MedC: A Literature Analysis System for Chinese Medicine Research

Xin Li[1(✉)], Yu Tong[1], and Wen Wang[1,2]

[1] Department of Information Systems,
City University of Hong Kong, Kowloon, Hong Kong
{Xin.Li,yutong}@cityu.edu.hk, wwang222-c@my.cityu.edu.hk
[2] School of Management, Xi'an Jiaotong University, Xi'an, Shannxi, China

Abstract. Chinese medicine research documents a significant amount of knowledge. However, compared to Western medicine, there are limited studies that take advantage of and summarize findings based on the Chinese medicine literature. This paper builds a literature analysis system based on information extraction and visualization technologies, which allow users to select and analyze a subset of Chinese medicine literature. The system provides complex search functionalities and makes a set of analyses (summary statistics on medicine/disease/acupuncture points, medicine co-occurrence analysis, and acupuncture point analysis) available to support Chinese medicine scholars and alleviate their workload. The system may facilitate Chinese medicine research and theorization.

Keywords: Chinese medicine · Text mining · Visualization · Literature analysis

1 Introduction

Chinese medicine is a treasure of Chinese culture. It reflects Chinese people's long-term wisdom in treating diseases and pursuing a healthy life. In its long history, Chinese medicine developed many effective disease treatments, which is documented in medicine monographs. Since the 1950s, Chinese medicine researchers have been systematically publishing cases, studies, and findings in journals.

This paper argues that it is necessary to examine the existing Chinese medicine publications, since they contain a large amount of empirical evidence from practice and hold the potential to derive theoretical explanations for Chinese medicine. Literature analysis is widely used in Western medicine to summarize previous findings, develop theories, and direct future research. However, literature analysis on Chinese medicine is rare, and difficult to many Chinese medicine scholars. To ease their research using the literature analysis approach, it is necessary to develop a system that provides automated literature analysis and visualization.

In this research, we propose a literature analysis system for Chinese medicine that is developed based on information extraction and knowledge mapping techniques [1, 2]. In this system, we collect Chinese medicine publications (especially journal papers), and extract disease, medicine, treatment, and acupuncture point information from the

X. Zheng et al. (Eds.): ICSH 2015, LNCS 9545, pp. 311–320, 2016.
DOI: 10.1007/978-3-319-29175-8_29

text of papers using text mining methods. We develop functionalities for descriptive statistics, medicine co-occurrence network, and body-based acupuncture point co-occurrence visualization. With this system, scholars can search and select some papers in their areas of interest and generate visualizations to help them digest the selected publications and have an overview of the empirical evidence published in that area.

We built a prototype system (http://medc.is.cityu.edu.hk/). We collect the abstracts and meta-data of about 1 million Chinese medicine papers published since the 1950s in the prototype. The system works in a cloud-computing fashion in which scholars can make multiple analyses and save the results. The system is open to the public.

2 Literature Review

2.1 Literature Analysis in Western Medicine

Medical literature records human experience in fighting diseases. The development of information technologies makes it possible to automatically extract and summarize information from medical literature. Information extraction and knowledge mapping reduce the work required to read and comprehend medical literature. By integrating findings from multiple medical literature, statistical methods and visualization tools can help further reduce errors in information extraction and identify the most important scientific discoveries. Text mining and knowledge mapping have been widely used to analyze and understand Western medical literature [1].

By the end of the last century, the National Library of Medicine (NLM) had supported a number of studies on searching and mining of medical literature [3]. NLM built the PubMed (formerly called MEDLINE) system, which is now the world's largest literature database on life sciences and biomedical information. Many knowledge extraction and mapping tools were developed based on this system to extract genetic relations [2], gene information [4], and other types of information [5]. The data extracted from MEDLINE literature are also being used to carry out further biomedical research [6–8].

2.2 Literature Analysis in Chinese Medicine

In recent years, Chinese scholars have gradually adopted text mining and knowledge mapping techniques to analyze Chinese medical literature [9]. For example, Zhou et al. proposed to mine both Chinese medicine and MEDLINE literature for gene functional networks [10]. Zheng et al. studied the shared biological networks between rheumatoid arthritis and coronary heart disease by simultaneously using Chinese medicine literature and MEDLINE literature [11]. In Taiwan, Fang et al. extracted Chinese medicine information (in English) from MEDLINE literature and developed a database that includes the links between Chinese medicine, genes, and disease [12]. Xue et al. developed a database for herb molecular mechanism analysis based on information extracted from literature [13]. In the past years, there is a group of related scholars who published about 40 papers analyzing various diseases in Chinese medicine literature [14, 15]. Nevertheless, for most Chinese physicians, it is a challenge to learn text mining and knowledge mapping skills and use them to digest literature.

