, and constitute the basic framework of the navigation system of disciplines.

The six master disciplines are communication, materials physics and chemistry, signal and information processing, design arts, business management, mechanical and electronic engineering.

In order to facilitate readers to understand the situation of these disciplines, the navigation has in detail described each of the subject group of professional disciplines in Beijing's position, their research direction, personnel training location, the situation of teachers, teaching experiment, the basic information researching practice.

The composition of key disciplines navigation system modules includes such types:

- library resources: According to the characteristics of disciplines, we select the collections targeted Chinese and foreign language databases, and refer to readers' use of these databases in recent years, and in accordance with the utilization from high to low or-

der. These collections databases often use local mirror, mirror Education Network (or the use of educational networks of foreign database mirroring, or use of the education network dedicated access) and local Own Special Libraries (Featured Collection Library, features special library), access speed and response time are more ideal, is the use of the database of choice for the reader. Settings include the database name, brief description of access methods (link to my digital library portal in the relevant hospital presentation). Meanwhile, to broaden the reader to understand and use the vision of specialized databases, lists of trial related to the professional database, through its use of information and feedback from readers, whether for the purchase of the database, and provides our library basis for decision making.

- campus resources: reconstruct the Institute academic resources organized by the relevant functional departments of the lead organization, the characteristicly self-built teaching units Resource Kit (databases), the reaction of the latest research disciplines and commitment to the teaching practice of laboratory disciplines (website), in accordance with disciplines to information restructuring, and then provide one-stop link to guidelines databases, lists of trial related to the professional database, through its use of information and feedback from readers, whether for the purchase of the database, our library provides basis for decision making.
- Other resources: making the full use of the Internet open access platform, we bring the related Internet resources of the disciplines together, and provide cooresponding guidance. Including free access to the Chinese and foreign periodicals websites, Chinese and foreign professional websites that have representative subject areas of expertise and certain influence .

6 Conclusions

The project focuses on the practice and exploration of technical implementation mechanism and workflow management on the subjects digital resources. Through the investigation and the communication and cooperation with relevant institutions, we thoroughly get the discipline and teaching digital resources and collection system integrated into the library's Web site section, and set the establishment of the "integration of academic resources in campus," and established the work to update and maintain mechanisms on time.

We have analysized and researched the key disciplines information resources, for the six master's degree granting disciplines, we have reorganized the digital resources among campus. In the meantime, we have brought together the representative and influential professional Internet websites, making the formation of the basic framework of library key subjects navigation system.

Journal of the institute is like the window reflecting the teaching and scientific research of institute. After several communication and adjustments, we have established the cooperation relationship and the rapid response mechanism with " Journal of Beijing Institute of Graphic," with the sharing of the electronic version and timely content provision included in the latest issue . The Division of Social Science Edition and the Natural science version of the Journal content can be released online in a timely manner on the library web site.

Teachers usually published academic work publications mostly on paper, only through the library staff's manual collecting, there are often not complete accumulation. As far as possible, we are trying to search for each teacher's academic achievements, such as from Division of Personnel of institute to get teachers directory, then go to the National Library General Bibliography or CALIS (China Academic Library Information Resources Sharing Project) bibliographic database search of teachers books published, and then collected the books from book dealers. Now, the" Institute's Publication Library" owns the accumulation of academic achievements reached a more comprehensive scale. We collect paper, based on the library's digital network platform through the establishment of a "Institute's Publication Library" database, while the teachers completed monographs published in special database, fully searchable by CNKI. Papers published by the teachers can be searched by CNKI and form the integration of academic achivement and building a database, an important manifestation of this work is the collection in the retrospective searching the papers of teachers published in the past in Journal of Beijing Institute of Graphic Communication through CNKI.

We have established the mechanism of collecting graduate thesis electronic version of Dissertation with the graduate school department. The electronic versions of Dissertation are wholely collected by the graduate school department when the graduates leave the school. We library systemly collection the database of dissertations , and then library build on our own network platform and pay attention to the protection of intellectual property protection and safety technology.

Existing Problems: Some important special academic resources left in other Campus areas, such as Chinese Editoral Research Information Center and the Beijing Publishing Industry and Culture Research Base to build special characteristics database is in progress, so our library portal currently do not integrate their core resources, after many consultations, we look forward to the information center and research base will complete the construction of database resources and then come to be well known and shared of the resources.

References

[1] Zhan Yu.,Integration of Network Academic Reseouces in Universitites. [J]. Modern Information, 2008(9):101-106

[2] Ma Wenfeng,Du Xiaoyong. Trends of Digital Resources Intergration [J]. Library and Information Services, 2007(7):66-70.
[3] Xie Luqing. On the Intergration of Academic Library Digital Resources with the Campus Network Teaching Resources.[J].Sci-tech Information Development & Economy, 2007(3): 12-13

Scientific Research

The Construction of OA Platform with Integration of Multi-national OA Journals Resources and Its DRM Function

——Based on International Cooperation on the Next Generation of NSTL

Ying LI, Bing LIANG, Changqing YAO, Yunhua ZHAO, Xiaodong QIAO

Information Technology Support Center, Institute of Scientific and Technical Information of China, Beijing, China

Email: liying@istic.ac.cn

Abstract: Open Access and Digital Rights Management (hereinafter referred to as OA and DRM, respectively) are two key issues of information resource dissemination under digital environment. As a next generation of science and technology information resource dissemination system, it should be carefully designed for the purpose of achieving balanced development of such information resource dissemination while simultaneously implement its two functions, which are resource consumers-oriented OA and resource owners-oriented DRM. This is also one of challenging subjects that should be resolved by next generation of National Science and Technology Digital Library (hereinafter referred to as NSTL). On the basis of next generation system of NSTL and with fund support of Construction Project of NSTL, This study designs a service platform that enables to integrate multi-national OA journals resources, and introduces its DRM Function. Through conducting international cooperation with multi-national providers of OA resources and institutes related DRM, we try to probe into and implement OA resources integration service and intellectual property protection. This article firstly introduces the framework of OA platform in the next generation of NSTL. Secondly it focuses on its function of XML metadata exchange and update of multi-national OA journals. Thirdly it discloses rights information acquisition service, and it further gives conclusion and future implementation plan at the end.

Keywords: OA platform; architecture design; NSTL; IPR; DRM; J-STAGE; DOAJ; KISTI; SciELO; DovePress

1 Introduction

Under the environment of OA activities getting worldwide attention, we must carefully think about utterly new or improved function of OA when designing update of science and technology resource service system, aiming to meet users' requirement on resources access to the greater extent. However OA does not mean rights regarding resources may be ignored as imagined by resources owners. In fact DRM technology has to be selected if we decide to effectively implementing mechanism of protecting rights of such owners. Next generation of NSTL must also design a way to meet challenges coming from simultaneous implementation of two key functions of OA and digital rights protection in its system.

Though NSTL is mainly required to provide users of China mainland with net delivery service for foreign language science and technology documents, its current function of OA resources service just plays a weakened role due to less cooperation opportunities with OA resources suppliers in China or abroad. Meanwhile it basically does not have service function based on DRM technology. We, for improving service of OA resources, have particularly designed multi-national OA Journals resources integration platform through international cooperation with suppliers of multi-national OA resources. Meanwhile function of access to resources rights information is especially introduced to help rights management application system implement rights management on the basis of rights information acquired.

This article will introduce designs of OA platform structure in next generation of NSTL system around aforesaid target and focus on function of integration of its OA resources. Specifically this article will give details on function of exchange and update of XML metadata among multi-national OA resources as well as function of acquiring rights information of resources. Firstly, we will briefly talk about NSTL system.

2 Overview of NSTL[1]

One of basic strategies set by Chinese government is to ensure successful supply of science and technology documentary resources. Only the country has the ability to ensure supply of such documents under condition that subscription fee continues to increase, which also reflects the value of NSTL.

NSTL, a virtual organization approved by State Council, was founded on June 12 of 2000. Nine members units coming from science, industry, agriculture and medicine areas jointly provide science and technology information resource service system in the network environment.

978-1-935068-76-1

From10 years ago, NSTL has provided science institutes and researchers of China mainland with science and technology documentary information service and support national innovation.

NSTL is playing a great role in national strategic ensuring system of science and technology documents, including: support other science and technology information institutes to increase their service level, promote development of social science and technology information service, help users and other institutes increase ability of using such information by providing support to various public information services, adequately present supporting and radiation function of NSTL. NSTL has defined its development strategies as follows in the new era:

- Enhance providing and development of science and technology documentary resources for building a nation-level provision base of such resources.
- Improve service level of its resources, enhance resources and extensive service, as well as derivative service and open application, build a mechanism of open integration and fusion of various resources, support adequate using of such information, perform deep development and integrated service and manage to become service hub of national science and technology documentary information system.
- Extend and promote information services of co-development, sharing and joint cooperation of various resources, support building of other documentary service systems, advance research on application, conduct demonstration and public extension and manage to become support center of national science and technology documentary information service development.

3 Integration Design of OA Platform in NSTL System

NSTL's current network service system V3.0 was officially put into operation since on June 12, 2010, and its detailed service is shown in Fig. 1. Its so-called service functions of OA journals include: open journal integration revealing and retrieval system integrating two functions of journal browse and retrieval. System provides two browse modes of A-to-Z journal title list and subject category, which may be switched and prompted via general and detailed information of such journals for further understanding all information of one journal including 15 kinds of information of journal name, ISSN, subject, subject category, reveal level and etc. meanwhile user may conduct journal retrieval for journal name, ISSN, subject, publisher and all fields. Furthermore system has specially set interactive hotline column for making maintenance and update of system easy and provide better service to user, by which user may use New Resource Recommendation Function to recommend better OA journals or exchange and respond with system on time by function of Suggest and Advice. System will be integrated in one interface for helping better application.

Subjects of journals collected by this system cover 17 fields of agriculture, forestry, industry, commerce and medicine. Now system has collected nearly 5000 journals data, including 44 Chinese journals in English version and 130 other journals that will not be used until they are registered or applied.

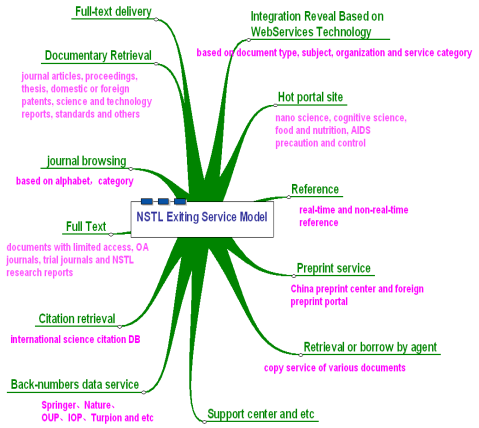


Figure 1. Services Provided by NSTL's Current System

As OA articles have become important document resources, OA resource service currently provided by NSTL needs to be integrated and enhanced. Therefore multi-national OA journal resource integration platform has particularly designed to help uniform retrieval and application of such resources via international cooperation with suppliers of multi-national OA resources (see fig.2). Specifically function of exchange and update of XML metadata among multi-national OA resources has uniformly retrieved documents of multi-national OA paper institutes in OA platform of NSTL.

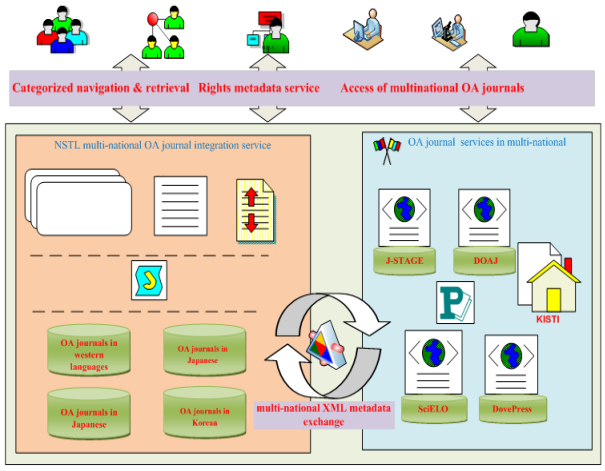


Figure 2. Framework of Next General NSTL Multi-national OA Journal Resources Integration Platform

4 Key Functions of OA Platform

Table 1. Key functions of OA journal resources service platform of next general nstl system

Function		Description
Multi-national XML Metadata Exchange		Automatically identify websites that provide metadata of electronic journals such as J-STAGE (Japan), DOAJ (Sweden), KISTI (South Korea), SciELO (Brazil) and Dove-Press (UK); automatically update documentary DB of this searching website.
Rights Metadata Service		Acquire rights information of resource based on description of rights metadata，or providing reference service of such information. Rights metadata principally may describe any complex rights information for managing intellectual property protection (IPR) in digital environment.
Page Generation Function by Different Categories		
1)	List of different journal documents	Page generation function of different journal documents list
2)	Detailed information of different Journal documents	Form Page generation function of different journal Documents
Document Retrieval Function		
1)	Document retrieval	Retrieve bibliography data (selective retrieval of journal name, ISSN, subject, publisher and all fields)
2)	Document list Check	For retrieving result, represents list of documents checked, CSV format download and printing functions.
3)	Detailed Information of Documents	presentation of linking to Full-Text information
4)	Full Text Link	Generating full-text linkage from Key information of full-text
New Resource Recommendation Functions		User may either use New Resource Recommendation Function to recommend high-quality OA journals to system, or interactive and feedback with system on time.
Other Functions		Log Management: acquiring retrieval queries of retrieval system.

4.1 Multi-national XML Metadata Exchange Function

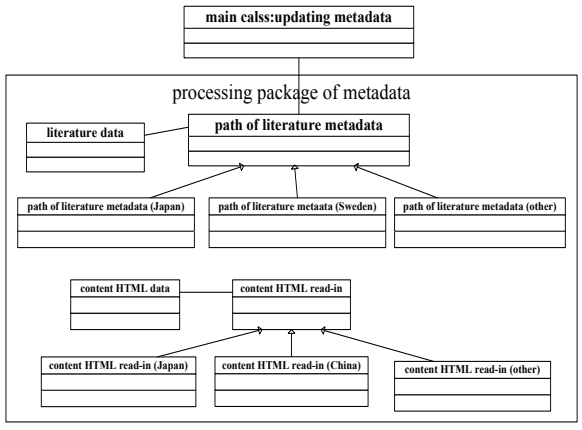


Figure 3. Diagram of several classed of Multi-national XML Metadata Exchange Function

The main goal of the OA Platform is to implement multi-national XML metadata exchange; the system function is acquisition of updating metadata. Acquisition of updating metadata is batch processing function to get information of OA journals site for retrieval. It automatically confirms OA electronic journals metadata provided on Japan, Sweden, South Korea and other countries' sites, once a day. If have new metadata, search site of OA platform provides bibliographic DB automatically updated. Class diagram of this function is shown in Fig. 3.

4.2 Introduction of DRM Function

Rights management is now facing great challenges as a result of freedom and virtuality of network publication of document resources. It has been clearly stipulated by China Copyright Law that oblige of work should own right of disseminating his/her work in the net. Regulation on the Protection of the Rights to Network Dissemination of Information beginning to be implemented since July 1 of 2006 used foreign and domestic experiences and was specially designed to deal with network dissemination. However these laws were confronted with some unavoidable questions, for example, difficulty of proving proprietorship, difficulty of proving infringement against property and high cost of collecting proof. All these questions had left negative effect on publication, transaction, circulation of these network digital resources and even some had generated actions of infringement and seriously affected national innovation efforts. It really made against healthy development of network publication. Therefore DRM technology should be used to resolve above problems.

We may understand that so-called rights means property information owned or endowed by resources. Generally rights description information includes intellectual property, copyright or other kinds of rights.

