Designing and Implementation of Expertise Search & Hotspot Detecting System

Wu Chen, Wei Lu, Shuguang Han
Center for Information Resource
Wuhan University
Wuhan, China

***Abstract*—Modern scientific researches are facing two kinds of problems: how to select proper research themes, how to get related experts and works. Expert search and hotspot detecting are two methods used broadly to settle these problems. Expert search can help resolving the problems of expertise retrieval, expert evaluation, collaborator choosing and so on in particular research fields, while hotspot detecting is very important for researchers to identify research fronts and promising themes in these fields. We first summarize the present researches in the world of these two issues, and then design an expertise search & hotspot detecting system which can identify experts' and institutions' expertise automatically, and also detect research hotspots and trends dynamically in particular fields. At last, we implement the system with traditional medicine as the selected field.**

Keywords-component; expertise identification; hotspot detection; traditional medicine

I. INTRODUCTION

Multi-regional or cross-border co-operation are becoming more and more important in scientific researches nowadays. This brings along some challenges such as how to evaluate researchers' expertise, how to select reliable collaborators and etc. On the other hand, how to determine appropriate research themes in a particular field? Because of research's continuity, researchers have to find related works, experts and institutions to a theme, but how to do this automatically? These problems' resolution is very important to the scientific development and research progressing. Expert retrieval and hotspot detection methods are usually used to solve these problems internationally. This is exactly what we do in this paper. By designing an expertise search & hotspot detecting system we do this in an automatic manner.

The second section of this paper will introduce some related works, and the third is about thoughts in constructing this system. The system's architecture and detailed implementation will be presented in section four. At last we conclude our work and give some possible future directions according to the shortcomings of our system.

II. RELATED WORK

Expert search ranks experts according to their relevance to a given query (or topic), using various documents and resources which represent experts' expertise. It is different from traditional information retrieval because it returns expert list relating to a topic rather than document list. Expert search is one of the hotspots in vertical information retrieval field. TREC (Text Retrieval Conference) started enterprise search track in 2005 with expert search as one of the submissions. Some of the research did by the participants represent the latest progress of this special field [1]. And also there are some commercial expert search systems already in use, such as SmallBlue [2] (IBM), People Finder [3] (CSIRO). One typical expert search method is as below: first search for documents relevant to a given query using traditional information retrieval methods, then merge scores of documents related to an expert as this expert's relevance to the query. Finally, we could get a list of experts ranked by their relevance. Using this method, we had designed and implemented an expert search system [4] with Wuhan University as an example(WHU-ES).

On the other hand, bibliometrics and co-citation analysis methods are broadly used to detect research fronts and hotspots in a specific field [5-7]. If the quantity of research results related to a specific theme is more than others, we can infer that this theme draws more attention and might be a hotspot. And even going further, co-citation analysis for frequently cited articles and then social network maps drawing, can help searchers getting a better understanding of the linkages between research topics and grip of hot research themes [8]. In the light of the fact that keywords contain plenty of semantic information, word frequency analysis and co-word analysis are significant methods to hotspot detection [9]

III. SYSTEM CONSTRUCTION

This paper tries to identify researchers' and institutions' expertise intelligently and detect research hotspots and trends of a specific field dynamically, by analyzing researchers and institutions' research results, especially articles. It also tries to realize uniform retrieval and knowledge management in a specific field which can provide information about other researchers and also research topics for researchers.

Commercial systems like SmallBlue and People Finder although provide function to rank experts, they doesn't do a good work in mining the linkage between experts. This article uses visualization methods and social network analysis to represent the networks of experts using their co-occurrence in

Supported by grants from the National Natural Science Foundation of China (No. 30370229), the Humanities and Social Sciences Project of Ministry of Education of China (No. 09YJA870021) and the National "985" Project of China (No. 985YK003).

articles. While for hotspot detection, in spite of traditional documents classification and statistics functions, the designed system will be able to draw trends curves. Established classification system has a shortcoming as it cannot change in real time, this new function overcomes it. And it can also show particular theme's trends.

