


Structuring the Chinese disjointed literature-based knowledge discovery system: The key technologies to success

Journal of Information Science
38(6) 532–539
© The Author(s) 2012
Reprints and permission: sagepub.
co.uk/journalsPermissions.nav
DOI: 10.1177/0165551512461104
jis.sagepub.com


Qing Qian

Institute of Medical Information, Chinese Academy of Medical Sciences, China

Na Hong

Institute of Medical Information, Chinese Academy of Medical Sciences, China

Xinying An

Institute of Medical Information, Chinese Academy of Medical Sciences, China

Abstract

The existing evidence shows that Chinese disjointed literature-based knowledge discovery is largely limited to traditional Chinese medical literature instead of the entire medical literature available in Chinese. Some limitations exist in current research when put into practice, for many reasons, including the data all coming from a single source. During the process of knowledge discovery, the problem of excessively large intermediate and target concept sets still needs attention. This Chinese medical disjointed literature-based knowledge discovery system was constructed by applying research to the concepts of extraction, pruning and sorting algorithms in an endeavour to solve the existing problems associated with disjointed literature-based knowledge discovery systems. The system was tested and verified in accordance with the classic magnesium–migraine hypothesis of Dr Swanson. Future improvements and development of the system are also proposed.

Keywords

Chinese medical literature; disjointed literature-based knowledge discovery; knowledge mining

1. Background

In 1986, Don. R. Swanson, professor at the University of Chicago, discovered a connection between two non-interactive retrieval literature groups, fish oil and Raynaud's syndrome using physiological changes of Raynaud's syndrome as intermediate terms. The research concluded that fish oil could be used to improve the symptoms of Raynaud's disease, which was supported by subsequent studies. Based on the discovery, Swanson first introduced the concept of discovering new relationships within a bibliographic database, that is, literature-based discovery theory. He pointed out that public, but disjointed, biological literature may harbour a large amount of undiscovered public knowledge. If these fragmented, or isolated, data were logically brought together and properly interpreted, useful information would be revealed. In his ABC model, two disjointed literature clusters (*A* and *C*) can be linked with intermediate terms or literature (*B*), and the process of finding the connection between *A* and *C* is called disjointed literature-based knowledge discovery. This method of knowledge discovery is not only a milestone for intelligence science research, but also a valuable tool to retrieve the associated data that conventional search methods normally ignore.

Based on the knowledge discovery model proposed by Swanson, some applications have been developed during about 20 years, as well as many data mining systems, such as the Arrowsmith [1] system by Smalheiser and Swanson, the DAD

Corresponding author:

Xinying An, Institute of Medical Information, Chinese Academy of Medical Sciences, No. 3 Yabao Road, Chaoyang District, 100020 Beijing, China.
Email: anxinying@yahoo.com.cn

(drug–adverse drug reactions–disease) system [2] by Weeber and Klein, the LitLinker system [3] by Yetisgen-Yildiz and Pratt, the Manjal [4] system by Srinivasan, the TransMiner [5] system by Narayanasamy and Mukhopadhyay, and the BIOTLA system [6] by Hristovski et al.

In China, the disjointed literature-based knowledge discovery method of Swanson was first applied in an attempt to promote the development of traditional Chinese medicine by Xu Jianyang et al. in 2005 [7]. Shao Yunfeng and co-workers tried to construct a Chinese medical disjointed literature-based knowledge discovery system based on the Arrowsmith System [8]. Gao Hongjie et al. adopted the subject headings frequency-based statistics in the open disjointed literature-based knowledge discovery system and analysed the associated connections of traditional Chinese medical literature on diabetic nephropathy treatment, resulting in a new hypothesis [9]. Zhang Linwei and co-workers attempted to modify and apply the Swanson method to frontier defence intelligence research [10]. In 2006, Liu Yao et al. constructed the tagged corpus of traditional Chinese medical literature and described the language knowledge database construction of the ancient Chinese medical literature from the perspective of traditional Chinese medical semantic dictionary [11]. In 2007, Zhang Jun provided a text analysis method in which characteristic vectors were extracted from a self-constructing Chinese medical dictionary. Following statistical classification, disjointed Chinese medical literatures were obtained from the dictionary [12]. Several studies were also performed by Huang Shuiqing et al. in an effort to apply the disjointed literature-based knowledge discovery method to social science research [13–16]. Leng Fuhai and co-workers brought forth the covert-linking and ternary co-words between literatures to improve the disjointed literature-based knowledge discovery method [17, 18]. Zhang Yunqiu made use of the integrated approaches to filtering, bidirectional term frequency, medical subject heading (MeSH)-based weighting and literature cohesion-based weighting to solve the main problem of low-quality intermediate concept sets [19]. Although considerable efforts have been made by many Chinese scholars in the field of non-interactive literature-based knowledge discovery research with some fruitful results having been accomplished, there are still some limitations or problems remaining to be solved:

- (1) The existing research and applications of the Chinese disjointed literature-based knowledge discovery are largely limited to the traditional Chinese medical literature instead of the entire medical literature in Chinese. This will lead to some research limitation because the data all come from a single source.
- (2) As the intermediate and target concept sets are too large, often leading to the masking of true, useful target concept sets, it is important to select an appropriate filtering strategy with regard to the intermediate concept set so as to improve the quality of valuable target concept discovery. Appropriate strategies should also be used when calculating the intermediate or target concept.
- (3) In theory, Swanson's methods are applicable to Chinese literature analysis. In practice, however, some unexpected problems may occur owing to the fact that Chinese literature is different from English literature in both grammatical structure and semantics.

In view of the reality that there has not been a mature medical disjointed literature-based knowledge discovery system available in China, this study intends to describe and explore technologies crucial to the design and development of Chinese medical disjointed literature-based knowledge discovery system (CmedLBKD).

2. Framework and key technology research of the CmedLBKD

2.1. Framework of CmedLBKD

Based on the open disjointed literature-based knowledge discovery theory proposed by Swanson, the study explored strategies for improvement and for better application of the proposed methods in the Chinese medical literature environment. The resulting CmedLBKD system was intended to achieve such a smooth application. The framework of CmedLBKD system is illustrated in Figure 1.

The concrete process is as follows:

- (1) *Construction of the start concept set* – China Biology Medicine disks (CBM) and PubMed are considered authoritative and comprehensive medical databases. Users enter a specific search term online. If the search term is a subject heading, the system outputs the subject heading and its narrow terms, and forms the start concept set *A*. If the search term is not a subject heading, the system then submits the term to PubMed and CBM for literature retrieval, and recommends a group of related concepts to the users. Simultaneously, the system filters the set using the Unified Medical Language System (UMLS) semantic network to help users locate the start

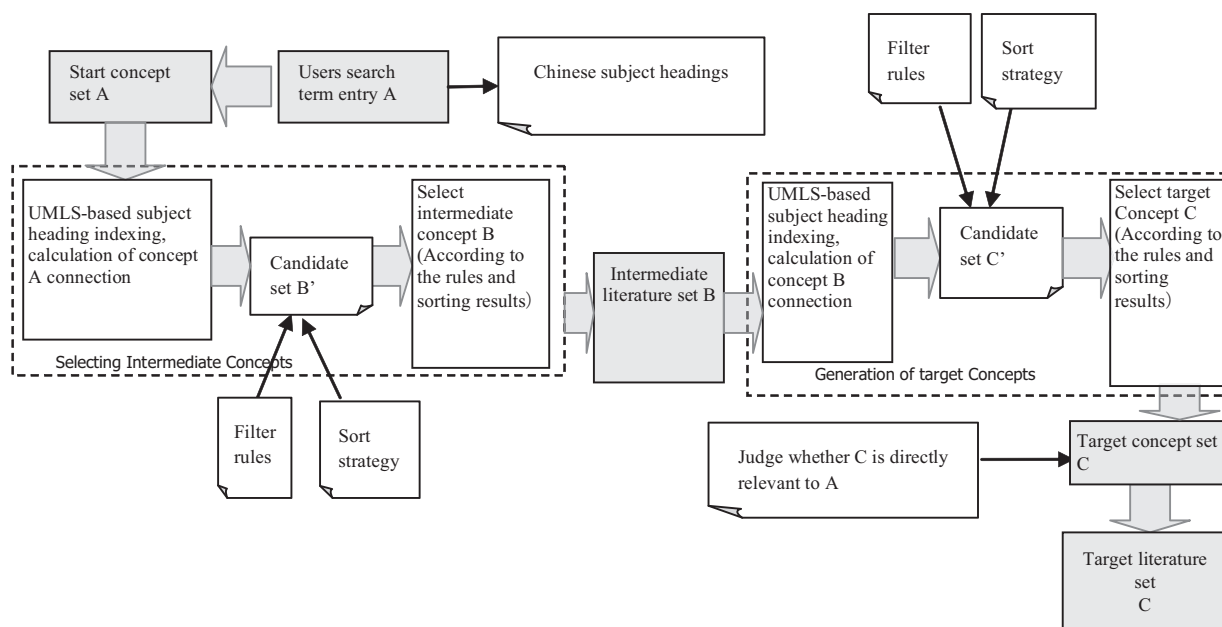


Figure 1. Framework of the open Chinese disjointed literature-based knowledge discovery system.

concepts quickly and obtain the start concept set A . Then, the start literature set A is constructed by retrieving the start concept set A from the database.