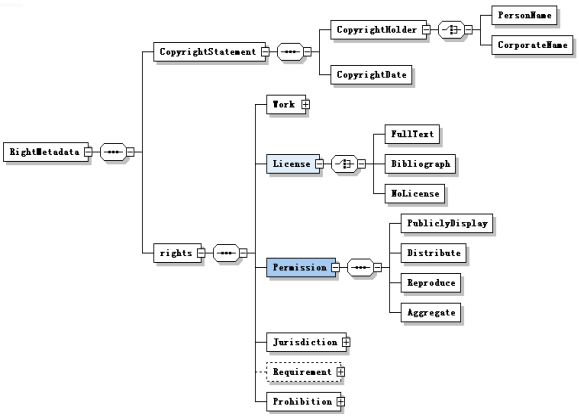


Figure 4. Rights Metadata Based on License Framework

Different document resource systems should identify their rights metadata system on the basis of their business requirement. Resource targeted by this study is those key science and technology journals under Ministry of Science and Technology (intended OA). The following two typical modes of providing resources have been studied when studying rights metadata design: Knowledge Sharing Mode (CC—Creative Commons) and E-com-

978-1-935068-76-1 © 2011 SciRes.

merce Mode. As building idea of elite science and technology journals achieved OA on the basis of copyright contract, this article principally selects design concept provided by Knowledge Sharing Mode in the experiment. As shown in Fig. 4, open level based on OA is classified into three access licenses for designing Schema used in rights metadata test.

- Full text OA
- Bibliography OA
- No license

We will henceforth design relatively complex rights metadata Schema for supporting development of application DRM system on the basis of actual property requirement of key science and technology journal OA and various content resource systems.

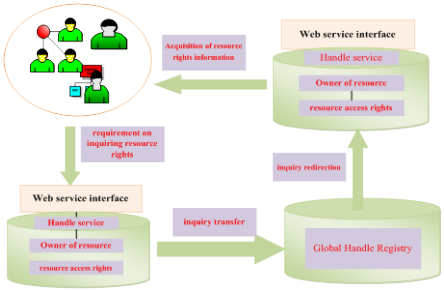


Figure 5. Process of Acquiring Resource rights Information

Resource rights information acquisition service is particularly developed on the basis of above rights metadata and its concept chart is shown in Fig. 5.

5 Conclusion: Future work

NSTL has preliminarily constructed its own open service system notwithstanding it is still far to an improved work. It must generally integrate current resources subscripted and OA resources in future steps for enhancing ability of retrieving and finding information and techniques. In a result we should focus on following works when designing, constructing multi-national OA resources integration platform and implementing DRM function:

- Test exchange of metadata with J-STAGE
- Extend exchange of metadata to systems of DOAJ, KISTI, SciELO and DovePress
- Adequately integrate Institutional Repository function with integrated supply of OA resources
- Implement DRM application system

References

[1] http://oainfres.caas.net.cn:8080/NSTL_OAJ/

[2] Guo Xiaofeng, Li Ying, and Sam X. Sun, “Federated Content Rights Management for Research and Academic Publications Using the Handle System,”http://www.dlib.org/dlib.html

Scientific Research

Study on Construction of Information Service Environmental Oriented Science and Technology Decision Support

Lixiao GENG, Feng WU, Yanzhong ZHANG, Hongyong GUO

[1]*School of Management, Hebei University of Technoloy Tianjin, China, 300401*

[2]*Hebei Institute of Science & Technoloy Information Shijiazhuang, China, 050021*

Email: lixgeng@yahoo.com.cn, guohongyong@263.net

Abstract: Nowadays the influence of science and technology on economy, politics, military, culture, education and ideology of the world and even on the whole society has great significance. The orientation, domain and level of the decision making of science and technology investment play important roles in the investment of the science and technology. The research object of this paper is services support environment of information oriented decision making of science and technology. The collection and processing of science and technology information of science and technology projects, organizations, staffs, high and new tech enterprises, experimental condition, literatures, patents, administrative organizations, intermediary services agencies, geographic data, meteorological data, planting resource and industrial market information, the analysis and induction of various of features of science and technology resource, the exploration of influencing factors of government for decision making of science and technology and the construction of the services support model and platform of science and technology information based on DSS enhance the quality, efficiency and level of decision making of science and technology, providing strong support and guarantee for scientific decision making of science and technology.

Keywords: information service environmental; science and technology decision support; decision support system; information service model

1 Introduction

Science and technology resources which supporting innovation activities of the whole society have became important fundamental conditions of the country. The optimization, restructuring and sharing of the fundamental conditions of science and technology resources becoming the government's task of development which has the most priority are very important for the promotion of the level of national scientific and technological research, the enhancement of competitive advantage of science and technology and the achievement of sustainable economic development of the country. At present, the large number of scientific and technological resources, are widely distributed. But science and technology decision-making requires a unified technology resource allocation. For this purpose, the paper attempts to establish decision-making information technology service environment.

2 Objectives of The Construction of Science and Technology Information Services Environment

The objectives of the construction of science and technology information services environment are the improvement of the environment for innovation, the enhancement capacity of sustainable development and providing strong support for the key breakthroughs of long-term development of science and technology by building service platform of science and technology resources which is public welfare, basic and strategic. It is designed to enhance the supply of scientific and technological innovation in public service, to optimize the quality of innovative resources and to build the platform of communication between government, enterprises and individuals, is an indispensable basic condition of scientific and technological innovation, belongs to the category of public goods of whole society and share common areas. The maximum benefit of informationization comes from the most widely shared information, the most efficient circulation and the deep-level mining meanwhile the information resources are the source of information. The information resources are only gradually standardized in the process of informationization, accessing to a wide range of consistency, to achieve the true sharing of information resources. In this regard, foreign countries has been advanced experienced on study of the standardization of information system, acceleration of the conversion rate of international standards combined with China's basic national conditions and giving priority to the development of informationization to meet the needs

Supported by Hebei Science & Technology Information Processing Laboratory Open issues, Database planning and construction of science & technology decision-making information

of basic data elements dictionary, information classification and coding standards, to improve the development and utilization of information resources is necessary.

3 Problems of Information Technology Resources Environment

3.1 "Silos" Formed by the Construction of Network Platform

The network environment provided by network platform which application of sharing of information resources to technology is required including hardware, software facilities is the basis of the construction of scientific and technological information services environment. The current information network platform has built a number of document information and integrated with science and technology information web site, provided a safe, stable and reliable network platform based on for the knowledge and dissemination of scientific information, scientific and technological information for the construction and use of resources and the environment. Since the era of new technologies and knowledge the limitation and inadequacy of emerging of the rapid increase of information, capacity of transmission channel, information processing, information service, and building and other information resources make the lag of the content of science and technology information in business organizations and information resources to deepen the process of Information resources and core business development become even more prominent so that the network independent individual to become "Silos". It is more and more difficult for users facing ocean of information to extract knowledge using to take effective action to make good decision-making.

3.2 The Deficiency of Capacity of Acquisition and Integration for Information

Information resources which are the premise of enhancement of value and capacity of service of the information through the process of acquisition, accumulation, processing, handling, storage are the subject of exchange of information resources running on the network platform. Since the construction of information resources related to technical standards and system norms, deficiency of technical standards for construction must make the information services environment be confronted with technical barriers so that information services environment cannot be integrated, resulting in that construction of network information resources and services is far from the height of the integrated information resources. It is difficult to provide support to meet the needs of development.

3.3 Issues of Operational Mechanism of Existing Information Service

Since the information resources lack of unified program, sorting, intersecting, collection and organizational structure can be used because of the fragmentation of Science and technology information service system, it has to build a channel to share the information resources among the information resource owners to protect intellectual property rights of stakeholders and interests of all on the share chain and the achievement of the use of information resources. At present, several obstacles of the use of information resources by service system and agreement still be left, the problem of distributed information resources and low efficiency are difficult to solve fundamentally.

4 The Mode and Content of Science and Technology Information Service

The collection and processing of science and technology information of science and technology projects, organizations, staffs, high and new tech enterprises, experimental condition, literatures, patents, administrative organizations, intermediary services agencies, geographic data, meteorological data, planting resource and industrial market information, the analysis and induction of various of features of science and technology resource, the exploration of influencing factors of government for decision making of science and technology and the construction of the services support model and platform of science and technology information based on DSS enhance the quality, efficiency and level of decision making of science and technology, providing strong support and guarantee for scientific decision making of science and technology.

4.1 Construction of Database about Scientific Data Resources

The Government spends a lot of money for scientific research and business investment, the industry has a certain degree of project funding. In order to build the communication platform among government, research institutes, engineering technical centers, universities, other institutions, high-tech enterprises and the park project, enhancement of the accuracy of efficiency of government investment, statistics of data revenue allocation and the establishment of scientific data resource database will be important. Since the declaration of Science and Technology Agency projects, and other links have been receiving information while the existing projects for collecting and collating, facilitating the exchange of project information and shared.

4.2 Construction of Research Archives Database

Base on the processing of the existing research staff and scientific experts' research archives, the advantages of current professional resources processing systems and teams would be brought into full play, the scope of the

research archives collecting and collating would be expanded. Unified planning, management and electronization of research files formed within the process of universities and research institutes' research and establishment of a complete database of scientific research files facilitate the government, enterprises, universities and research institutes to use scientific research produced.

4.3 Construction of Science and Technology Resources Database

Science and technology resource is an important part of science and technology information, including natural science and technology resources and technology literature resources. Natural science and technology resources include geographic data, meteorological data, agricultural planting resources, etc.

Scientific literature resources include scientific research management services organizations, technology intermediary service organizations and scientific and technological information research institutions. In order to further meet the requirements of scientific literature for technological innovation in the period of "the Twelfth Five-year Program", a lot of funds need to be invested each year for construction of literature resources, the range of services, service permissions and increment supplements of literature species in great demand should be widened and enlarged to provide scientific literature services further for enterprises, research institutes and universities.

Establishment of decision-making model base, decision-making knowledge base and decision-making expert database and inquiry of government technology decision making are needed after the storage of the content which be converted and integrated of the database used with data warehouse and analysis and induction of the characteristics of various types of technology resources to provide decision support service platform, to improve the quality, efficiency and level of technology decision-making and to provide strong technological support and protection for scientific decision-making.

5 Model and Mechanism of Science & Technology Information Service

For the realization of the long-term services of scientific and technological, an integrated service management system and operation mechanism must be established. Systems and mechanisms should be gradually improved with pilot projects and experiments. It should be fully aware of that information service environment is feasible for the existing scientific and technological resources are mostly state-owned assets and mostly established under financial support. The key is to raise awareness, enhance the awareness of decision support technology. It is import to raise awareness and to enhance the awareness of decision support technology.

Meanwhile, interest factors should also be considered in the implementation, models and methods can be implemented should be explored. Because of the cost, unit of interest and departmental interests in the process of development, protection and operation coordination is need for the system exploration and innovation of management and operation mechanism and win-win situation Figure 1.

5.1 Construction of Environmental Science and Technology Information Service Can be Divided into Five Levels, Data Sources Layer, Data Processing Layer, Data Storage Layer, Data Analysis Layer and Data Exchange Layer

1) Data exchange layer is the externalization of the overall service environment

To achieve multi-media information management through the introduction of multimedia database technology.

To provide communication platform for the government, enterprises and individuals

2) Resource analysis layer is responsible for leading and coordinating the work of data integration and publishing

The basic policy of reviewing and approving the data resources to support.

To promote and monitor the implementation of integration of data resources.

To coordinate the resolution of the data resources published in the major issues in the work of various units.

To coordinate the resolution of the data sharing service for the community on major issues evaluation and implementation of incentive measures on information services in accordance with relevant provisions of the platform to work units within the project.

3) Data storage layer is the support of data and protection of the information service platform, integration of resources should work normatively under the guidance of the expert group

To receive the Data processing layer data resources and to publish the latest data in a timely manner.

To provide fundamental environment of storage, backup, management and sharing service for the aggregation of the resource data including the environment, mass storage environment, data service environment, off-site data backup and disaster recovery is necessary.

To research and develop the platform of data integration, management and shared services.

To research and develop of database construction and shared services standards.

4) Data processing layer to collate basic data

To review and to control the quality of centralized management of data resources, to sort and to classify according to subject characteristics and requirements and to release data resources of central management according to the requirements of the authors and institutions and resource integration of the Group of Experts promptly.

978-1-935068-76-1 © 2011 SciRes.

To provide technical support and training for the information services of the node Units database construction and management.

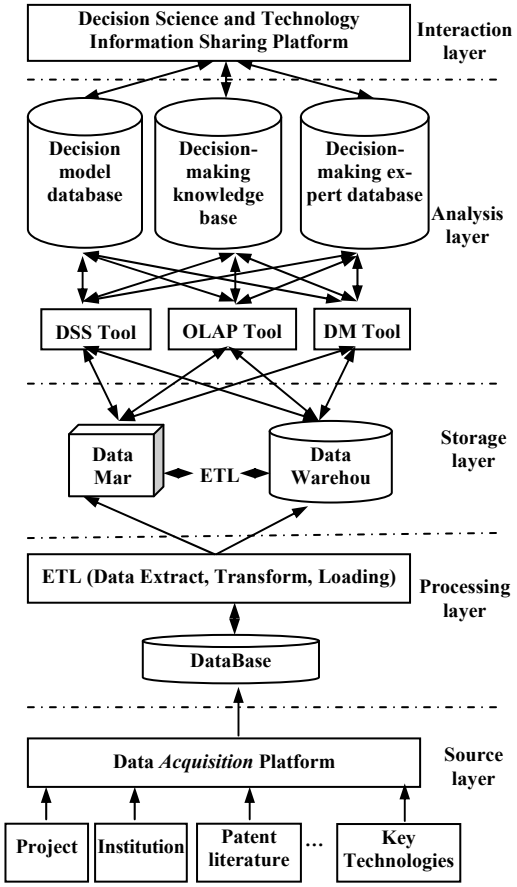


Figure 1. Information Service Model of SDSS

To organize and facilitate the development of standards and norms related to data management information services .

5) Basic data resources provided by data acquisition layer

To collect, process data, construct database and to update maintenance work.

To evaluate and control the quality during the production process of data and, to ensure data quality.

To release data hierarchically for the public, providing shared services.

To gather data to the data center according to the requirement .

5.2 Innovation Mechanism of Science and Technology Information Resources

Establishment of innovation mechanism science and technology of information resources would strengthen sharing mechanism between cooperation units, optimize sharing of cooperation between organizations, promoting the development of scientific and technological information industry and services. The core of innovation mechanism of Science and Technology information resources include knowledge creation, dissemination and application.

Innovation vector includes research institutes, intermediary institutions, universities and enterprises. Mechanism of innovation is the sharing in the field of science and technology in the information environment, supported by the government and laws and regulations and constraints play a role, through a standard format to promote the unity and the market environment. Science and Technology information resources innovation also be impacted by some external factors such as the global innovation system, regional infrastructure, innovative ideas and culture, regional technological innovation ability.

6 Conclusion

Decision Support Systems mainly service for government, base on the scientific data of project information, integrate the collected government reports, statistical data and information from abroad which will be analyzed and processed by analysis tools such as DSS, OLAP, DM, providing information support for government decision-making according to the requirement of decision-making. It is needed to build science and technology information service environment, to demonstrate concentratedly scientific standards, to integrate various scientific and technological resources, to achieve the function of comprehensive, multi-level information services, portals and unified user authentication of platform, to access control the database based on the role for providing the technology services information retrieval and business applications of whole network resources for single industry and cross-industry and abundant one-stop, integrated application services of technology information resources for governments and scientists and a global resource integration and application services platform that is safe and stable for administrative departments.With policy guidance and mechanism innovation, promote the resource sharing within the genesis pits and large scientific instruments and equipments and ensure the accurate and efficient investment of government to research institutes and enterprises for R&D, create a support environment of scientific and technological innovation.

References

[1] HUANG Tao, The Information Processing in the Construction of Sci-tech Information Resource Database

[2] SHUI Jun-feng, CHEN Shu-xiao, GUO Mao-lin, WU Jin-wang, WANG Zong-yan. Research on a Governmental Decision Support System Based on Data Warehouse

[3] Codd,E.F. Providing OLAP(On-Line Analytieal Proeessing) to User-Analysis.An IT Mandate. E.F. Codd and Assoeiates, 1993.