With comprehensive consideration of similar systems' functions and users' need, the system designed in this paper can be departed into four functional modules: expertise retrieval module, navigation module, hotspot detecting module and data maintenance module. The framework of these modules is shown in Figure 1.

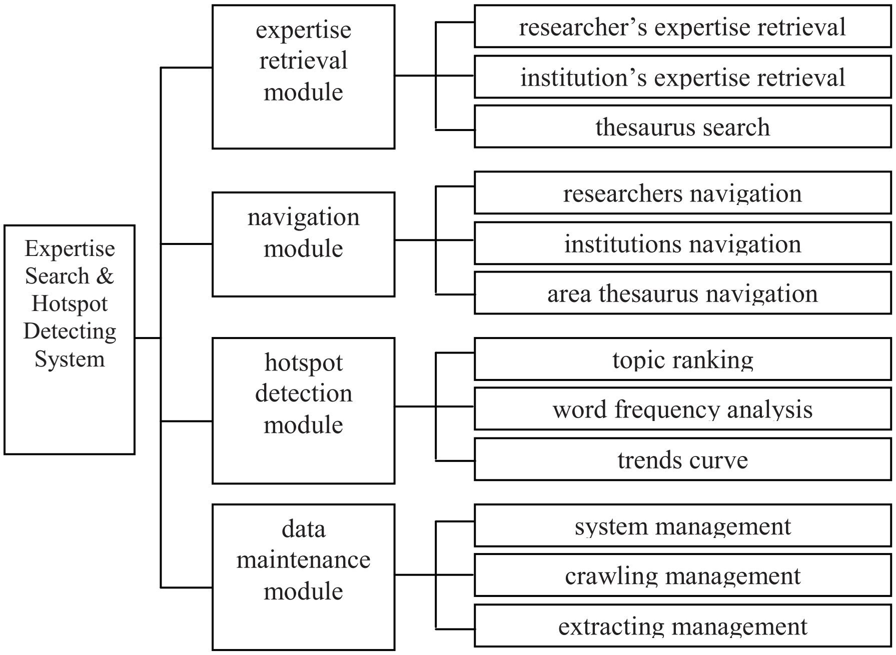


Figure 1. Framework of system modules

A. *Expertise retrieval module*

Users often hope to get a list of experts or institutions ranked by their relevance to a topic by submitting a query to the system. And they may want to be able to browse experts' and institutions' personal information such as published articles, web documents, specialty in every aspect, and also social network maps. This module is design for this user need. It can be divided into three sub-modules: researcher's expertise retrieval, institution's expertise retrieval, thesaurus search. The first two is the same with a small difference in returning results. Researcher's expertise retrieval returns researchers related to a topic while the other returns institutions. Thesaurus search sub-module return extra keywords semantically similar to user provided query, and this can help users expanding and properly choosing query words.

B. *Navigation module*

User activities can be divided into two categories: retrieve and browse [10], according to the manner they use a retrieval system. User might not only want to search a specific researcher or institution, but also want to browse all researchers and institutions. This module is designed to let users be able to get researchers' and institutions' information in browsing manner. By selecting a researcher of institution, users can get information about it just as expertise retrieval module does. There is a small by significant difference. Social network map will not be restricted to experts (institutions) related to a specific topic as the first module dose, just because the user doesn't provide a topic. The map will contain all experts (institutions) who are related to the selected expert (institutions).

C. *Hotspot detection module*

Goal of this module is automatically identifying research hot topics, representing trends, and detecting potential hotspots. It contains three sub-modules: topic ranking according to their hotness degree, word frequency analysis under each topic, and trends curve of topics. The first sub-module computes quantity of articles related topics. It first categorizes all the articles in a specific field into a number of topics, ranks these topics by their article numbers, and creates an article number curve for each topic. The second module performs word frequency analysis under each topic. The last module creates two time series maps by computing user provided query's occurrence every year. The first map's y-axis is articles' absolute number and the second relative number, while combining the two maps can overcome impacts of total article number's difference every year.