From the knowledge mapping perspective, prescription analysis is related to literature analysis. In 2011, the Institute of Automation at the Chinese Academy of Sciences and the Chinese Academy of Traditional Medicine developed a prescription analysis software, TCMISS (http://www.tcmnd.com/detail.aspx?id=441). Although it uses basic visualization methods and does not have time-mining functions, it still significantly improved the ability of Chinese scholars to analyze prescriptions. This software supported over 30 papers to analyze prescriptions of some famous doctors [16].

Despite the above efforts, in general, existing efforts to analyze Chinese medicine literature are relatively simple. These studies usually report words frequency, word correlation analysis (network visualization) for a selected area or disease [14, 17, 18]. There are no generic Chinese medicine literature analysis systems for scholars.

This paper takes text mining and knowledge mapping approaches to fill the gap. We develop a Chinese medicine literature analysis system to support scholars who have no information processing experience conducting literature analysis.

3 System Architecture

Following the design science paradigm, we develop a Chinese medicine literature analysis system (named MedC) and evaluate its effectiveness in helping Chinese medicine scholars. Figure 1 presents the general framework for this study, including testbed development, system development, and user studies. Before system development, in the testbed development part we collect Chinese medicine literature from CNKI (the largest literature database in China), which will be mined and analyzed. We also collect medical knowledge bases on medicine, diseases, and syndromes to facilitate the text mining task. After system development, for the user study we will interview domain experts and conduct studies with students to validate the proposed approach. This paper focuses on reporting the system development component, which is shown in detail in Fig. 2.

Information Extraction Module Design and Development. After collecting the Chinese medicine literature, we built an automatic information extractor based on existing research in Chinese NLP. We convert the collected papers to text format and build a heuristic filter to filter out irrelevant information, such as header on each page, and keep only paper content. We conduct Chinese word segmentation and POS tagging by using the ICTCLAS package (http://ictclas.nlpir.org) developed by NLPIR. We incorporate Chinese medicine-specific dictionaries, which are built upon the collected medical knowledge bases, to extract terms such as disease names, symptoms, etc. We conduct rule-based term aggregation and disambiguation to address typos and synonyms in literature, so that each term has a unified identifier in our system. (This step is supported by a thesaurus of medicine and disease names collected and compiled by us.) We then extract co-occurrence relations between terms within and across sentences and paper sections. We pre-process our collection and conduct information extraction for later visualization.