[4] WANGXiu-kun, YANGNan-hai, ZHANGZhi-yong. A Scheme Based on RBACof Database Secure Access Control Designing and Realizing. Dept. of Computer Science & Engineering, Dalian University of Technology, Dalian Liaoning 116023, China

Scientific Research

Information Ecology for Knowledge Communication in Knowledge Asymmetry Settings

Chammika MALLAWAARACHCHI, Xiangxin ZHANG

School of Management, Jilin University, Changchun, China

Abstract: In this research note we illustrate the relevance of the notion of Information Ecology for knowledge communication. We outline where and how elements of organization are crucial for knowledge communication. We introduce key elements of Information Ecology's for knowledge communication.

Keywords: information ecology, organizational learning, knowledge communication, knowledge transfer

1 Introduction

New technology, new markets and legal aspects make knowledge communication processes in organizations are increasingly complex. Knowledge communication, insights, skills, experiences, cannot easily communicate as facts or figures. These processes, however, are more complex than simply transfer of information. The types of knowledge, know-how and know-why communication, is crucial for the decision making and also play a vital role for organizational sustainability. In fact, however, the asymmetry of knowledge due to the organizational information politics, information culture and behavior, and available technology is crucial in terms of knowledge communication.

However, humans play a remarkable role in knowledge communication processes than technology. But, only few organizations have already noted that the technology is just a tool uses for knowledge transfer. Now organizations isolation functional character is dying instead the evolving concept is playing a vital role. In another way, organizations become treating as eco-systems.

Eco-system around us teaches that all components: humans, resources, communities, knowledge and environment are inter-related and one component's behavioral changes will effect to the entire co-existent of the eco-system. In organizational eco-system changes affect to knowledge creation, sharing, diffusion, communication, transfer and disposal etc and list goes on. But we presume that having advanced technologies in place will that fulfill organizations knowledge communication needs. However, that is not the case. It has shown that many organizations have been focusing on technology but not other elements of the organization's information ecology. Therefore, this paper is intend to introduce information ecology model that "the complete information environment of the organization" for knowledge communication.

This paper is organized into four sections. The next section describes what knowledge and knowledge communication is. The section three discusses the concept of information ecology and its main components: information politics, information culture and information technology while the ways in which how to use them for knowledge communication. The final section ends with conclusion and implication challenges.

This work was supported by "985 Project" of Jilin University

2 Knowledge and Knowledge Communication

2.1 Knowledge

Knowledge is the capacity to act. As Peter Drucker notes knowledge is information that changes something or somebody either by becoming grounds for actions, or by making an individual (or an institution) capable of different or more effective action. So knowledge has to communicate effectively and efficiently. It is not an individual task. That is a series of process: from data to information to knowledge to intelligence and to wisdom. The figure 1 depicts how sub-units inter-related in knowledge communication process.

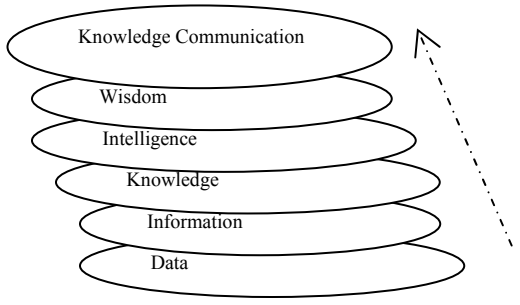


Figure 1.conceptual idea of knowledge communication chain

As we know, population statistics is just data sets. Knowledge sources (e.g. databases, help desks) are repositories of information but not knowledge because they do not have the capacity to act. In fact, knowledge cannot be managed. For example a testing taste of wine is knowledge. To use that knowledge requires a certain level of intelligence. However, have the wisdom; you can use knowledge through collaborative intelligence. Wis-

978-1-935068-76-1

dom is an effective use of intelligence and then it tells you what to pay attention to and what to communicate.

The nature of knowledge Schultze and Leidner[1] describe as metaphors while knowledge views as an object, an asset, and as situated practice. These differentiations are based on how knowledge operationalized. Knowledge as an object refers to rules, explanations and problem solution sets that independent from an individual. As we know knowledge based systems are closely associated with object perspective knowledge. Asset based knowledge is residing within an individual and composing individual expertise, competence and job experience. As Rogers[2] explains knowledge is something that occurs as a result of knowledge creation or knowledge transfer processes. The knowledge as situated practice describes knowledge being socially constructed and shared among people who are working in the same professional field.

Polanyi[3], Nonaka and Takeuchi[4] differentiate knowledge into explicit and tacit knowledge. Explicit knowledge is easily transferred knowledge like publications while tacit knowledge likes "crafting a violin". Boisot[5] categorizes knowledge as codified and un-codified. The term codified refers to readily prepared knowledge that for transmission purposes while un-codified is knowledge that cannot be easily prepared for transmission purposes, for example experiences.

Nonaka and Takeuchi[4] emphasize that how tacit and explicit knowledge are important in knowledge creation and communication. Nonaka et al. [6] proposed a SECI model (S–socialization, E–externalization, C–combination, I–internalization). It represents on knowledge converting process of an organization: [1] converting from tacit into explicit knowledge, [2] explicit into tacit knowledge, [3] tacit knowledge into new tacit knowledge and [4] explicit knowledge into new explicit knowledge. They believe that the successful transfer of tacit into explicit knowledge depends on the use of metaphor, analogy and mental model. Then Davenport et al. [7] define knowledge as "information combined with experience, context, interpretation and reflection". The knowledge management process must give an equal attention to knowledge storage, distribution and integration to achieve significant organizational improvements.

2.2 Knowledge Communication

Knowledge communication is tools and systems developed for supporting knowledge management. For example, As Wilmotte and Morgen [8] note group of legal experts briefing a management team on the implications of new regulations on their business model, Creplet et al [9] urge consultants on business strategies present the findings to the board of directors to devise adequate measures.

So it is understood that, knowledge communication will take place either face to face or media based interactions for successful transfer of knowledge. The figure 2 depicts, however, how knowledge communication takes place in organizations.

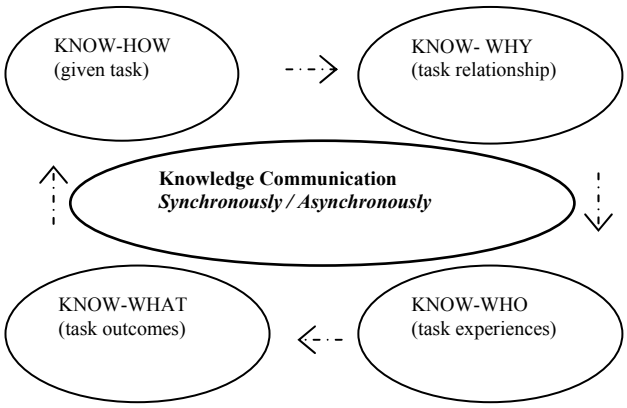


Figure 2.conceptual idea of Knowledge Communication sphere

Knowledge communication can be started at any position in the circle because it is more than communicating information or emotions. It requires convey a context, background, and basic assumptions based on the task going to be performed. Therefore, knowledge communication eases in synchronously or asynchronously approaches. Synchronously approach refers to real-time interactions, such as an instant messaging, videoconferencing that is face to face. An asynchronously approach usually is on media-based tools such as an email, voicemail, discussion forums and shared drives.

It is understood that knowledge communication does not only differ in terms of what to communicate is but also how to communicate. The process of knowledge communication hence requires more reciprocal interaction beyond the organizational environment. We have to bring eco-system into the organization to create a specific type of knowledge communication context so that information can be used to re-construct insights, create new perspectives, or acquire new skills because we still communicate only information and emotions but not knowledge.

3 Information Ecology

An ecological approach to knowledge management was introduced by Davenport and Prusak with the concept of information ecology. They focus on the humans centered information management model in the information environment, the organizational environment that surrounds it, and the external environment of the market place. According to Davenport and Prusak [10] the key premise of information ecology is organizations need to focus beyond the machine engineering focus on the technologies of information. The complete information environment addresses all the firms values and believes

about information (culture); how people actually use information and what they do with it (behavior and work processes); the fit falls that can interferers with information sharing (politics); and what information systems are already in place (technology). The key proponents of information ecology have made an interesting case for focusing on information rather than on the hardware, software or telecommunication networks.

Therefore, we assume: an organization mission; intranet, information management plan; information culture; information politics; physical setting; information staff; and information handling is important elements for knowledge communication for any organizations. Among them, we presume that, information politics, information culture, and information technology are the most important elements for knowledge communication.

Information politics involves the power of information providers and the governance responsibilities for its management and use. Information culture composes an organization's information behaviors and determines that how much those involve value information, share it across organizational boundaries, disclose it internally and externally, and capitalize on it in their businesses. An information behavior includes information sharing which requires the removal of various political, emotional, and technological barriers to encourage the exchange of information among organizational participants. Information technology is already existing, servers, PCs, networks, databases and applications.

3.1 Information Politics

Indescribable side of an organization is the information politics. It shows how information [knowledge] is important as well as powerful in competitive advantages because many organizations reports are Greek to the other areas and all of them are Greek to top management too. The main reason for such a phenomenon is no knowledge communication atmosphere within the organization so that all are fear losing knowledge.

Therefore, information politics can defines in terms of power, agendas, and fights and flights that are related to organizational information and information technology (Travica, Bob [11]); Davenport et al [7]. According to Travica, information and information technology can influence the aspects of information politics (e.g., control over data and information technology is a source of power) as well as being influenced buy it (e.g., the agenda of a certain party determines design and use of data and information technology). Then it is obvious that organization information and information technology constitutes high political stakes. By having some special knowledge that others consider a resource, the knowledge holder can influence thought and behavior of others.

We, as seekers and beneficiaries of information, assume that information and communication technology supposed to stimulate information flow and eliminate hierarchy but that was not. For example, in an information communication system for a minor content changes has a big communication [permission] process. This is a real situation in many organizations! Under these situations you do not have common understanding about your organization's knowledge communication mission and goal so cultivate it. You bring "equal trust" in inter-organizational and intra-organizational levels and also promote "equal responsibilities" for knowledge communication.

Given the situation, table 1 shows five models of information politics are prevailing in many organizations. It is likely to have proponents for more than one of the information politics in many organizations knowledge communication is concerned.

Table 1. Models of Information Politics (adapted from Davenport et al information politics)

Model	Explanation
Technocratic Utopianism	A heavy technical approach to information management stressing categorization and modeling of an organization's full information assets, with heavy reliance on emerging technologies.
Anarchy	The absence of any overall information management policy, leaving individuals to obtain and manage their own information
Feudalism	The management of information by individual business units or functions, which define their information needs and report only limited information to the overall corporation.
Monarchy	The definition of information categories and reporting structures by the firm's leaders, who may or may not share the information willingly after collecting it.
Federalism	An approach to information management based on consensus and negotiations on the organization's key information elements and reporting structures.

Models specify: technocratic utopianism ignores politics, anarchy is politics run amok, feudalism involves destructive politics, monarchy attempts to eliminate politics through a strong central authority and federalism treats politics as a necessary and legitimate activity by which people with different interest workout among themselves a collective purpose and means for aching it.

Therefore, knowledge communication is concerned, what kind of information politics models people in the organization is holding, which model currently predominates, and more desirable as well as way to achieve have to be considered. However, be careful in one thing adapting more models will confuse both information managers and users and same time it will confuse scarce resources as well. Therefore, organizations better choose one model and more continuously work towards to achieve it.

3.2 Information Culture

Organization culture is equally elusive with the infor-

978-1-935068-76-1 © 2011 SciRes.

mation culture because values, assumptions and beliefs are common attributes. As Peppard and Ward [12] point people are a vital resource for best performance of organization therefore the people will be a major determinant of organizational culture. Davenport and Prusak [10] identify information culture in terms of a pattern of behaviors and attitudes that express an organization's orientation towards information. For example, information cultural attitudes are preferences for facts or rumors where as behaviors include information sharing and preferences for types of communication channels. Those distinguish in between information culture concerning to the group and organizational level and information behaviors that are demonstrated at the individual level such as searching information and using them.

The mixture of information and organizational culture would therefore appear to be an integral part of a knowledge organization. It could be argued that to become a knowledge organization it is not only necessary to fulfill the requirements of an information culture but to encourage a culture of organizational learning too. At the same time, to become a knowledge organization it is necessary to instill and nurture successful information culture. Therefore, organizations better first adapt to information culture and then communicate with the components of information culture. To build a healthy information culture the elements such as participate, processes, and information is vital. It shows that integration of organization culture and information culture. And it is suitable to consider further information and knowledge management tools to see progress towards fulfilling information cultural needs. At this stage the information culture is no longer distinguishable from the organizational culture.

As Bloor and Dawson [13] explain an organizational culture is generally agreed to be historically and socially constructed, holistic and difficult to change. Those represent the dynamic interaction of several factors, such as operating and cultural systems, historic and social contexts, external organization context and professional culture and the environment. The operating and cultural systems refer the organizational infrastructure of both technologies and personnel while the shared values and beliefs of the people in the organization is a cultural system. The founder of the organization's vision and values and to past infrastructures and influences refer to the historic context of the organization and the expectations, norms and values of the society in which an organization finds itself are social context. External organizational environment refers to both the business of the organization and the market in which it operates. Professional culture and it sub-cultures and also environment are a large influence on an organization's culture. Professionals have access to two sets of cultural values that of their profession and the organization while sub-cultures can exist in harmony in an organizational culture. Then building harmony in and among the organizational culture, the key aspects, which communication flows, cross organizational partnerships, internal environment, information systems management, information management, and process and procedures are important for build a healthy information culture.

Therefore, information culture, structural, communicational, and social and psychological, is an important element for a healthy information atmosphere. On aspect of structural, organizational hierarchies, functional differentiation and separation, workflow directly or indirectly cause for information culture. Therefore, it affects knowledge creation and knowledge communication activities.

3.3 Information Technology

In organizational context, information technology represents a wider meaning. In particular, the existing information technology such as information applications and information systems are only one segment and there is a very active and vast domain of information technology exists outside the organization. For example, individual customers maintaining networks are very stronger than just an organization's customer details database. Because interpersonal knowledge communication can brings more potentials than just an organization's information system.

As we hinder earlier, having good knowledge on people, it will comprehend maximum benefits from technology of the organization. Because a clear understanding of how people interpret, perceive and act on information, technology can thrives benefits of using them. People at the center of information ecology! People have a profound influence on the functioning of all other inanimate elements in the information ecology. The perception of organizational structures and roles by people has a marked impact on the way an organization functions. Therefore, changes in the information technology must be planned in such a way that the perception of all stakeholders affects the performance of the system positively.

Rapid development of information [computer] technology and competitive intelligence is a big challenge for an organization's existing technology. The usefulness of information technology increases only if the existing systems are flexible enough to evolve in response to the opportunities and the requirements of the external environment. For example, business strategists focus on the internal structure of the organization to match the requirements of the market. To respond quickly to the change in market conditions, a flexible and evolving internal structure is required. Therefore, a static and inflexible information system will not help sustain growth and it will become a stumbling block in the way of organizational excellence by limiting the potential for growth. On the contrary, if the internal structures are

flexible and evolving, the propensity of the organization to transform to the market challenges will be high.

So, information ecology is against the idea of technology change will improve the information environment instead it emphasizes technology is an only one component of the information environment and often not the right way to create change. It describes the interaction of the systems in the overall organization eco-system, rather than depict information systems on technical architecture and engineering drawings. Therefore, this is information engineering not the system architecture. It means organizations better to understand social happenings both in marketing and technological perspectives.

Other component of information technologies represent in information ecology is the convergence of different but link technologies. For example, instructional designs, multimedia technologies, computers and telecommunications technologies, content expertise, business and industries linkages. Integrating, however, these technologies into large-scale systems are a complex and difficult task because rapid internal changes and more capabilities are continuously becoming available. Designers must learn the new capabilities and integrate them into their systems on the fly. New capabilities are not always compatible with old ones, sometimes requiring wholesale redesign of the entire ecosystem, and increasing the cost of implementation.

4 Conclusion

This exploratory study shows that in knowledge competitive advantages, existing concepts on knowledge communication have been drastically changing. Knowledge is growing organism and cannot be managed. Therefore, elements of information ecology: information politics; information culture; and information technology is vital in knowledge communication. Information ecology brings potentials to encounter greatest challenges in knowledge communication facing in organizations now because they just communicate only information and emotions. And also this study offers preliminary yet valuable theoretical and practical ideas for "till young" the concept of information ecology.