D. *Data maintenance module*

The module provides system administrators with a web interface to maintain and modify the system, such as change data repository, modify thesaurus, set data crawling cycle, modify the map between institutions' formal names and frequently-used names.

IV. System principle and complement

As a crystallization of human wisdom, traditional medicine is receiving more and more attention from researchers. Undoubtedly, informationization in this field will promote its development tremendously. This is why we choose this field as an example to implement our expertise search & hotspot detecting system (referred to as Tr-Med for short).

A. System Principle

When implementing Tr-Med, the problems of getting lists of all researchers and institutions as well as their papers in traditional medicine field must be solved. Tr-Med crawls academic papers in traditional medicine automatically with the help of thesaurus (MeSH), and then extracts researchers' and institutions' names from certain fields of these papers. MeSH is used to index, categorize and search medicine related documents as an authoritative thesaurus in medicine field, which is worked out in 1960 by NLM (The United States National Library of Medicine). Web of Science, Medline databases contain academic papers which can best represent the latest progress in medicine field. We first select some keywords from MeSH manually which are related to traditional medicine, and then use them as query words to get records from Web of Science and Medline databases. Finally, we extract researchers' and institutions' names from crawled records and normalize them. The problem mentioned above is settled.

As for expertise identification and retrieval, Tr-Med takes advantage of a two-stage method [11]. It is described below: given a query, first get related documents, and then for each expert/institution merge its related documents' scores [12]. Detail description can be found in reference [4].

Tr-Med takes paper quantity and word frequency as indicators of topic's hotness degree. Lager quantity of them means the related topic is hotter, while higher growth rate of them means the related topic is potentially hot. On one hand, Tr-Med creates a data cube of records with three axes: published year, author country, topic. Users will be able to get certain topic's hot degree based on this cube. On the other hand, time series map is created base on occurrence number in certain fields of records such as title, abstract and keyword. These maps provide a better understanding of the future research trends and the trend of a particular topic.

B. Implementation

Tr-Med's implementing flow is shown in Figure 2. Detailed progress can be divided into five steps: data crawling, information extraction, experts database and institutions database construction, hotspot analysis preprocessing, user interface designing and retrieval results visualization. This section will show the detail of each step.