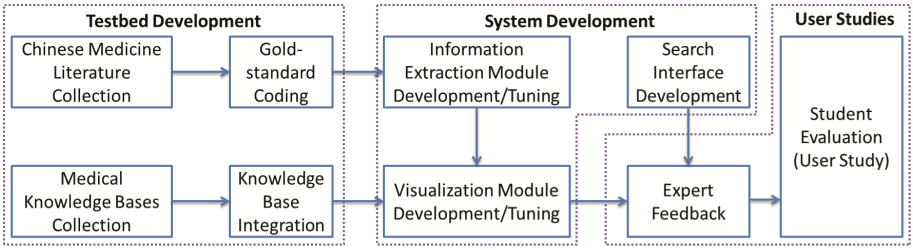

Fig. 1. The framework for the study on MedC system

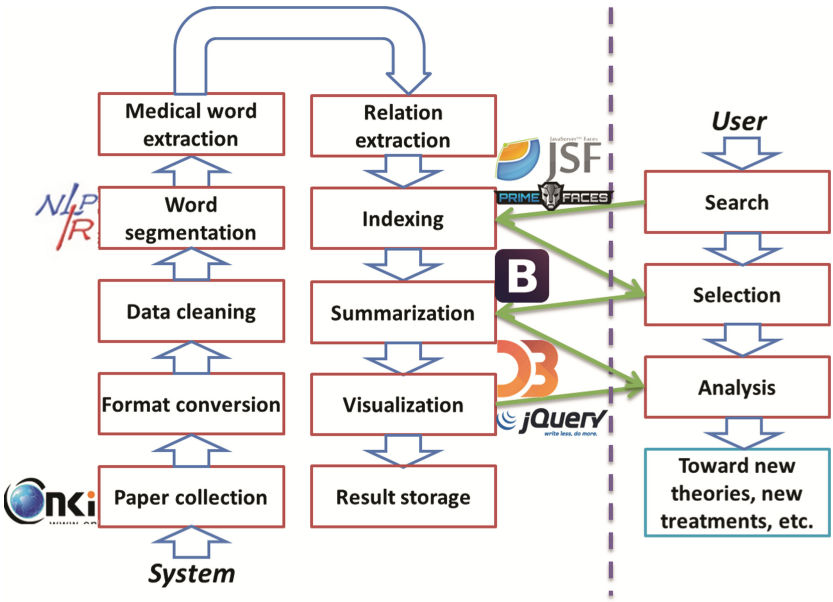

Fig. 2. System development

Search Interface Development. Different from previous researchers' work, we do not specify any scope of literature. We leave the selection of literature to be analyzed to researchers, i.e., potential users of our system. To facilitate users finding the literature they are interested in for visualization and analysis, we develop a Web interface that contains complicated literature search and selection functionalities. The interface is built on Html, Ajax, Java, and JSP techniques. It is based on the JavaServer Faces (JSF) framework[1] using primefaces[2] with the help of bootstrap[3]. The webpage logic was

[1] https://en.wikipedia.org/wiki/JavaServer_Faces.
[2] http://primefaces.org/.
[3] http://getbootstrap.com/.

written with the help of JQuery[4]. The system also has basic account management and report management functionalities to ease the use of visualization modules.

Visualization Module Design and Development. The visualization module is a key component of the project. We customize an open source package, D3[5], to build the interface. The visualizations are designed considering Chinese medicine researchers' requirements. Specifically, we provide summary statistics of diseases, medicines, and acupuncture points, in which we highlight the inter-correlations among the three types of entities. We provide medicine co-occurrence graph visualization together with their medicine properties (药性, Chinese medicine classifies medicines to four types: cold, hot, warm, and cool; note that such classifications may not match their chemical properties) and GuiJing (归经, a unique construct in Chinese medicine to annotate which part-of-human-body the medicine belongs to; note that the part-of-human-body in Chinese medicine does not always match anatomy). We provide an acupuncture point visualizer based on the human body. While we are still developing other visualization modules, the three visualizers provide us unique perspectives on understanding the selected literature.

4 System Implementation/Functionality

4.1 Testbed

We developed a spider and crawled all the abstracts and meta-data of CNKI publications (the full-text of papers requires purchase). By November 2014, we collected the abstracts of 1,098,014 articles, including 1,081,864 journal papers, 2,995 doctoral dissertations, and 12,118 master theses. Our customized text segmentation system contains a lexicon with 979,079 unique words integrated from multiple sources. By applying the text segmentation tool to the 1 M papers, we extract 494 unique disease names occurring 672,226 times, 877 unique medicines occurring 669,604 times, and 331 unique acupuncture points occurring 151,925 times.

Chinese medicine literature abstracts usually do not contain enough information on treatments, prescriptions, and symptoms of the subjects, which are necessary for this research. In the future, we will purchase full-texts of papers to improve the system. For a case study, we retrieved asthma research papers with the help of some Chinese medicine scholars. Such literature is used to illustrate the functionalities of the system. The system also allows users to upload full-texts of papers for analysis.

4.2 System Functionality

Literature Search and Selection. The MedC system provides several searching functions for Chinese medicine scholars. Users can start with simple search functionalities,

[4] https://jquery.com/.
[5] http://d3js.org/.

such as title, keyword, and abstract. The system also provides advanced search functionalities. Users can use author, acupuncture point, medicine, disease, and year to search the results. After the initial query, the related publications are listed as shown in Fig. 3, on which users can further shortlist publications by applying filtering criteria, including time, journal names, and words in title/abstracts. On the shortlisted search results, users can browse and select papers into an analysis candidate list. The left half of Fig. 4 shows the analysis candidates. Here users can search and select publications, so they have the flexibility to analyze any subset of the 1 M publications in our system. We assume, after this step, the users can identify a small set of related documents that are worthwhile and relevant for further analysis.