In practicable point of view, the use of information ecologies may face humans, social and institutional, and also technological challenges.

Human's influence is the most challenge for an organization to use information ecology because working with emotional responses to artificial knowledge agents obviously will become conflicts. Historically, knowledge and intelligence have been the distinguishing characteristic of human beings. Now computers are able to take over some of the routine and even expert functions of knowledge work. In some areas computers excel over human capabilities. A related problem is the very natural human fear that computerized information ecologies come dangerously close to replacing humans in knowledge work. This conflict is an intensely emotional matter. Like other emotional issues it remains largely suppressed in organizations. However, the conflict and deep emotional resistance among managers and staff can sabotage implementation of information ecologies, or at least reduce their effectiveness.

The next challenge is social and institutional environment of an organization. This stage carrying of information ecologies is more challenging and more conflict. In this level, it effects to the whole organization's functions, tasks, and performance. At the same time, it changes career prospects and earning potential of members. The organizational members require retraining due to changes in structures and systems, and also installation of new equipments.

Above problems compound by the lack of comprehension of what information ecology represents. In some senses, information ecology is an extended virtual brain of the organization. The knowledge function and access of this ecology extends well beyond the cognitive capacity of individual humans or even departments and divisions. It represents a new form of organized complexity that many managers and workers find incomprehensible. It falls outside the collective cognitive map of the organization.

References

[1] Schultze. U, Leidner. D. E. "Studying Knowledge Management in Information Systems Research: Discourse and Theoretical Assumptions". MIS Quarterly Vol 26(3). 2002.

[2] Rogers. E. "Diffusion of Innovations" 5th ed, The Free Press: New York. 2003.

[3] Polanyi. M. "The Tacit Dimension". Garden City:New York. 1966.

[4] Nonaka. Ikujiro, Hirotaka. Takeuchi "The Knowledge-Creating Company: How Japanese Companies Create the Dynamics of Innovation, Oxford University Press: New York. 1995.

[5] Boisot. M. "Information and Organisations: The Manager as Anthropologist. Fontana/Collins: London. 1987.

[6] Nonaka. I, R. Toyama, N. Konno. "SECI, Ba and leadership: A Unified Model of Dynamic Knowledge Creation. Long Range Planning 33, 2000.

[7] Davenport. T.H, Eccles. Robert. G, Prusak. L. "Leadership and Organizational Studies, Management of Technology and Innovation Information Politics". Sloan Management Review, Fall 1992.

[8] Wilmotte. R.M, Morgan. P.I. "The Discipline Gap in Decision Making". Management Review. September, pp.21-24 1984.

[9] Creplet. F, Duouet. O, Kern. F, Mehmanzapir. B, Munier. F. "Consultants and experts in management consulting firms". Research Policy. 30, pp.1517-1535, 2001.

[10] Davenport. T.H, Prusak. L. "Information Ecology: Mastering the Information and Knowledge Environments", Oxford University Press: New York. 1997.

[11] Travica. Bob. "Information Politics and Information Culture". home.cc.umanitoba.ca/~btravica/vita.html, 2010.

[12] Peppard. J, Ward. J. "Mind the gap: Diagnosing the Relationship between the IT, Organization and the Rest of the Business". Journal of Strategic Information Systems, 8. 1999.

[13] Bloor. G, Dawson. P. "Understanding Professional Culture in the Organizational Context. Organisation Studies, 15(2) 1994.

978-1-935068-76-1 © 2011 SciRes.

Research of News WEB Data Extraction Based on Text Feature

Feng WU[1], Yongfeng DONG[2], Junhua GU[2]
[1]*Hebei Institute of Science & Technology Information, Shijiazhuang, China*
[2]*Hebei University of Technology, Tianjin, China*
Email hebwf@126.com

Abstract: With the rapid development of Internet, the speed of the news releasing is accelerated. What is more, it makes the Internet the best place to gather the most news. How to extract information from Web effectively and accurately has become a hot issue. Web information extraction technology arises at the historic moment. An automatic extraction method of news information is put forward which is based on rules combing with the statistics rules, according to the demands. The method can extract user interested information conveniently and its extraction formulation process does not need artificially participation.

Keywords: news page; dom tree; clustering; similarity; chinese word device

1 Introduction

With the popularization of Internet technology in China, Network media plays an increasingly important role in people's life. Because network news is very informative and up-to-date, people choose the network approach to check news. It makes the research about network information extraction particularly important due to the popularity of network news. Therefore, more and more researchers devote themselves to web information extraction work.

In recent years, scholars both at home and abroad has obtained certain achievements and designed various multifarious information extraction methods. At present, Information extraction is mainly divided into two main trends: one based on the rules of information extraction method and one based on the statistical information extraction method [1, 2]. Based on the analysis of these two methods, this article decided to combine them. Through making full use of the characteristics of news pages whose structure is clear and extraction rule has certain rule, we generate the extraction rule. In the process of generating extraction rules, we use statistical methods to summary the characteristics and link information of news content to help to generate extraction rules.

2 Overall Process of Extraction Method

This paper mainly studies the automatic extraction of news information, namely manual-needless annotation training record set. The extraction process mainly includes web clustering, rules structure and information integration [3,4].

1) Web clustering

Many well-known news portals usually adopts fixed template to publish news. If we can get the templates of some websites, we can extract information more quickly and accurately. Web clustering can well solve the problem. Sorting out news pages of the same templates and separating the news pages of different templates. It makes a small difference within the class and a big difference between the classes.

2) Rules structure

In the results of web clustering, we do the common analysis to the same cluster web set and extract the optimum selection rules. This article only introduces the progress of generating extraction rule for news headlines and content.

3) Information integration

According the generated extraction rules, we use the HtmlParser functions to achieve web information extraction and save the results, which convenient the data integrating and the subsequent value-added services, to database. Figure 1 shows the overall process.

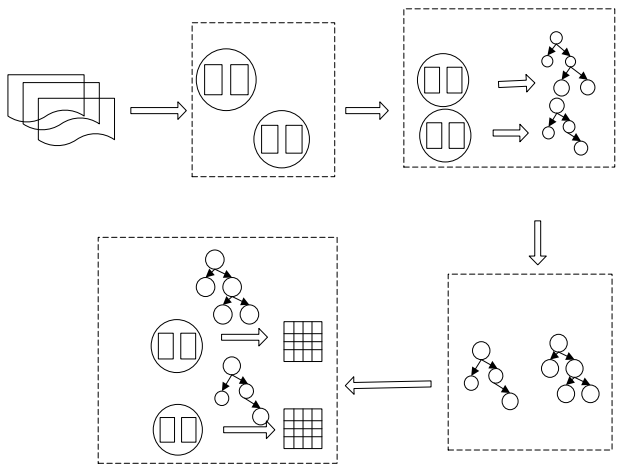


Figure 1. Main flow of information extraction system

Identify applicable sponsors: Open project of Hebei Science & Technology Information Processing Laboratory; Scientists and engineers service businesses (2009GJA10020)

3 News Text Extraction

This paper first positions standards text node based on Chinese feature symbols, and then combines DOM tree node link information to find nodes which fit the threshold setting in advance. At last, we extract news text. Standard Text Node is a kind of node which contains the most disconnected characters and periods in the paragraph of webpage text. Below are the main steps:

1) Constituting DOM tree

Firstly, we should convert the webpage into DOM tree and traversal the tree, then remove nodes which have nothing to do with text, such as Script, Form and so on.

This can accelerate the extraction rate and improve the extraction accuracy.

2) Looking for standard text node

Recursion traverse the node in the DOM tree, and look for text node, then record the text node's length, the order in the webpage and the detailed links from root node to this node. The information of node link refers to the node's whole route from root node to its father node. As shown in Figure 2. For example, the link of node text1 is Html->Body->table->tbody->tr->td->b, the paper can find distribution consistent text node through the link information.

Using the following below calculation method to calculate the weight of text node:

$$W = textLength \times stopCount$$

textLength is the length of text node, stopCount is the number of period which is contained in the text node. This method not only considered the information of the node's length, it considered the density situation of the period which is contained in text. The text node is sorted by the value of M from large to small and then we choose the node whose value of M is the largest as the standard text node, named Node0.

3) Looking for other text nodes similar to the link of standard text node

Extracting the link of Node0 and comparing it with other links of text nodes. In the general website structure, the different paragraphs in text content or the links of information are not completely same; some contents are surrounded by some modification marker, so they have one more layer markers than the other text information. To avoid these modified text content been filtered out as noise, we introduce threshold $K_{i,j}$ to measure the similarity of link i and j. $K_{i,j} = 2S_{i,j} / (L_i + L_j)$. $S_{i,j}$ is the number of the same link layers between i and j. L_i and L_j correspond to the length of the link i and j. When calculate the value of $S_{i,j}$, if the root node in link i is different with one in link j, $S_{i,} = 0$, and $K_{i,j} = 0$; Otherwise, $K_{i,j} = 1$. If defined as above, Threshold will has transmissibility, related to both the link of fathers close degree and the depth of the leaf nodes. It is more tallies with the actual situation.

Extract the text node which fits the threshold setting in advance and display them according to the order in the webpage.

4) Looking for hachmonite node

Experiments show that when the threshold is 0.88, this method for extracting gets the highest accuracy. But testing of WebPages which contain image finds that the effect is not very good. It is mainly because that picture instructions often mingled in text. Based on such problems, we find out the public father node of text nodes as the head of the text node. The text information which is covered by the head of the text node may contain the information which is not belonging to the text information. Such as video, images and so on. Because the instructions of picture or video usually are different from the text font styles, we can use this feature to exclude image and its instructions. Through such treatments, we can improve the accuracy rate of the text information. Figure 3 is the whole progress of text extraction.

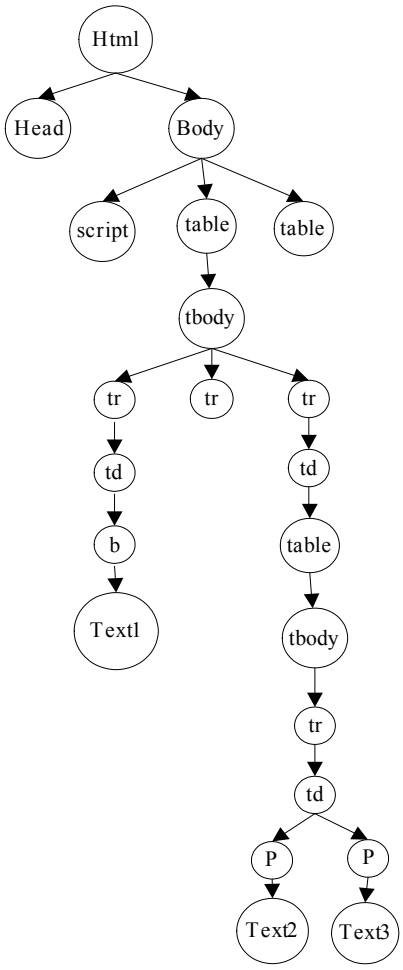


Figure 2. HTML tree representation

978-1-935068-76-1 © 2011 SciRes.

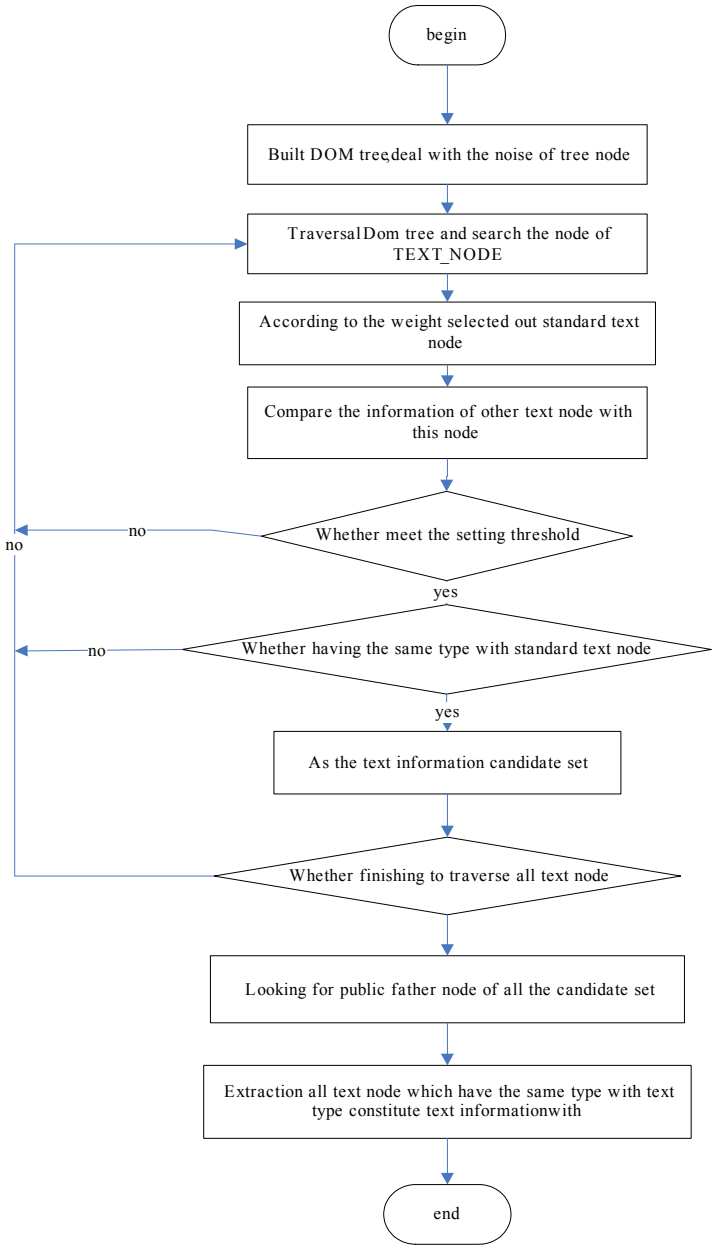


Figure 3. Main flow chart of extraction news text

4 News Headlines Extraction

News titles usually display in the list of news. The list will only show parts of words for the long titles and instead the rest of the title with an ellipsis. So if you want to get the accurate title, you should extract it from the page of news details. Literature based on C4.5 takes decision tree as machine learning extraction model to extract news title. And it does some Sample training and extraction based on title's features in the web page. Although websites after training have a high extraction rate, the not-trained websites have a low extraction rate. On the basis of referencing the experimental features, we design a method based on labels matching and a extraction algorithm combining IKAnalyzer.

4.1 Distribution of News Headlines in Webpage

Through observing large news web portals, such as Sina, Netease, Tencent as well as university and government portals, we get a conclusion: an html page title tends to appear in web pages of the text. Because this is where visitors mainly concern, web editors often modify the news headlines by amplify the font. The news headlines mainly distribute in <title> or <hl> except some special circumstances. But we can get the conclusion: once the tag of <hl> exists, the title of news must be the

content it contains. For some famous website, the title is the content that the tag of <title> contains.

Based on the above characteristics, we put forward the headlines extraction algorithm. The article fully makes use of the tag of < title> and <h1> to find the title of web news. If this method doesn't work, we can use IKAnalyzer Chinese parting-words device to find the title from text.

4.2 Headlines Extraction Based on IKAnalyzer

First select the candidate from the title's node set in DOM tree, then use IKAnalyzer participle device to divide the candidate node into word, next use the index function to set the extraction of text as index area, and index the text with the candidate nodes of title as the keywords, record the hit numbers. At last, choose the candidate node whose hit number is the largest as the title node. The steps of extracting title are shown in Figure 4 and Figure 5.