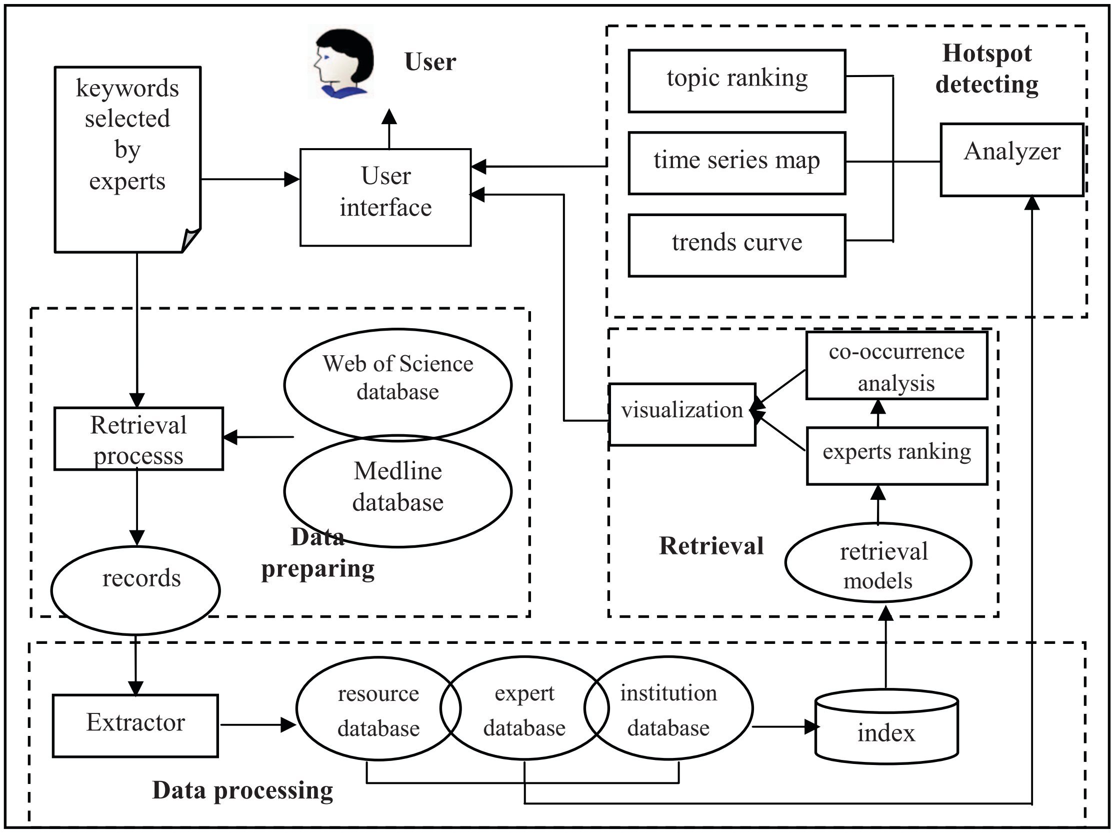


Figure 2. System implementation flow

1) Data preparing

To accomplish the first data crawling of our system, 139 keywords are selected from MeSH by experts of traditional medicine manually. And then we use these terms as queries to search articles from Web of Science and Medline database with a time limit from 1995 to 2009. Finally we get 107824 records. Of course data preparing must be done every certain time after Tr-Med established. This can be done by users with the help of data maintenance module.

2) Information extraction

Records crawled from Web of Science and Medline databases are in HTML format. Regular expression is employed in this step to extract contents of certain fields such as Title, Author, Address, Source, Doc Type, Abstract, Keywords, ISSN, Published and etc.

3) Experts database and institutions database construction

Experts database construction is based on contents extracted from Author field of crawled records. The raw extracted names are usually in abbreviated form. So problems might exist because some certain names may be in the same abbreviated form, and one name can be abbreviated as different forms. In the first situation, Tr-Med treats two researchers' names as different if their institutions are different, even if their names' abbreviated forms are the same. In the other situation, the problem is solved with the help of users' feedback. Users can merge names with different abbreviated forms if they are certain that these names refer to one same person, in the data maintenance module. Institutions database construction is completed with three sub-steps. First, process extracted content of Address field and store string within a certain range of words like University, Association as an institution name. Second, construct a map between normalized institution names and common names. Normalize names extracted using this map. Finally, store institution names into database.

4) Hotspot analysis preprocessing

As mentioned above, a data cube is created in Tr-Med. The first axis is published year. Value of this axis is extracted from Published field of each record. The second axis is author country. Its value is the first author's country. The last axis is topic and is of most importance. We assume that a paper's topics are represented by query words which retrieve it out. So, if paper A is retrieved out by both Pharmacognosy and Ethnopharmacology, A's topics are G01.273.118.598, G01.703.015 and H01.158.273.118.598 (see Figure 3).