Fig. 3. Literature search and selection

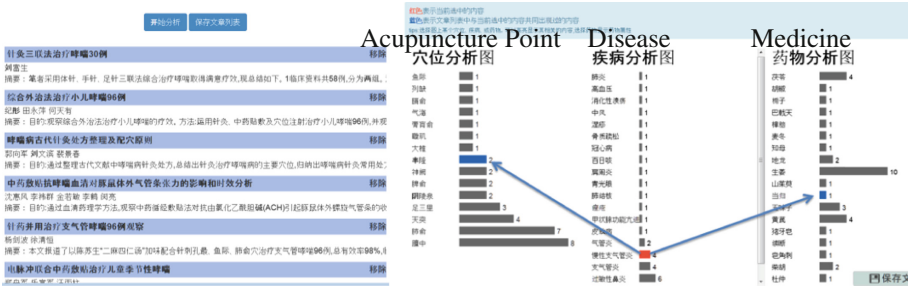

Fig. 4. Analysis candidates and summary statistics

Summary Statistics. The system provides three ways to visualize Chinese medicine literature. The summary statistics function reports the occurrence frequency of acupuncture points, diseases, and medicine in the selected literature. Since our collected dataset is abstracts, the co-occurrence relations are at the abstract level. The right part of Fig. 4 illustrates this functionality; by selecting any term on the bar charts, the co-occurrence of two other types of terms in the charts will be highlighted. As a result, the highlighted bars show the co-occurred elements surrounding a specific disease and medicine in the selected literature. This analysis can aid researchers to have some initial idea of the disease and treatment relations in existing literature.

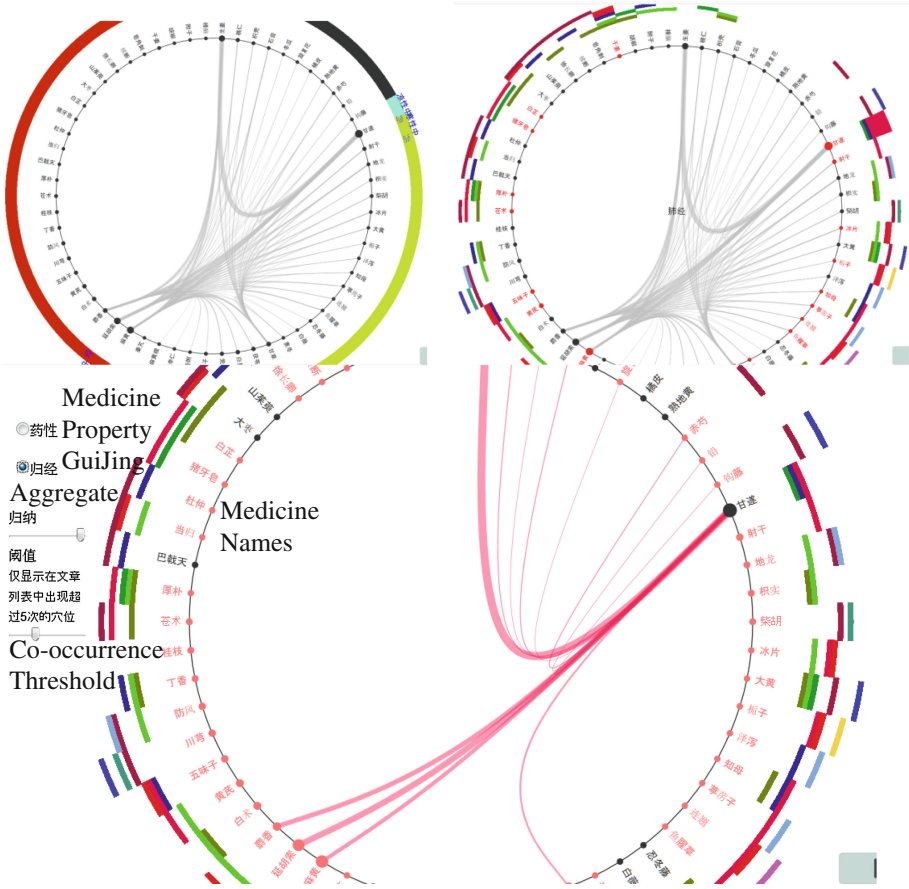

Fig. 5. Medicine co-occurrence graph

Medicine Co-occurrence Graph. Figure 5 shows the medicine co-occurrence graph, which shows the medicines that are often used together. Here, we visualize the co-occurrence network in a ring-layout. We organize medicines according to their properties as shown in the upper left part of Fig. 5. The system can also show the GuiJing of medicines using colors outside of the rings, as shown in the upper right part of Fig. 5.