4.3 Extraction Method Based on the Label Matching and IKAnalyzer

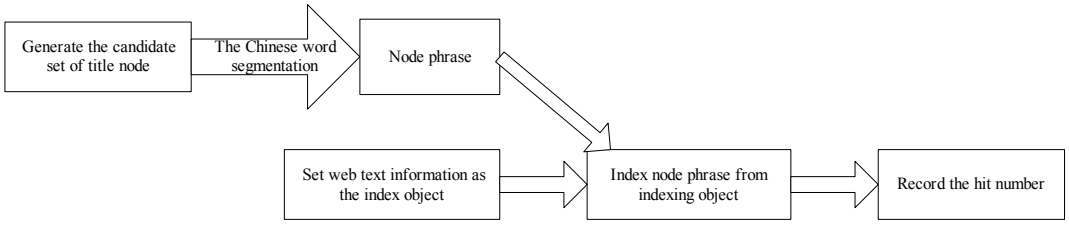


Figure 4. Main process of IKAnalyzer

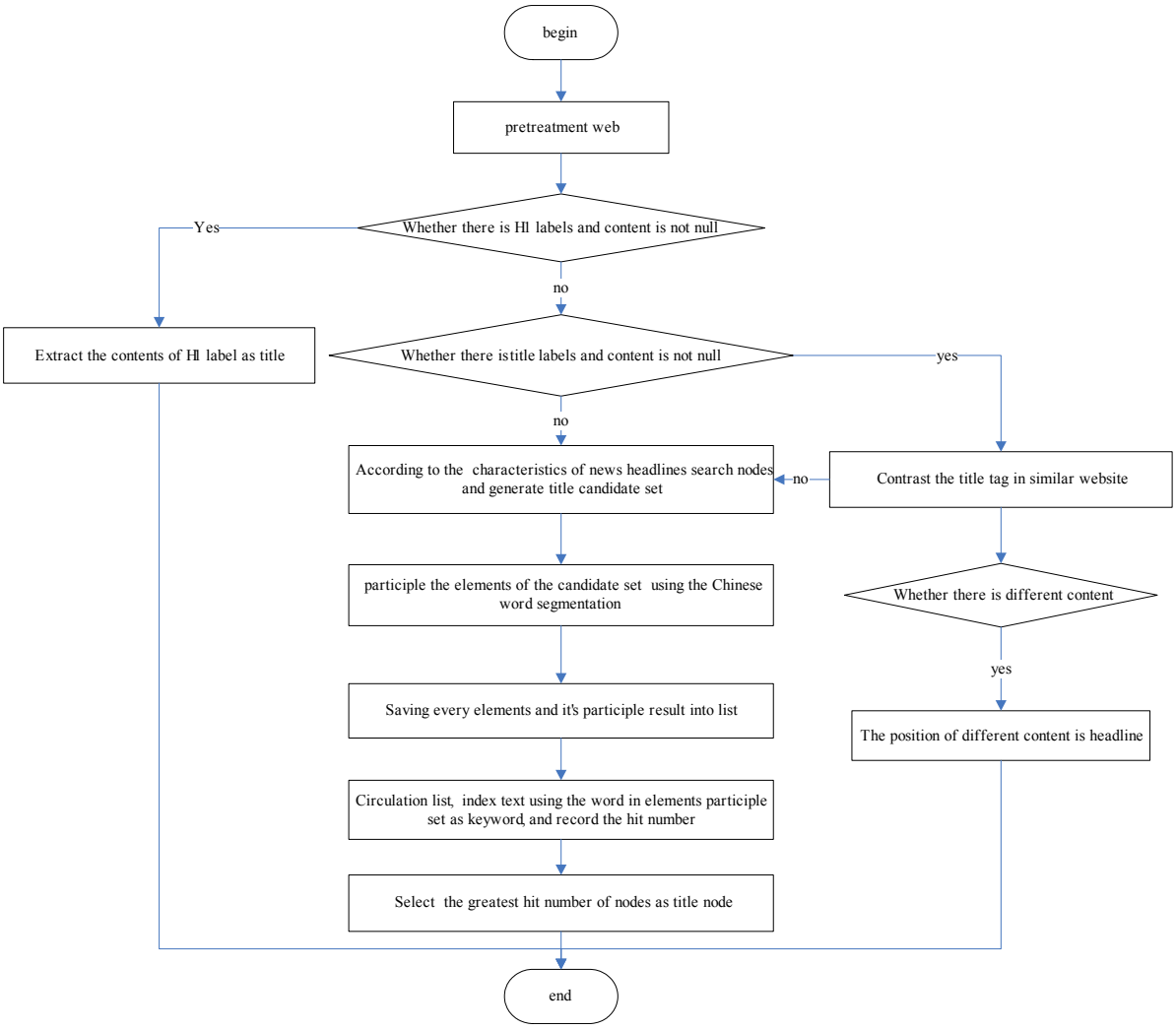


Figure 5. Main flow chart of Extraction web title

978-1-935068-76-1 © 2011 SciRes.

The method based on IKAnalyzer can extract the title of news, but this method puts text of news to memory and then index among full text in the memory. If the text is informative and memory is small, the speed of extraction will be affected. This article makes full use of the characteristics of tag in the webpage and puts forward an algorithm about extracting headlines. This algorithm unifies the extract method which based on the tag characteristics and IKAnalyzer. The extraction rate is much quicker than IKAnalyzer. Extraction accuracy and generalization are more accuracy and efficient than pure application for label matching extraction method.

5 Extraction Result Assessment

5.1 The Evaluation Standards of Information Extraction

The performance of information extraction mainly can be measured by two evaluation indexes: Recall(*R*) and Precision (*P*).

$$\text{Recall}=A/B$$

$$\text{Precision}=A/C$$

A is the number of correct information points which is extracted, *B* is the number of all correct information points, *C* is the number of all information points. In order to evaluate the performance of the system, we introduce the weighted average of recall, precision and index *F*. Below is the calculation formula.

$$F=\frac{(T^2+1)*P*R}{(T^2*P)+R}$$

T is the relative weight of recall and precision, $T=1$ says both are the same important, $T>1$ says precision is more important, $T<1$ says recall is more important.

5.2 Experimental Results and Analysis

Using web crawlers to crawl 100 pages for each website restrictively and analyzing the effects of the news page. Through the experimental observing, we find the accuracy can reach 100% for extracting these websites. For text extraction, the structure of website in cultural field and its content is clear and the capacity of information is not much. So the effect is good for extracting in these websites. But when testing the website which contains much more noise like media, the effect drops. Table 1 shows the result.

Table 1. Result of testing sites

The name of Website	R	P	F
Hebei university of technology	98%	100%	99%
Hebei natural science fund	94%	94%	94%
Hebei industry and information hall	83%	100%	91.5%
sina	74%	90%	86%
average	87%	95.6%	91.3%

6 Concluding

Through studying the features of news site and the existing extraction systems, analyzing their advantages and disadvantages, this article puts forward a certain commonality extraction algorithm using the advantage of the fixed format of news page. This algorithm uses the feature of text to locate text area, and uses the Chinese word segmentation to find out the news title in the text accurately. Experimental results show that this algorithm has certain feasibility.

References

[1] Bing Liu, Web data mining, Beijing:Tsinghua University press, pp.249-250, 2009.
[2] YangYong Gui, The key technology research and implementation of Chinese information extraction,Master degree theses, Beijing: Beijing university of posts and telecommunications, 2008.
[3] Tie Xu and JiaNing Geng, The reaserch of web information extraction method[J], Information technology,vol.4,pp. 112-114, 2009.
[4] Reis DC,Golgher PB,Silva AS,et al.Automatic Web News Extraction Using Tree Edit Distance[C]. In: Proceedings of the 13th International Conference on World Wide Web,2004, pp.502-511.
[5] Qing Zhu and XiaoXu Lv, HTML title extraction based on machine learning, Micro computer information, 3rd, vol.26,pp. 15-16,2010.

Scientific Research

Research of Web Crawler and Web Information Extraction

Yongfeng DONG[1], Bin GAO[1], Hongyong GUO[2]

[1]*Hebei University of Technology, Tianjin, China*

[2]*Hebei Institute of Science & Technology Information, Shijiazhuang, China*

Email: Dongyf@hebut.edu.cn

Abstract: With the rapid development of Internet and growing larger of web data, it is an urgent problem how to extract information from the web fast and efficiently. In order to make more fully and effectively use of web information, we get into the research specific to web information collection and information extraction technology. The information collection technology has included the web page grabbing, the extraction of URL and its optimization, as well as the strategy of preventing repeated grabbing and other key technologies. Based on these, this paper does research into the information extraction technology which is specific to the extraction of sample pages of information acquisition. According to the actual requirements, we design and implement an Information Extraction System based on Htmlparser. This system uses the web structure feature of tag as an information extraction rule template. The simulation shows the system has high accuracy, recall rate and practical application value.

Keywords: information collection; information extraction; htmlparser; fatures tag

1 Introduction

Along with the rapid development of Internet, Web information is explosive growth. This brings users a lot of difficulties to find the information they need. The search engine is an important way to solve this problem. As the basic component of search engine, the Web information collection played an important role. The Internet is a vast repository of information and it contains a lot of potential, valuable information. However, it has been more and more difficult to obtain the user needed information from the mass information. In this case, Information Extraction Technology (Information Extraction, referred to as IE) came into being. The traditional IE mainly uses the natural language understanding technology, aiming at realizing the extraction of plain text. Along with the boom of Internet, the information extraction technology which based on Web has become the focus of research in the field of information extraction. This paper first introduces the Web information collection and its simple implementation. On this basis, we launch the web information extraction technology research and discuss the data extraction method based on the structure of HTML. Considering the actual application requirements, we design and implement a Web information acquisition and processing system based on Htmlparser.

2 Web Information Collection

2.1 The Theory of Web Information Collection

Identify applicable sponsors: Open project of Hebei Science & Technology Information Processing Laboratory; Scientists and engineers service businesses (2009GJA10020)

Web information collection is using the link relationships between Web pages to get the page information automatically. This process is basically accomplished by the Web Crawler. Web Crawler usually consists of three parts: Spider, Controller and Web Library, as shown in Figure1.

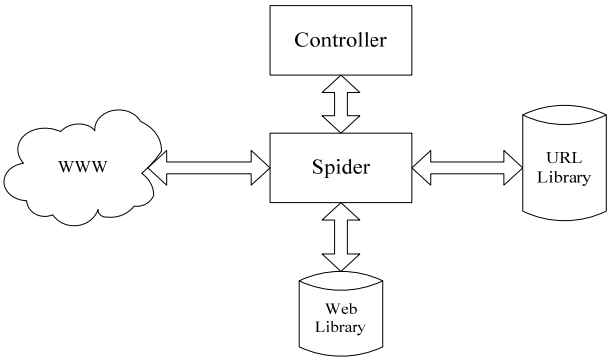


Figure 1. Theory of information collection

2.2 Web Collection Crawler

In the aspect of Network Information Collection, Web information crawler is the key point. Web crawler can be divided into common crawler and focused crawler. Through the use of the deep traversal strategy of common crawler combing with Regular Expressions, we design a simple program of crawler. The design of crawler Program flow chart is as follows:

Crawler can mainly be divided into Page analysis module, URL list module, Timing acquisition module and grabbing & preventing repeat module. The specific

978-1-935068-76-1

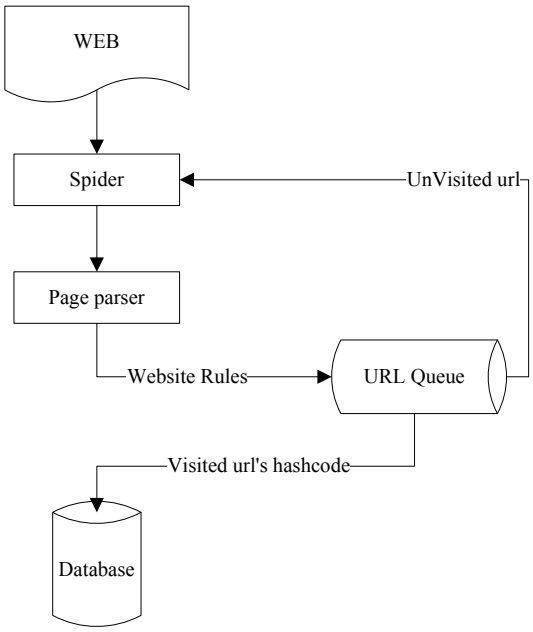


Figure 2. Flowchart of Spider

functions are as follows:

1) Page analysis module

Firstly, Web Crawler obtains a URL link and uses the URL class of Java to read the date source, then uses Htmlparser to read all the links in source combing with regular expression to form web page sets conformed to the rules. At the same time, in order to improve efficiency of crawl, we use multi-threading technology to accelerate the timely grab.

2) URL list module

This module is mainly responsible for recording the visited URL lists and unvisited URL lists of the pages which have been climbed by crawler. For the climbed URLS, we conduct a URL initialized processing. URL initialization include URL case conversion, URL the ".", ".." and other treatments. Ultimately, it turns into a legitimate URL.

3) Timing acquisition module

Because some sites have anti-crawler strategy, in a certain period of time, we can not scan them many times without interval. Given this, we set the timing acquisition and monitor the real-time information of web sites by setting the time cycle of crawler.

4) Grabbing&preventing repeat module

In order to avoid the repeat of crawler and improve the efficiency of crawl, we compare the grabbing situations of Web sites through warehousing URLS. During the time of database storage, we calculate the HASH value of URL string to form the URL eigenvalues. By this way, we can reduce and optimize the data access time of database queries, significantly improving the efficiency.

3 Web Information Extraction Technology

3.1 Overview of Information Extraction

The initial research about Information Extraction begins in the mid 60s of the 20th century. Information Extraction use syntax and semantics to obtain information of specific format from the text. In the end of 1980s, Message Understand Conference plays an important role in promoting the development of Information Extraction. In December 2000, ACE (Automatic Content Extraction) evaluation meeting launched officially, which further promoted the development of Information Extraction Technology.

Now, the popular Web Information Extraction Technology includes Information Extraction technology based on Hidden Markov Model and Information Extraction method based on Ontology etc. Although the above methods have been successful on the Information Extraction, they are mostly based on complex mathematical models and make it hard for projects to implement.

3.2 Information Extraction Process

Through the study of existing information extraction methods, this article puts forward a method which is based on the method of Html Web information automatic extraction. Before the information extraction, we make the Web document parsing a syntax tree, then form extraction rules semi-automatically. It will achieve the fast and accurate extraction of web in specific topics. The main steps are as follows:

1) Page optimization

The page optimization disposal system includes page tags repair and page noise processing. Through the repair and compensation of tags, the pages become web document accord with XML standard. Hence, it is easy to form DOM tree of pages and extract the information. The noise processing is mainly to eliminate advertising information and display the styles of web information, which hinder information extraction etc. So it can speed up the efficiency of information extraction.

2) The Formation of Dom tree

After pretreatment of the Web page, we use the tree component of ZK (ZK is a web application development framework based on Ajax/Xul/Java) technology to form the Web DOM tree structure. As shown in Figure3.

3) Formation extraction rules

We define a feature tag as the extraction rule for a certain set of pages. Because the information of Web Dom tree is usually located in a tag attribute, we define a standard tag:

(tagName tagAttr=“tagValue”) + tagId

Among them, tagName is the tag name of information; tagAttr is the tag attribute; tagValue is the attribute value of tag, tagId is serial number of standard tag. As shown in Figure3.

4) Information Extraction.

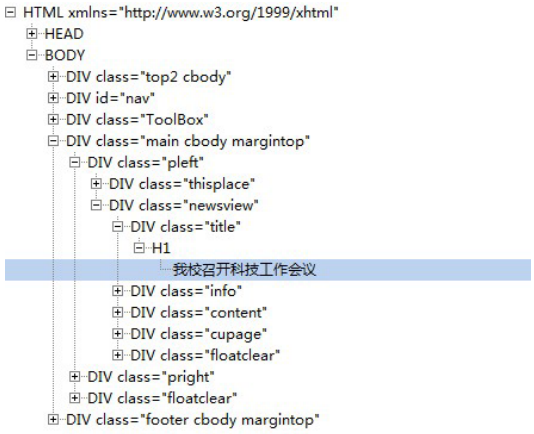


Figure 3. DOM of Web

According to the template rule in rules library, we achieve the accurate, rapid extraction of the similar web information

3.3 The Conversion Treatment of Information

We can conduct a variety of transformations to information by using extraction rules. To the extracted information, it can be converted into text or tag information chosen by users and then it can be saving to the database. Finally, the system displays the extracted information in database.