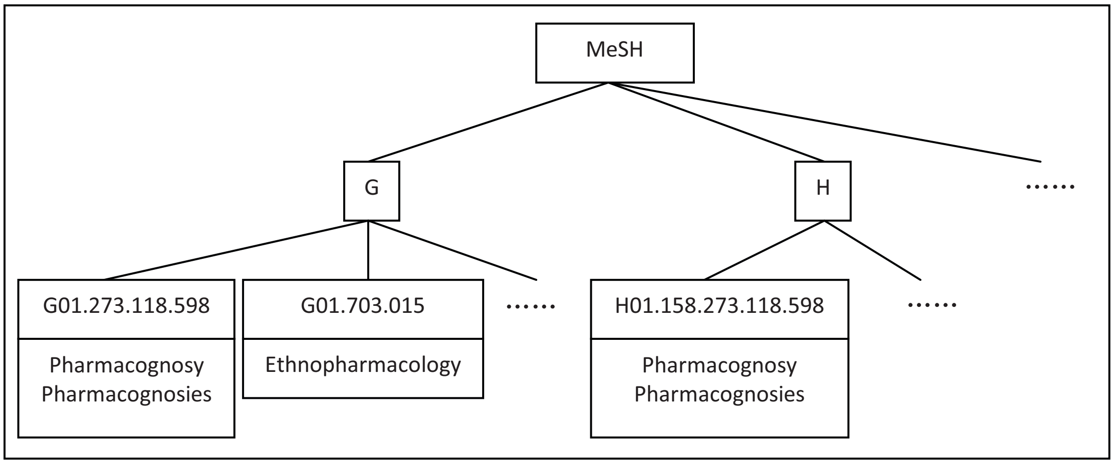


Figure 3. Part of MeSH

5) User interface designing and retrieval results visualization

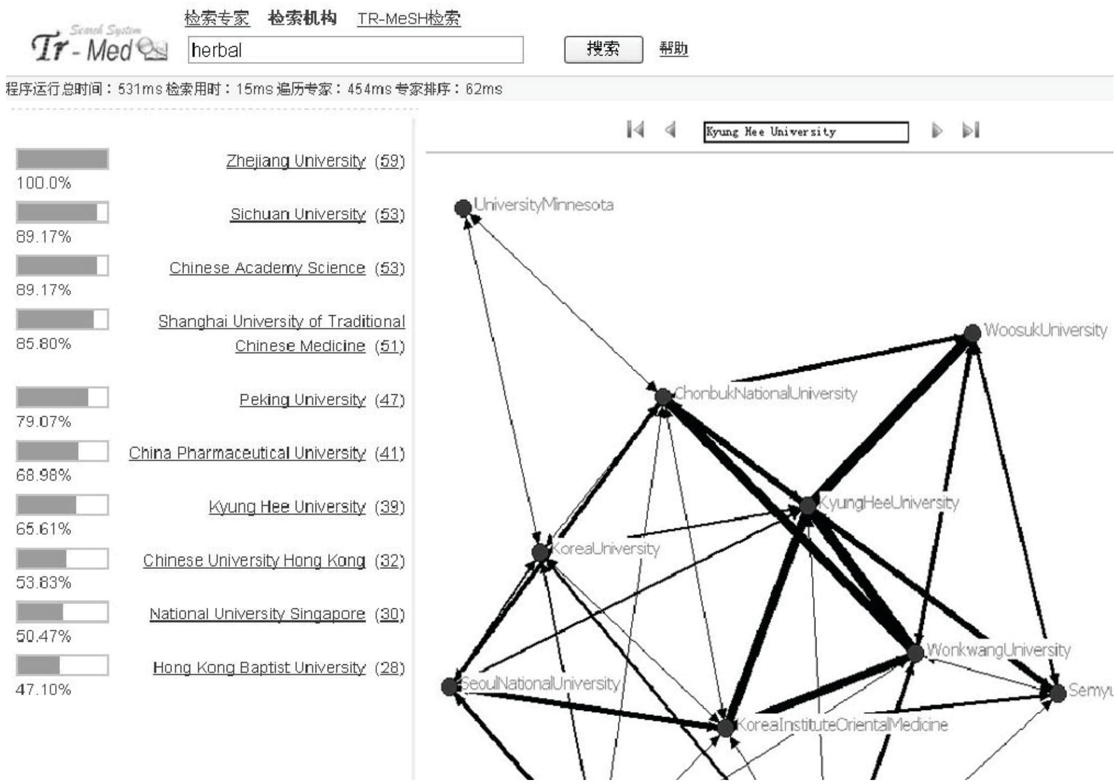
Tr-Med provides three kinds of user interface to users: retrieval interface, navigation interface and hotspot detecting interface. Users submit a query to the system by retrieval interface, and get result of experts or institutions list as well as their social network map. Figure 4 shows the result of query word "herbal". The left column is the list of institutions, while the right column is their social network map created by NetDraw [13]. In this map, dots represent institutions and lines represent linkage between each pair. The bolder the line is, the closer the two institutions are.