The interface can filter out links with a small co-occurrence frequency and group similar types of links closer, which makes them more distinct on the interface, as shown in the lower part of Fig. 5. Moving the mouse over a medicine will show its GuiJing and highlight its related medicines. This function allows researchers to identify the interesting relations among medicines, such as the effective collocation or prohibition of using medicines, and conduct follow-up analysis.

Acupuncture Point Graph. Figure 6 illustrates the acupuncture point graph, which places acupuncture points according to their position on the human body. The visualization highlights the co-occurred acupuncture points using links, which can also be filtered by occurrence frequency. Placing the mouse on acupuncture points will highlight their meridian (经络, which are groups of related acupuncture points in their locations and treatment effects) and their closely related acupuncture points. By using this function, researchers can analyze acupuncture points across the meridian, which allows them to explore other theoretical explanations of acupuncture points.

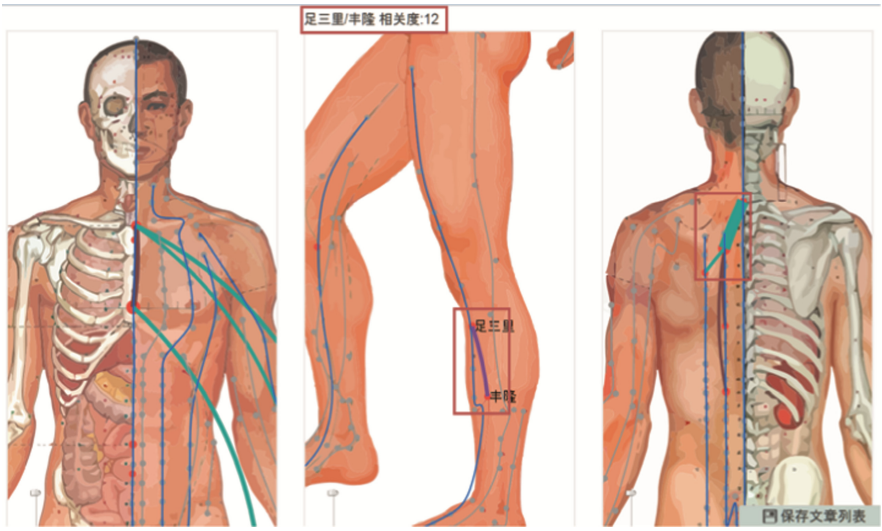

Fig. 6. Acupuncture point graph

Other Functionalities. Users can save the analysis results and retrieve them from the system later. Users can also upload new Chinese medicine papers onto this system.

5 Discussion and Future Work

This paper develops a tool (MedC) to help Chinese medicine scholars retrieve literature, extract key information, and analyze their relations. The system contains more than 1 million Chinese medicine publications and works in a cloud-computing fashion. Users can access it using a browser. MedC can be used to support scholars to comprehend the

key information from literature and direct their explorations in Chinese medicine research. This paper discusses the development and functionalities of the system.

In order to validate the effectiveness of our proposed approach, we need to conduct follow up user studies to evaluate system functionalities. Such evaluation, which is missing in this paper, is critical for a design science paper, especially for the development of design theories. We plan to recruit Chinese physicians and conduct an interview regarding the usefulness and ease-of-use of the information extraction, visualization, and knowledge integration components. We will also conduct another round of evaluation among senior students majoring in Chinese medicine. These senior students have experience in participating in clinic decisions and hence can be considered as junior-level practitioners in Chinese medicine. The purpose of the evaluation is to evaluate how much the platform has changed the efficiency and effectiveness of physicians' work. We will adopt a randomized, between-groups design to conduct the evaluation.

In future research, we will continue improving the system, including information extraction effectiveness and visualization functionalities. Nevertheless, it is one of the earliest efforts to reduce manual efforts in Chinese medicine literature analysis. It allows scholars without IT background to employ state-of-the-art text mining and visualization methods to analyze empirical evidence related to their interest.

Acknowledgements. The research is partially supported by National Natural Science Foundation of China grant 71572169, GuangDong Natural Science Foundation grant 2015A030313876, and CityU SRG 7004287.