4 The Information Collection System

Through the analysis of information collection technology, we design a system based on Htmlparser.

4.1 Introduction of Htmlparser

Htmlparser is a HTML parsing library written by pure Java, mainly used for analyzing the web documents. Html parser can use three ways to parse and filter html, Lexer, Filter as well as visitor.

There are three types of nodes in Htmlparser: RemarkNode, TagNode and TextNode. Htmlparser will read binary data stream, then conduct code conversion and lexical analysis and other operations, generating a node set of hierarchical tree structure. Htmlparser mainly uses the Node and Tag to express HTML.

Node is the foundation of forming a tree structure to express HTML. All the data representations are the realization of the interface Node. Node defines the page object expressed by page tree structure and the method to getting the parent, child, sibling nodes. It defines the method that associates nodes with the corresponding html text and defines the corresponding start-stop position of nodes.

4.2 The Structure of Information Collection

Through using the above Web information collection technology combining with the information extraction method, we preliminary design a relatively usable Web information collection system. The design process is as follows:

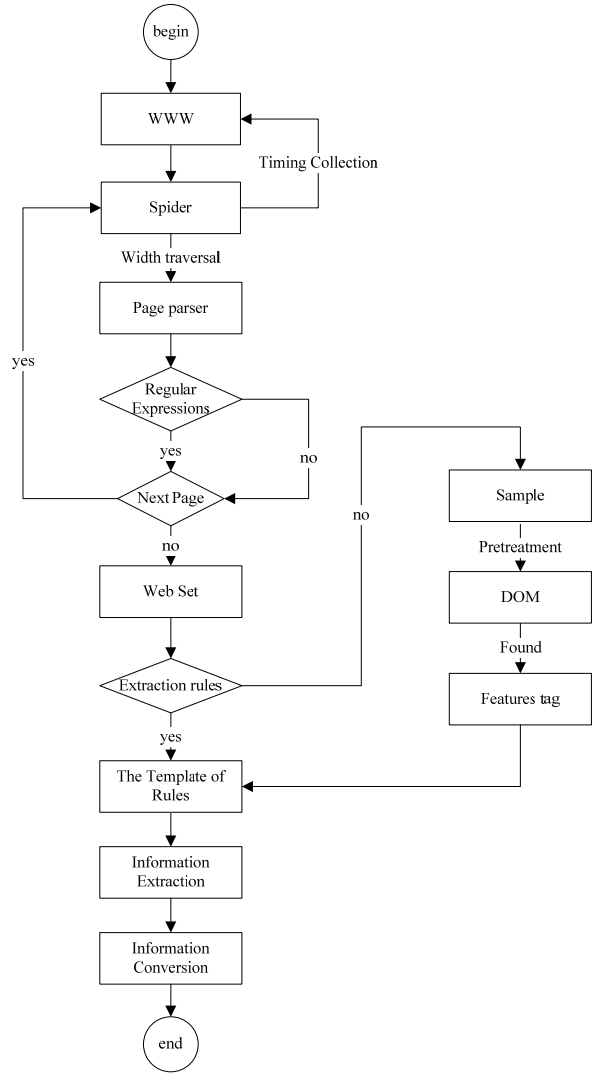


Figure 4. Information collection process framework

5 Experimental Result and Analysis

In order to test the performance of the system, the test environment and experimental platform are given in Table 1.

Table 1. Experimental environment and platform

Test Environment	CPU	Memory	
	2.0 GHZ	2.0 G	
Test Platform	Operating Environment	Database	Operating System
	Eclipse 3.4	Sql2005	Windows 7

Using the prototype system, we test 400 pages from

978-1-935068-76-1 © 2011 SciRes.

four different websites. The efficiency of the system is tested by two indicators: the recall rate and the precision rate. The Experimental data are shown in Table 2. The data show that this method can remain a higher precision; what is more, it has a high recall rate. Of course, to some websites of complex structure, data item, the effect is not very ideal.

Table 2. Test prototype website

URL	Recall	Precision
http://www.hebut.edu.cn/html/xiaonaxinwen/	98%	100%
http://www.tstc.gov.cn/zhengwugongkai/yw/	94%	94%
http://news.qq.com/newsgn/zhxw/shizhengxinwen.htm	74%	90%
http://military.people.com.cn/GB/1076/index.html	87%	95%

6 Summary

This article introduces the information collection technology and related Web information extraction technology, combined with the actual needs of the design information acquisition system. The system has good recall and precision rates through the analysis of experimental data, so it is suitable for small and medium-scale information extraction.

References

[1] Hu Guoqing, Li Jianhua, "An automatic extraction method of Web design and implementation of information". Computer Engineering.China, vol.1, pp.38-40, 2009, 2009, 1:38-40.
[2] Li Shuchen. Research of the Internet Information Acquisition and Processing Platform. Beijing Jiao tong University, 2009.
[3] Shao Hui and Li Fang. "Study and Implementation of Dynamic Web information Extraction based on tree model Algorithm." Computer Applications and Software. China, vol. 24, pp.99-144, October 2007.
[4] Liu Yaqing. "Web Information Extraction Based on Hidden MarkovModel", ComputerEngineering.China, vol35. pp.25-27, 2009.
[5] Zhang Yanchao, Liu Yun, Li Yong and Shen Bo. "Study of Web Information Extraction Technology Based on Automatically Generated Template".Jou Rnal of Beijing Jiao Tong University. China, vol35, pp.40-45, October 2009.
[6] Huang Ying and Huang Zhiping. "Design and Implementation of Web Information Extraction Based on HtmlParser". Journal of Jiangxi University of Science and Technology. China, vol28, pp.26-28, December 2007.

Financial Competitive Intelligence Systems in Financial Enterprises in China: Their Establishment

Juanjuan WANG[1], Feng CHEN[2]

Institute of Scientific and Technical Information of China, Beijing 100038, China

Email: wjj263115@sina.com, chenf@istic.ac.cn

Abstract: The competition of China financial market has been getting fierce due to the in-depth reform of the industry and its fast integration to the global financial market. This paper falls three parts: discussing the necessity and importance of establishing financial competitive intelligence system in China's financial enterprises; anglicizing the functions of financial competitive intelligence system; and suggesting possible actions during the process of establishing financial competitive intelligence system in China's financial enterprises.

Keywords: competitive intelligence; competitive financial intelligence; competitive intelligence system; financial enterprise

1 Introduction

Competitive intelligence has been gradually set up in China and fast developed since the 90's of last century when it was introduced to China from abroad. Financial competitive intelligence system belongs to the domain of research of enterprise competitive intelligence. Although the competitive intelligence has shown its great value in competitive industries, the studies on financial competitive intelligence system is still at its primary stage in China.

Financial industry is a special sector providing financial services and products by commercial banks, investment banks, insurance companies as well as other financial institutions. China's accession to WTO has been more than ten years, China's financial industry, including insurance, commercial banking and investment banking, has completed their integration to the international market system, meanwhile, more and more international financial enterprises have entered China's market. Therefore, Chinese financial enterprises are not only facing fierce competition from domestic rival, but also the challenges from international financial enterprises[1]. Financial competitive intelligence can help China's financial enterprises obtain the best practices from their international rivals: a) abilities of adapting dynamic financial market; b) managing possible financial risks, c) learning the innovative financial products and services; d) expanding their market and improving efficiency[2]. An effective financial competitive intelligence system can improve competitiveness for the enterprises, it's of extremely importance for China's financial enterprises to establish such a system.

2 The Necessity and Importance of Establishing Financial Competitive Intelligence System

2.1 The Definition of Financial Competitive Intelligence System

Competitive intelligence research is about competitive environment, competitors and competitive strategy, it is not only a systematic, timely, actionable information product, but also a process. Honorary chairman of competitive Intelligence branch of China Science and Technology Information Institute, Bao Changhuo pointed out that: competitive intelligence system is a strategic decision supporting and advisory system, which based on human intelligence and used information networks as a means, to attain the goal of the company, enhancing the competitiveness[3].

Financial competitive intelligence system is a human-computer interaction system which improves the competitiveness of financial enterprises by creating intelligence products for financial enterprises decision-making system.

2.2 The Necessity and Importance of Financial Competitive Intelligence System

2.2.1 Competitiveness Enhancement of the Financial Enterprises

Up to now, competitive intelligence system has not been established in Chinese financial enterprises, due mainly to their lacking awareness for competitive intelligence. While on the contrary, multinational financial enterprises attach great importance to the collection and analysis of competitive intelligence.

There are many foreign financial firms establishing offices or representative offices in China which collecting and analyzing China's economic and financial competitive intelligence before and after China's accessing to WTO[2]. In the information society, who controls information, who will controlls the fate of the enterprise.

1. Juanjuan WANG, female, Master student of ISTIC from 2010 to now, competitive financial intelligence.86-10-58882556;

2. Feng CHEN, male, research fellow, doctor, competitive intelligence, technology foreight, S&T strategy and policy. 86-10-58882556

978-1-935068-76-1

Facing challenges domestically and abroad, China's financial firms should have make full use of advantages on wide network coverage, localization, culture, government support, historical data, etc. to establish effective financial competitive intelligence system.

2.2.2 The need for risk management

High risk is a main feature that financial enterprises different from ordinary enterprises. Such as, banks and insurance companies have to bear the risk which other social organizations, economic organizations and individuals may not want to take, which is one of the reasons for their existence. Therefore, risk management is the routine and primary task for financial enterprises. Financial risks can be divided into two types: systemic financial risk and non-systemic financial risk. In order to avoid systemic risk, financial institutions need to track domestic and international economic and financial developments, and focus on their internal risk control mechanisms. To reduce the impact of non-systemic risk, financial firms need to track the global politics, economic and financial developments, and predict the possible trend in the future.

The subprime mortgage in the States has caused global financial and economic crisis, and greatly changing the world financial system and market. It gave the world a strong warning for financial risk management. Financial competitive intelligence can help financial enterprises to improve their ability to prevent internal risks, also can monitor and quantitatively analyze the impacts from external environment, therefore, advice them to take precautions to prevent risk in the first place.

2.2.3 The Need to Improve the Quality of Information

The ultimate goal of competitive intelligence is to improve the competitiveness of enterprises. Intelligence is the fundamental of all the competitive intelligence processes for its quality determines the quality of intelligence system, such as intelligence analysis, and intelligence products. The financial industry has a highly developed information system, its information is based on rapidly changing environment and involving multi-aspects, which come from government departments, industry regulators, financial market, operating companies, issuers, underwriters, investors, customers, etc. Diversification and vast amount of financial information significantly increases the processing and analyzing difficulties. Many financial enterprises have established data warehouse system with the main purpose to collect customers' information, and some of them have established management information department, however, most of the studies on competitive intelligence still stays in the collection phase, and no in-depth research and analysis for competitors. Moreover, the connection between competitive intelligence and strategic management decision and operations is not able to be established satisfactorily.

Establishing an efficient financial competitive intelligence system would be enable the financial enterprises not only to collect and process internal information systematically and intelligently, but also be able to help them track competitors and domestic and international economic and financial environment changing.

3 Main Functions of Financial Competitive Intelligence System

Competitive intelligence research is about competitive environment, competitors, competitive strategy. Financial competitive intelligence could analyze economic and policy environment for financial enterprises, tracking industry trends, monitoring competitive activities. It plays an important role in financial business decisions, maintaining and enhancing competitive advantage, etc.

3.1 Monitor Domestic and International Political and Economic Environment and Monitor Industries Environment

China's financial industry can not develop with detachment from the world; the world's financial industry also needs China. No financial enterprise can discuss development and competitiveness without considering the world economy and international financial system. For example, though the 2008 global financial and economic crisis did not start in China, it has greatly impacted China's economic and financial system. So in the era of financial globalization, Chinese financial companies should not only focus on domestic economic and financial environment, but also monitor the relevant foreign countries and regions' development trend and potential impact on China's financial industry, then gain experience from them and avoid mistakes.

The developments of domestic industry affect companies' operation directly. Financial enterprises should monitor the development trend of the industry, compare the differences between Chinese and foreign industry, and predict future trends, then avoid risks and search for business opportunities.

3.2 Market Demand Research

The financial industry is customer-oriented[4]. Understanding and satisfying customer needs plays an important role in enhancing customer satisfaction and loyalty.

For example, if insurance companies want to sell a new insurance product in a region, first, they must know the region's social, economic, legal and political environment and customs, people's income levels, and financial management concepts and so on. Then the company determines whether to enter the region and the corresponding marketing strategy through collecting and analyzing the information of competitive intelligence system. Meanwhile, competitive intelligence and corporate decision-making is not just a one-way action, continue using the above example, if competitive intelligence system

found that another insurance product was popular in this region, then the managers can determine the next competitive intelligence research point according to this intelligence product: the reasons why this product is popular. This reflects double-side interaction between competitive intelligence and corporate decision-making in marketing research. Enterprises can innovate financial products and services, and find new marketing entry point through this interaction.

3.3 Monitor competitor

The structure of financial market in China is: Medium and large players take dominant position while small players grow rapidly.

In fact, oligarch competition is the fiercest type of competition. In a market with oligarch competition, the top players must keep a close watch on the movements of their competitors in order to obtain information of other players' business, products and marketing tactics and provide corresponding solution based on their own features. By using competitive intelligence, these players can grasp other players' strategy, improve competitive advantages and core skill continuously, improve product innovation and service and develop new business and market.

The development direction of large financial organizations leads the development trend of entire industry. In the first place, a small player should monitor the strategies of large players thoroughly and systematically so as to guide their own strategic objectives. Secondly, a small player should monitor the innovation in business and product of large players and absorb their experience and duplicate their products. Finally, a small player should absorb the experience of large player in technical innovation and risk management in order to survive in the market.

3.4 Assessment of competitive ability

Financial industry is the market where funds converge; it affects all the organizations and individuals in economy. The stability of financial industry relates to the stability of society. Compared to a non-financial enterprise, a financial enterprise has a more profound and lasting influence on society. If financial industry is in trouble, it would not only affect the players in it, but also cause a series of problems in society, economy and politics, such as panic in citizen, run on a bank, crash in stock market, devaluation, unrest in country and descent of housing price. So the assessment of competitive ability is crucially important. A financial company has a large number of branches and complex organization structure, so it must collect information of its network through financial competitive intelligence system to transform the advantages in network, business data and fund into corporation competitive ability. Such activities would improve the ability of corporation management and risk control of financial enterprise in our country. A platform of financial intelligence not only can collect and analyze the data of subsidiaries, branches and departments, but also can realize the share among them, improve collaborative ability and locate the differentiations and core skills of a company.

4 Suggestions on establishing financial competitive intelligence system

4.1 Constitution of financial competitive intelligence system

We may use the basic framework of corporation competitive intelligence system as a reference to develop the framework of financial competitive intelligence system.

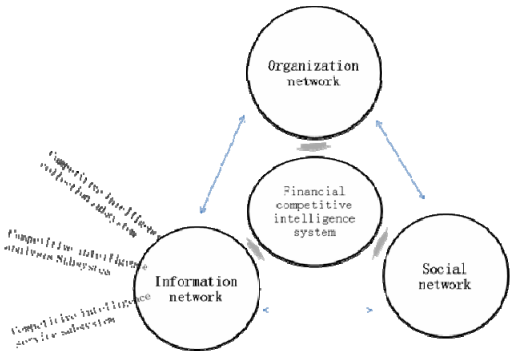


Figure 1. Financial competitive intelligence system

A financial competitive intelligence system includes all factors in decision process of a financial corporation. The factors, including organization network, information network and social network, are showed in Figure 1[3].

Organization network is the core of the financial competitive intelligence system, including the organization of competitive intelligence system, examination and evaluation system, policies and corporate culture. Social network refers to personnel, including intelligence employees and other staff concerned, expanding channels and sources of business intelligence by using a variety of ways and methods. Information network is an intelligence platform based on computer information network system, which could be divided into three subsystems.

A financial competitive intelligence system is an open system, it is able to adapt to different environments. Three factors in this system, including organization network, information network and social network, are not independent to one another. Each factor affects and relates to others.