By limit published year and country, users get a ranked list of topics within this restricted range with hotspot detecting interface. After clicking a topic in the left column, Tr-Med returns two time series maps representing this topic's development over all the years. This is done by invoking JFreeChart [14].

V. CONCLUTION

This paper designs an expertise search & hotspot detecting system, and implements it in traditional medicine field.

Shortcomings still exist in Tr-Med. In the future, we will pay more attention on crawling web data, patent data and so on. Optimizing the system's ability to normalize experts' and institutions' names will be done too. When analyzing word frequency, we found some words with very broad meaning. We are thinking of assigning weights to every word according to their specificity so that the result could be more comprehensive.



程序运行总时间：531ms 检索用时：15ms 遍历专家：454ms 专家排序：62ms

100.0%	Zhejiang University (59)
89.17%	Sichuan University (53)
89.17%	Chinese Academy Science (53)
85.80%	Shanghai University of Traditional Chinese Medicine (51)
79.07%	Peking University (47)
68.98%	China Pharmaceutical University (41)
65.61%	Kyung Hee University (39)
53.83%	Chinese University Hong Kong (32)
50.47%	National University Singapore (30)
47.10%	Hong Kong Baptist University (28)

Figure 4. Retrieval interface

REFERENCES

[1] http://trec.nist.gov/pubs/trec15/papers/cwi.ent.final.pdf.

[2] Ching-Yung Lin, Ehrlich.K, Griffiths-Fisher.V, et al. SmallBlue: People Mining for Expertise Search. Multimedia, IEEE,2008,15(1), pp.78-84.

[3] http://hal.archives-ouvertes.fr/docs/00/03/54/04/PDF/ADCS-03.pdf.

[4] Lu Wei, Han Shuguang. Design and Implementation of Organization Expert Search System. Journal of the China Society for Scientific and Technical Information, 2008,27(5), pp.657-663.

[5] Qiao Wenming, Suo Dawu. Bibliometric Analysis of the Subject Distribution-based Papers on Information Studies in China . Information Studies: Theory & Application, 2002,25(2),pp. 108-111.

[6] Ma Feicheng, Zhang Qin. Comparative Analysis of Knowledge Management Literature between China and Overseas:A Bibliometric Analysis . Journal of the China Society for Scientific and Technical Information,2006,25(2), pp.163-171.

[7] Ying Ding, Gobinda G.Chowdhury, Schubert Foo. Bibliometric cartography of information retrieval research by using co-word analysis. Information Processing and Management, 2001,37, pp.817-842.

[8] Chaomei Chen, Katherine McCain, Howard White, et al. Mapping Scientometrics(1981-2001). ASlST 2002 Contributed Paper, pp.25-34.

[9] Ketan K. Mane, Katy Börner. Mapping topics and topic bursts in PNAS. http://www.pnas.org/content/101/suppl.1/5287.full.

[10] Ricardo Baeza-Yates, Berthier Ribeiiro-Neto. Modern Information Retrieval. Addison Wesley,1999.

[11] Y. Cao, J. Liu, S. Bao, et al. Research on expert search at Enterprise track of TREC 2005. In Proceedings of the Text REtrieval Conference (TREC),2005.

[12] Jiepu Jiang, Wei Lu, Dan Liu. CSIR at TREC 2007 Expert Search Task. http://trec.nist.gov/pubs/trec16/papers/wuhanu.ent.final.pdf.

[13] http://www.analytictech.com/Netdraw/NetdrawGuide.doc.

[14] http://web.uconn.edu/~cdavid/netBeansJavajFreeChartpart1.pdf.