References

1. Chen, H., Fuller, S., Friedman, C., Hersh, W.: Medical Informatics: Knowledge Management and Data Mining in Biomedicine. Springer, Heidelberg (2005)
2. Leroy, G., Chen, H.C.: Genescene: an ontology-enhanced integration of linguistic and co-occurrence based relations in biomedical texts. J. Am. Soc. Inf. Sci. Technol. **56**, 457–468 (2005)
3. Houston, A.L., Chen, H.C., Hubbard, S.M., Schatz, B.R., Ng, T.D., Sewell, R.R., Tolle, K.M.: Medical data mining on the internet: research on a cancer information system. Artif. Intell. Rev. **13**, 437–466 (1999)
4. Wei, C.H., Harris, B.R., Kao, H.Y., Lu, Z.Y.: tmVar: a text mining approach for extracting sequence variants in biomedical literature. Bioinformatics **29**, 1433–1439 (2013)
5. Savova, G.K., Masanz, J.J., Ogren, P.V., Zheng, J.P., Sohn, S., Kipper-Schuler, K.C., Chute, C.G.: Mayo clinical text analysis and knowledge extraction system (cTAKES): architecture, component evaluation and applications. J. Am. Med. Inform. Assoc. **17**, 507–513 (2010)
6. Wong, A., Shatkay, H.: Protein function prediction using text-based features extracted from the biomedical literature: the CAFA challenge. BMC Bioinform. **14**, S14 (2013)
7. Li, J.X., Zhang, Z., Li, X., Chen, H.: Kernel-based learning for biomedical relation extraction. J. Am. Soc. Inf. Sci. Technol. **59**, 756–769 (2008)
8. Li, X., Chen, H.C., Li, J.X., Zhang, Z.: Gene function prediction with gene interaction networks: a context graph kernel approach. IEEE Trans. Inf. Technol. Biomed. **14**, 119–128 (2010)

9. Zhou, X.Z., Peng, Y.H., Liu, B.Y.: Text mining for traditional Chinese medical knowledge discovery: a survey. J. Biomed. Inform. **43**, 650–660 (2010)
10. Zhou, X.Z., Liu, B.Y., Wu, Z.H., Feng, Y.: Integrative mining of traditional Chinese medicine literature and medline for functional gene networks. Artif. Intell. Med. **41**, 87–104 (2007)
11. Zheng, G., Jiang, M., He, X.J., Zhao, J., Guo, H.T., Chen, G., Zha, Q.L., Lu, A.P.: Discrete derivative: a data slicing algorithm for exploration of sharing biological networks between rheumatoid arthritis and coronary heart disease. Biodata Min. **4**, 18 (2011)
12. Fang, Y.C., Huang, H.C., Chen, H.H., Juan, H.F.: TCMGeneDIT: a database for associated traditional Chinese medicine, gene and disease information using text mining. BMC Complement. Altern. Med. **8**, 58 (2008)
13. Xue, R.C., Fang, Z., Zhang, M.X., Yi, Z.H., Wen, C.P., Shi, T.L.: TCMID: traditional Chinese medicine integrative database for herb molecular mechanism analysis. Nucleic Acids Res. **41**, D1089–D1095 (2013)
14. Zhang, Z.H., Guo, H.T., Zheng, G., Feng, F.H., Li, S.W.: Exploration drug characteristics of Sjogren's syndrome using text mining. Tradit. Chin. Med. Res. **26**, 72–74 (2013)
15. Zha, Q.-l., Yu, J.-Y., Yu, F., Zheng, G., Guo, H.-T., Lv, A.-P., Yu, Z., Jiang, M.: Biological characteristic of five flavours categorized Chinese herbs used for cough treatment based on metabolism related mesh text mining. Chin. J. Basic Med. Tradit. Chin. Med., 616–618 (2010)
16. Liu, T., Liu, J.C.Z.T., X.F., Liu, X.S., Zhang, W.D.: Based on TCM inher-itance assist system to analyze the mdication experience of Liu Yun-Shan for treating Diarrhea in children. Clin. J. Chin. Med. **5**, 10–13 (2013)
17. Tan, Y., Yang, J., Zhao, N., Zheng, G., Cai, F., Jiang, C.Y., Guo, H.T., Jiang, M., Lv, A.P.: Regularity of Chinese and western medicine application for chronic hepatitis b with text mining technique. Chin. J. Exp.Tradit. Med. Formulae **17**, 232–235 (2011)
18. Zhou, C.Y., Chen, H.J., Tao, J.H.: Graph: a domain ontology-driven semantic graph auto extraction system. Appl. Math. Inform. Sci. **5**, 9–16 (2011)