4.2 Suggestions on establishing financial competitive intelligence system for financial corporations

4.2.1 Change in mentality

The majority of financial corporations in China ha-

978-1-935068-76-1 © 2011 SciRes.

ven't established their financial competitive intelligence system yet. The main reason for this situation is that financial enterprises have misunderstandings on the concept of financial competitive intelligence system. They do not recognize the importance of this system from a fundamental perspective. Competitive intelligence system is a sub-system of MIS. The goal of a competitive intelligence system is to assist a company to keep a sustainable advantage in development. The mechanism of financial competitive intelligence system is showed in Figure 2[4].

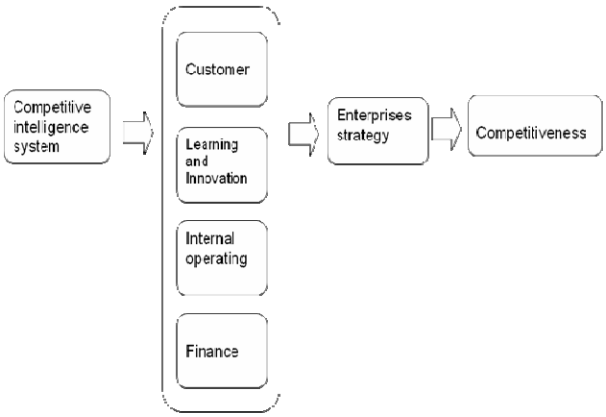


Figure 2. The mechanism of financial competitive intelligence system

An effective competitive intelligence system is likely to help financial corporations to improve their profitability, to cut unnecessary costs and to motivate innovation, in addition, it could also support strategic analysis and fact-based selection of strategy.

4.2.2 Organization of a specialized department

Deployment of the department of competitive intelligence system in a corporation can be 3 types: Dispersed, centralized and focused organization. Dispersed type means functions of the department are scattered into different departments or strategic business unit (SBU). Centralized one means the functions are all located in headquarters, while focused type means the functions are located in the most important department. Financial companies usually have complex organization structure and a large number of departments and branches. Furthermore, the original data often kept in branches which are widely dispersed in geography, so the best choice for a financial corporation would be a combination of dispersed type and centralized type[2,5]. Specifically speaking, this combined type has several characteristics: a) establish the department of competitive intelligence system in headquarters: the responsibility of this department is to collect, analyze and provide competitive intelligence to other SBUs of the financial corporation[6], b) divide the networks of the corporation into several regions, such as east China and north China, and then establish sub-departments in each region responsible for their own competitive intelligence system but should report to headquarters; c) at city or province level, apply a deployment type similar to focused type. Set a pole of competitive intelligence system in the most important department and make it report to upper-level department. Regarding different kinds of financial companies such as banks, insurance companies and security companies, a financial company has to choose different types of deployment according to its specific background and demand. Last but not least, the department of competitive intelligence system should establish good communication and double-sided interaction with other departments.

4.2.3 Establishment of financial competitive intelligence system and talent introduction and training program

Talents and MIS (Management Information System) are the two key factors of establishing a financial competitive intelligence system[7]. MIS of financial competitive intelligence includes collection, sort, release and management of information and also includes hardware, software; it relates to collection of intelligence, analysis and application, and realizes connection, exchange and share of intelligence. MIS is the physical base of competitive intelligence system. As a system of human-computer interaction, financial competitive intelligence system requires a decisive role of human intelligence. Financial competitive intelligence is related to multiple areas such as finance, information and management; it requires inter-disciplinary talents. Financial enterprises need to strengthen the training of inter-disciplinary talents. At the same time, these enterprises need to import talents who understand both competitive intelligence and finance. In a department of competitive intelligence, talents of finance, competitive intelligence and IT coexist. These specialists can exert their strength in different area and realize complementation in knowledge.

5 Conclusion

Competitive intelligence is viewed as the 4th generation of core competitive ability. Competitive intelligence is a must for enterprises to face the competition in the generation of knowledge economy and is an important tool for financial enterprises. In our country, financial organizations haven't established a sound financial competitive intelligence system. Collection of financial competitive intelligence lacks continuity and systematicness, and is unable to integrate scattered information of domestic and foreign economic environment, industry development, competitors and internal affairs. Managers of financial enterprises must change their mind, realize the importance of competitive intelligence, and establish effective and sound financial competitive intelligence system, so these enterprises have to organize a specialized department of competitive intelligence, implement a

professional financial competitive intelligence system, and train and import talents of competitive intelligence.

References

[1] Lu Wei. The Empircal Study on Relation Between Investment Funds Holding and Stock Volatility. Science & Technology Progress and Policy, 2004(8): 76-77.

[2] Ma Haibo. Competitive Information Systems in Commercial Banks in China: their Establishment. Library Journal, 2004, 23(8): 23-25.

[3] Bao Changhuo, Xie Xinzhou. Enterprise Competitive Intelligence System. Huaxia Publishing House. Beijing, China, 2002: 1-263.

[4] Wu Xiaowei, Xu Fuyuan, Wu Weichang. A Competitive Intelligence System Success Model and Its Empirical Study. Journal of the China Society for Scientific andTechnical Information, 2005, 24(4): 473-480.

[5] Luo Ying. On problems about Competitive Intelligence Model of Commercial Bank Based on Value Chain Analysis. Master Dissertation of Tianjin Normal University , 2010:12-16.

[6] LI Yuan, LI Shun, LU Xinyuan. Commercial banks competitive intelligence system:their establishment and application. Researches in Library Science, 2009(6): 99-100.

[7] YU Haojin. The Build-Up and Application of Competition-Based lnformation System of Commercial Banks. Finance Forum, 2006(3): 44-55.

978-1-935068-76-1

A Bibliometric Analysis of Scientific Production in Hepatitis C Research

Xiaoli TANG, Jian DU, Taotao SUN, Yanwu ZHANG

Institute of Medical Information / Medical Library,

Chinese Academy of Medical Sciences/Peking Union Medical College, Beijing, China, 100005

Email: tang.xiaoli@imicams.ac.cn

Abstract: In recent years, treatment for hepatitis C virus (HCV) has attracted much attention in medical sciences. This paper intends to explore the development of the field of research utilization in the field of hepatitis C, and to identify the structure of this scientific field. The total number of international SCI papers on hepatitis C showed a linear growth trend. The proportion of the papers from China was increasing steadily, but the overall proportion was quite small in the world. The international cooperating network in hepatitis C research field could clearly be divided into two sub-networks. Hot spots of a field of academic study or research can be found in analysis of the keywords map.

Keywords: Hepatitis C; HCV; bibliometric; mapping knowledge domains

1 Introduction

Hepatitis is an inflammation of liver, usually due to viral infection but sometimes due to toxic agents. Previously endemic throughout much of the developing world, viral hepatitis now ranks as a major public health problem in industrialized nations. The three most common type of viral hepatitis (A, B, and C) affect millions worldwide[1]. According to the World Health Organization (WHO), the global infection rate of hepatitis C is 3%, with an estimated 170 million carriers of HCV. In China, 3.2% of the general population is anti-HCV positive, and nearly 40 million patients, namely one out of 30 people, are suffering from hepatitis C. Among the hot topic of scientific research in 2011 predicted by an article from *Nature* published on January 6, 2011, *therapy on hepatitis C* is highlighted to be one of a new hot issue of biologic research. The public eagerly anticipated drug approvals in 2011 include a decision by the US Food and Drug Administration on telaprevir, which could provide relief for the 3% of the world's population infected with the hepatitis C virus[2].

Bibliometrics is the study dealing with the quantification of written communication which helps in the measurement of the published knowledge. Bibliometric analysis provide insight into the growth of literature , interrelationship among different branched of knowledge, productivity, authorship, degree of collaboration, and the flow of knowledge within a specified field of academic research. Findings have shown bibliometric methods to be one valid and reliable way of mapping the development and structure of a scientific field[3]. The bibliometric method uses empiric data and quantitative analysis to trace formal communications in the form of published literature, and to study the patterns of publication within a field. A key assumption underpinning this method is that research papers represent knowledge produced by scientific research[4]. Bibliometric mapping of science and technology is the method to visualize the field of knowledge and it is accomplished by creating landscape maps[5].

Research utilization is the use of research to guide clinical practice. However, little is known about the characteristics of the research utilization literature in the field of hepatitis C, including the development and organization of this field of study. In this paper an attempt has been made to identify the publication growth and collaboration research on hepatitis C. The article addressed the knowledge gap in this field of study by bibliometrically analyzing the research utilization literature in this field. Based on the literature data collected in the online ISI Web of Science database (WoS), this paper aimed to map the development of the field of research utilization in the field of hepatitis C, and to identify the structure of this scientific field, including the network of international collaboration, research fronts, etc. Thomson Data Analyzer (TDA) and Citespace software were used to do the bibliometric analysis.

2 Data and methods

As data sources for this study we used information resources the Science Citation Index (SCI) and the Social Sciences Citation Index (SSCI) databases, produced by ISI- the Web of Science (WOS), with type of publications restricted to articles and reviews. The time span was set as 1976-2010 , in order to discover the initial status of international research in the field of hepatitis C, and acquire a systemic view of the development and progression of the theme. The document retrieval was conducted on December 17th, 2010. The retrieval strategy was: Theme=(("hepatitis C" OR "hepatitis virus C") OR (HCV NOT "hog cholera virus")) AND Type=(Article OR Review). There are in total 37,418 literature records were

retrieved from WoS.

TDA, a powerful text mining tool to analyzed data from a variety of rich information sources, was applied to clean and analyze the literature data on hepatitis C.Beyond traditional bibiliometric analysis, this paper has done a co-word analysis on the development and progression in the field of hepatitis C by using the CiteSpace ,an online version of knowledge mapping analysis software, designed by Chaomei Chen, a Professor of Information Science and Technology College in Drexel University .We set coefficient appropriate adjustments of three parameters(C,CC,CCV) to carry out co-occurrence analysis of key words, analysis of co-citation documents. Hot spots of a field of academic study or research can be found in analysis of the keywords map, and the transfer trends of researching hot spots may be found in it. In this article, CiteSpace software is used to draw out the knowledge map of research field of hepatitis C to analyze the study situation of hepatitis C.

3 Results and Discussion

3.1 Data overview

In order to reveal the initial status and development trends of scientific research on hepatitis C, we analyzed the number of articles published worldwide and those by Chinese as well as its proportion over the past 30 years.

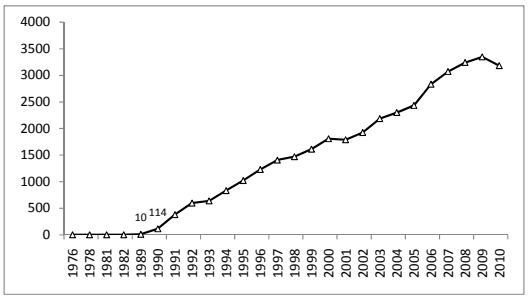


Figure 1. Distributions of world papers on Hepatitis C

Articles published worldwide on Hepatitis C showed a straightly increasing trend from 1989-2009 as in Fig 1. The first paper indexed by SCI was "Role of Hepatitis C and Hepatitis B in Transfusions", published on the INTERNIST in 1976 by FROSNER G[6]. In fact, in the early 1970s, researchers had already indentified the concept of "Non-A, Non-B Hepatitis", which was initially used by Lancet in 1975. It was worth noting that articles grew from 10 in 1989 up to 114 in 1990, which was due to the discovery of HCV with molecular cloning techniques by Chiron Company in 1989[7], and the confirmation of its pathogenic role in causing chronic and infectious NANB hepatitis during the International Symposium on NANB Hepatitis and Blood-borne Infectious Diseases in Tokyo, Japan. Since then an increasing number of researches had sprung up.

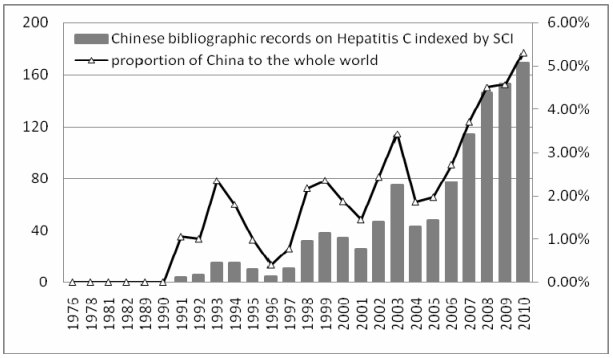


Figure 2. Distributions of Chinese papers and its proportion to the whole world on Hepatitis C

The first China's article indexed by SCI was "Investigation of Anti-HCV in 391 Serum Samples in China", published on CHINESE MEDICAL JOURNAL in 1990 by Professor Tao Qimin[8], a famous hepatologist previously working in People Hospital of Beijing Medical University. Since then, the number of papers showed a fluctuated increasing from 1991 to 2003, and a straightly increasing trend after 2004. The proportion of bibliographic records on HCV published by China to the whole world was more periodically increasing. Nevertheless, the ratio was quite small, merely 5.3% in 2010.

3.2 International Collaboration

Hepatitis C treatment is a common problem internationally in medical community, thus requiring more extensive international collaboration. It is one of the key findings and events that could emerge from the research world in 2011 predicted by Nature. In order to provide references for the consolidation of international cooperation and the selection and introduction of excellent researchers or laboratory equipments in the field of HCV, we analyzed the cooperation networks of the major countries (regions) and the institutions involved in the collaboration between China and other countries.

3.2.1 Collaboration network of the major countries

The cooperation network among the top 22 countries or regions published more than 300 articles during 1967-2010 was shown in Fig.3. Ranking by the number of papers, the top 10 countries are United States, Japan, Italy, France, Britain, Germany, Spain, Canada, Australia and China. These 22 countries formed 2 distinct networks. The first one, with United States as its center, consisted of the Americas (Canada), European and Asian countries. The central position of United States in the network indicated its key role in the field of HCV. United States and Japan have maintained the closest cooperation, followed by Canada, Germany, Spain, Sweden and Switzerland. The second, namely the "European group", consisting of Italy, France and the United Kingdom revealed a more intensive collaboration. However, apart from them, other countries could not form a distinct network. Although

978-1-935068-76-1 © 2011 SciRes.

Australia and China are among the top 10 countries, the degree of their participation in the international collaboration remained to be extended. Many Asian or developing countries (regions), such as South Korea, Egypt, India, Taiwan, Brazil, Greece and Turkey, are at the margin of the network. But compared with India and Brazil, China has a relatively more extensive and intensive collaboration internationally.

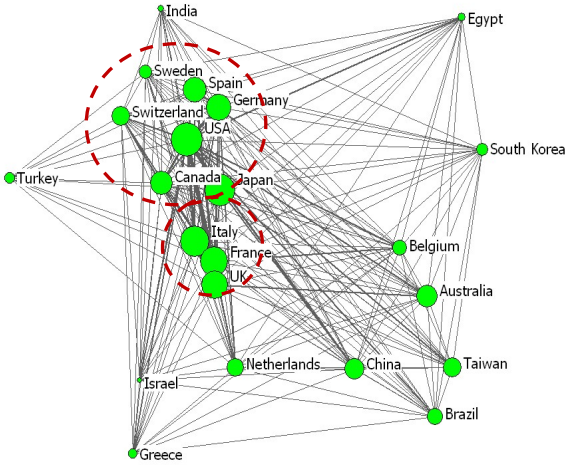


Figure 3. Cooperation network of the major countries with more than 300 articles on HCV

3.2.2 Collaboration network of the major institutions

The cooperation network among the top 36 institutions which is involved in research collaboration among China and other countries, with more than 6 articles on HCV during 1990- 2010 was revealed in Fig.4. The thickness of the line connecting two nodes represented the cooperation degree. The thicker the line, the more cooperative papers published.

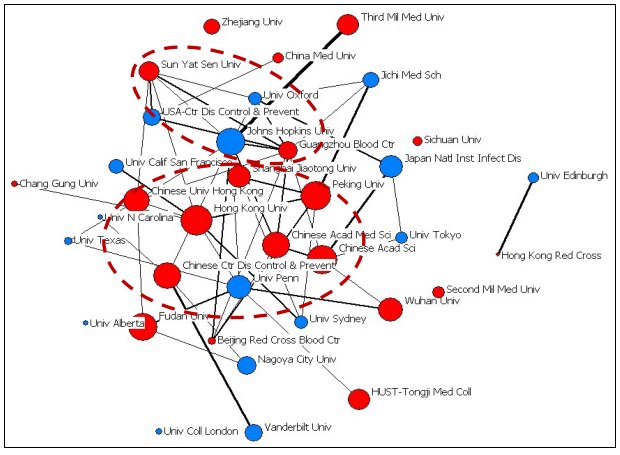


Figure 4. Collaboration network of the major institutions involved in research collaboration among China and other countries

Fig. 4. revealed that among the top 36 institutions involved in the international collaboration in China on HCV, 16 of those were located in mainland China, 7 in the US, 4 in Japan, 3 in Hong Kong, 3 in the UK, and 1 in Australia, Canada and Taiwan respectively. The center of the cooperation network consisted of 7 scientific institutions, namely Hong Kong University, Chinese University of Hong Kong, Shanghai Jiao Tong University, Peking University, Chinese Academy of Medical Sciences, Chinese Academy of Sciences, and Chinese Center for Disease Control and Prevention, indicating that China could play a central role in the international collaboration on HCV. The cooperation degree between Shanghai Jiaotong University and Johns Hopkins University is the most closed, followed by the Third Military Medical University and Johns Hopkins University. Located in Guangdong Province, SunYat-sen Univeristy and Guangzhou Blood Center also formed a distinct triangular cooperation with the US CDC. The partnership between China CDC and Vanderbilt University of US, or collaborations among University of Pennsylvania, Wuhan University and Fudan University were also clearly. However, Hong Kong Red Cross and University of Edinburgh yet established cooperative relationship each other outside the major collaboration network. Analyzing the cooperation condition of these institutions could provide valuable information and established “bridges” for the selection and introduction of overseas excellent intellectuals. Fudan University had more articles published through international collaboration, but only occupied a marginal area in the network, indicating its unilateral cooperation character, which was also embodied by Huazhong University of Science and Technology, Wuhan University, Second Military Medical University, Sichuan University, Third Military Medical University, China Medical University, Zhejiang University, and Sun Yat-sen University. By contrast, international cooperation of Shanghai Jiaotong University was more extensive.

3.3 Development of the main research fields

According to the distribution of world bibliographic records on hepatitis C, we divided years from 1976 to 2010 into 4 slices, namely 1976 to 1995, 1996 to 2000, 2001 to 2005 and 2006 to 2010. Using CiteSpace software, we carried out co-word cluster analysis of top 40 most occurred keywords during each slice in order to reveal the main research topics during each of them, as well as the development and evolution of hot spots and subjects on HCV.

In the slice of 1976-1995(Fig.5.), the biggest nodes were non-a, non-b-hepatitis and hepatitis-c virus, indicating that researchers were focusing on the identification of hepatitis C and the discovery HCV. What’s more, HCV genome, macro epidemiology of HCV and α-interferon for HCV treatment were also the hot research spots during this period. In 1996-2000(Fig.5.), the cluster formed

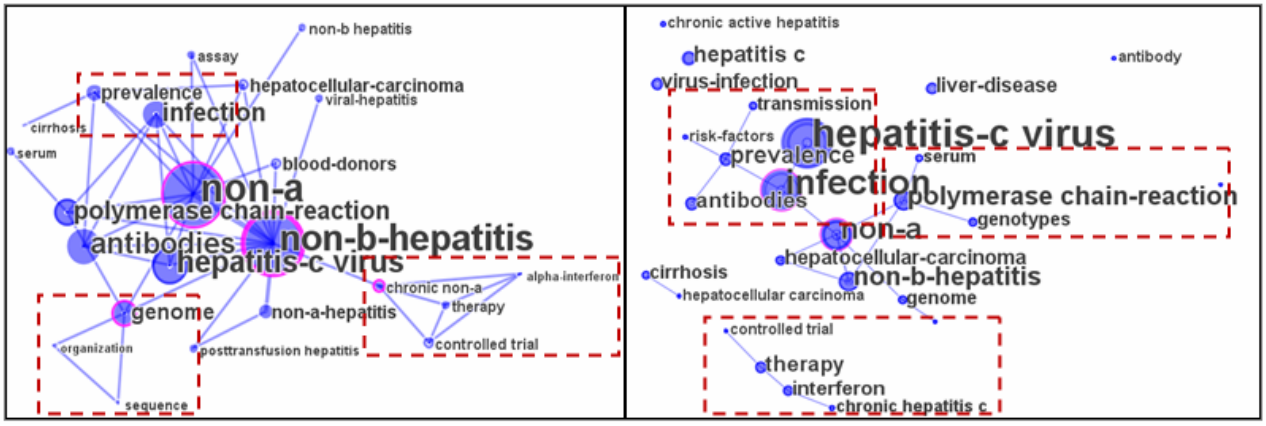


Figure 5. Co-word cluster analysis chart of top40 keywords in 1976-1995 and 1996-2000

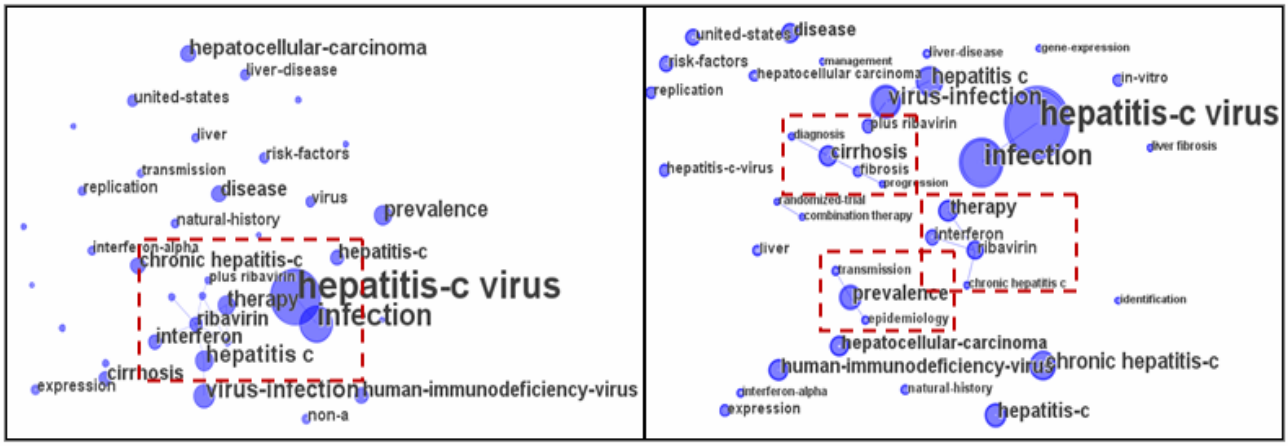


Figure 6. Co-word cluster analysis chart of top40 keywords in 2001-2005 and 2006-2010

by 6 keywords including hepatitis-c virus, infection, prevalence, antibodies, transmission and risk-factor located in the center of network, which demonstrated that researchers during this period paid more attention on the occurrence, transmission and spread of HCV, as well as the screening of risk factors and the prevention and treatment with antibodies. We could also name it as the clinical epidemiology of hepatitis C. In addition, network consisted of polymerase chain reaction, serum and genotype showed that genotype of HCV RNA was being furtherly analyzed and indentified. Considering the HCV genotype was clear between 1976 and 2000, these studies were making important progress on detecting the severity of HCV infection. Other information gained from the clusters in the lower part of the right figure is that interferon treatment of HCV continued to be a hot research spot. Controlled trials are carried out to find the best treatment methods.

Towards the new century, research topics were dispersed from 2001 to 2005(Fig.6). The treatment of HCV continued to be the main research field, but the core subject had converted to combined treatment with interferon and ribavirin. However, research subjects gradually centralized during 2006-2010(Fig.6.), which could be divided into several sub-fields. For example, more attention had been paid on the combined treatment with interferon and ribavirin, epidemiology of HCV, the relationship between HCV and cirrhosis, hepatic fibrosis became major topics again.

Quantitative analysis indicated that the research topics on hepatitis C showed significantly centralized trend, that is, during 1976-1995, the recognition of hepatitis C and the discovery of HCV were the first two centers, followed by genome of HCV, genotype of HCV, macro epidemiology of hepatitis C and α-interferon treatment for HCV. During 1995-2000, the studies were further deepened, focusing on clinical epidemiology of HCV and quantitative detection of different genotypes of HCV-RNA. Interferon treatment was a continuing hot research topic. Research topics during the first five years of the new century were dispersed, but the treatment of HCV continued to be focused. During the recent 5 years, sub-

978-1-935068-76-1 © 2011 SciRes.

jects of researches had been centralizing. More attention had been paid on the combined treatment. What's more, the relationship between HCV and insulin resistance, crystal structure analysis of HCV protease were recently occurred hot topics.

3.4 Research frontier analysis based on burst terms

In order to identify the words with a high concentration and high density characteristics in archives of literatures, we used CiteSpace to carry out burst terms detection on keywords. Time distribution of frequencies was analyzed as a criterion to extract words whose frequencies changed and increased rapidly from most commonly used words. The frontier disciplines and developing trends were analyzed not only by the degree of frequency but also by the trends of frequency.17 burst terms with maximum burstness about hepatitis C were shown in Table 1.

The frequency of 7 terms such as 'non-a', 'randomized trial', 'non-b-hepatitis', 'polymerase chain-reaction', 'b virus', 'antibodies' and 'genome' dropped during 2001-2004. However, some terms such as 'sustained virological response', 'insulin-resistance' suddenly appeared in 2010 which indicated the possibly new research field in the future. The frequency of crystal-structure first decreased and then increased during 2005-2010, which indicated a possibly new research field as well. In 2006, the research on crystal-structure of NS2-3 protease published in Nature was a significant progress [9]. The structure showed that NS2 is a kind of cysteine protease dimer, which might be helpful to clarify the function of NS2 in the lifecycle of the virus, as well as to design new drugs. According to the association between insulin resistance and HCV, a latest research in 2010 demonstrated the correlation between clearance of HCV RNA levels and improvement of IR which provided additional in vivo support for a causal role of HCV in the development of IR[10].

The most promising research frontier in HCV research was sustained virological response rate relating to the treatment of HCV patients. The research focus in this aspect has changed in the past few years. Firstly, it was combination therapy of PEG-INF and ribavirin, and then it focused on HCV NS3/4A Serine protease inhibitor and NS5B RNA dependent RNA polymerase inhibitor. After that it focused on combination therapy of HCV protease and polymerase inhibitors. And recently it highlighted in limiting the occurrence of HCV drug-resistant strand, minimizing the adverse effect and other combination therapies or strategies that still need scientific verification. Nature predicted telaprevir, which was a kind of HCV NS3-NS4A protease inhibitor, would be approved by US Food and Drug Administration as an eagerly anticipated drug in 2011. The frequency of interferon-alpha-2b plus ribavirin did not changed during 2004-2005 but decreased in 2006, which indicated that new combination therapies might be developed for example PEG-INF plus ribavirin. However, combination therapy of interferon plus ribavirin still continued to be a hot topic.

Table 1. Burst terms on hepatitis C

Keywords (Ranked by the value of burstness)	*Year(Grey cells shows the change of the frequency)*										Trends
	01	02	03	04	05	06	07	08	09	10	
non-a	■	■	■	■							decreasing
randomized trial			■	■							sharply decline
non-b-hepatitis	■	■	■	■							decreasing
polymerase chain-reaction	■	■	■	■							decreasing
b virus	■	■	■	■							sharply decline
sustained virological response										■	burst in
antibodies	■	■	■	■							decreasing
genotypes	■	■	■								fluctuated increasing
initial treatment			■	■	■						fluctuated decreasing
genome	■	■	■								decreasing
insulin-resistance										■	burst in
combination		■	■	■							decreasing
interferon-alpha-2b plus ribavirin				■	■	■					decreasing
randomized-trial						■	■				decreasing
antiviral therapy								■	■	■	decreasing
crystal-structure			■	■	■	■	■	■	■	■	increasing
steatosis							■	■	■	■	fluctuated decreasing

4 Conclusions

Through quantitative and content analysis on hepatitis C, we displayed the history, current situation, and especially the international cooperation network. Moreover, main research fields and the development and evolution during different periods were revealed. Hot topics and research fronts were discussed as well. The main results are as follow:

(1) The number of international SCI papers on hepatitis C showed a linear growth trend. In 1989, American company Chiron first discovered HCV by molecular cloning techniques which showed that HCV was the major pathogen causing chronic and infective non-a, non-b hepatitis. It was a landmark in this field and attracted broad attention so that the number of papers and researches on HCV increased rapidly. The proportion of the papers from China was increasing steadily, but the

overall proportion was quite small in the world.

(2) The international cooperating network in hepatitis C research field could clearly be divided into two sub-networks. The first was a network centered in America involving some developed countries such as America, Japan and Germany. Another was a small European group consisted with Italy, France and Britain. The numbers of papers from Australia and China were also large, but international cooperation was not enough. However, compared with India and Brazil, China had a relatively more extensive and closer international cooperation. Chinese institutions could play as centers in international cooperation in hepatitis C research fields, showing stronger foreign cooperation ability.

(3) The main research fields had undergone the following development and evolution from 1976 to 2010: during 1976-1995, the recognition of hepatitis C and the discovery of HCV were the first two most significant research fields, after which more researches on genome of HCV, genotyping of HCV, macro epidemiology of hepatitis C and α-interferon treatment were carried out. During 1995-2000, more intensive studies were conducted which focused on clinical epidemiology of HCV and quantitative detection of different genotypes of HCV-RNA. The best therapy of interferon treatment was a continuous hot topic. Subjects of researches during 2001-2005 were dispersed, but the treatment of HCV continued to be focused. During the recent 5 years, subjects of researches were centralizing. More attention had been paid on the combination therapy of interferon plus ribavirin. The epidemiology of HCV, the relationship between HCV and liver cirrhosis, fibrosis became major topics again.

(4) How to increase sustained virological response in HCV treatment were important research fronts. In addition, the relationship between HCV and insulin resistance, crystal structure analysis of HCV protease were recently appeared hot topics as well.

References

[1] Stedman. Stedman's Medical Dictionary. 27th ed. 2000. Lippincott; Baltimore. pp 808

[2] Richard Van Noorden , Heidi Ledford & Adam Mann. New year, new science. Nature looks at key findings and events that could emerge from the research world in 2011. [accessed on February 4,2011]. http://www.nature.com/news/2010/101231/full/4690 12a.html

[3] Lievrouw, L. A. (1989). The invisible college reconsidered: Bibliometrics and the development of scientific communication theory. Communication Research, 16(5), 615-628.

[4] Okubo, Y. (1997). Bibliometric indicators and analysis of research systems: Methods and examples. STI Working Papers Paris: OECD.

[5] Noyons E D(2001).Bibliometric mapping of science in a Science Policy context,Scientometrics,50(1),83-98

[6] Frosner G(1976). Role of Hepatitis C and Hepatitis B In Transfusions[J]. Internist,17 (5):256-257

[7] COLOMBO M, KUO G, CHOO QL, et al(1989). Prevalence of Antibodies to Hepatitis-C Virus in Italian Patients with Hepatocellular-Carcinoma[J]. LANCET, 2(8670): 1006-1008

[8] TAO QM, WANG Y, WANG H, et al(1990). Investigation of Anti-HCV in 391 Serum Samples in China[J]. CHINESE MEDICAL JOURNAL,103(8):616-618.

[9] Lorenz IC, Marcotrigiano J, Dentzer TG, et al(2006).Structure of the catalytic domain of the hepatitis C virus NS2-3 protease [J]. Nature, 442(7104):831-835.

[10] Delgado-Borrego A, Jordan SH, Negre B, et al(2010). Reduction of insulin resistance with effective clearance of hepatitis C infection: results from the HALT-C trial[J].Clin Gastroenterol Hepatol, 8(5):458-62.

978-1-935068-76-1