- (2) *Construction of the intermediate concept set* – among a large quantity of intermediate concepts, how to select a strategy to filter out the uncorrelated words and reserve those correlated to form the intermediate concept set and thus navigate to the target concept C is the technological key for the disjointed literature-based knowledge discovery process. The present study used text mining and subject indexing to extract concepts from the start literature set A , and selected the correct method to calculate the concept B candidate that is correlated to A . The concept B candidate was defined as the intermediate concept set B' . After pruning and sorting for B' , the system allowed users to configure a variety of sorting strategies in order to choose the important intermediate concepts that might lead to the discovery of a potential target concept. This process eventually determined the intermediate concept set B , and then retrieved the intermediate literature set B from the database.
- (3) *Construction of the target concept set* – through text mining and subject indexing, the concepts were extracted from literature set B . The correct methods were used to calculate the candidate concept C that is correlated to B . After pruning and sorting for the candidate concept, the system permitted users to configure a variety of sorting strategies, and eventually determined the target concept set C . After database retrieval, target literature set C was obtained, which was relevant to the users' interest. Users then could track the contents and background of the target knowledge discovery and further validate the hypothesis.

2.2. Key technology research

Within the framework of the Chinese medical disjointed knowledge discovery system, multiple key technology studies, including concept extraction, concept pruning and concept sorting, have been carried out in light of the uniqueness of Chinese medical literature.

2.2.1. Conceptual theme extraction. Subject headings are part of the standardized vocabulary that reflects subject concepts that are not isolated, but logically interacted, and therefore form a summary of the themes of the article. Subject indexing reduces the noise caused by the free-word analysis of literature, and is technologically more feasible, thus being convenient for developing disjointed knowledge discovery. However, subject heading alone sometimes may fail to identify new discoveries in the authentication phase. By weighting accuracy rate against recall rate, we focus on improving the accuracy of knowledge discovery over recall rate by switching to subject heading, as do most of the researchers abroad on the

study of disjointed literature knowledge discovery, and let the system regulate the retrieval words and recommend relevant concepts by means of ‘Chinese Medical Subject Headings’ (CMeSH).

CMeSH was organized by the Institute of Medical Information, Peking Union Medical College and the Chinese Academy of Medical Sciences. CMeSH integrates the MeSH of the National Library of Medicine in Chinese and Traditional Chinese Medical Subject Headings, in the use of Chinese medical literature indexing, cataloguing and retrieval [20].

In this study, concept extraction was divided into the following steps:

Step 1. Segmentation match. Based on ‘Chinese Medical Terminology’, the data sources (title, abstract and keywords) are segmented using the reverse maximal matching algorithm. The words matching the terms in ‘Chinese Medical Terminology’ will be extracted as keywords.

Step 2. Subject headings conversion. The extracted keywords are converted to standard subject headings using the ‘Chinese Medical Terminology’ subject headings mapping table.

Step 3. Subject headings duplicate deletion. Different keywords generated from segmentation may belong to the same subject heading because of the diversity of expression of a natural language. The system will merge duplicates and, in the end, keep only one subject heading.

Step 4. Subject headings weighting. The importance of a word in an article is, to some extent, decided by its occurrence in a specific context as well as its frequency. As a result, the system can weight the converted subject headings by calculating a threshold value based on their position in the article (title, keywords, abstract) and the frequency. Any subject headings that surpass this value are selected.

2.2.2. Concept pruning. The system needs to prune the concepts in the start concept set, the intermediate concept set and the target concept set in order to lighten the burden on the user and, simultaneously, to find the most significant target concept. The concept pruning adopted in the study included pruning the hypernyms and hyponyms from the start, the intermediate and the target concepts, and filtering the check tags, the broad terms and the UMLS semantic types.

- (1) *Hypernym and hyponym pruning* – the subject headings that are directly related to the start concept and intermediate concept are removed. Hypernym/hyponym pairs are of a hierarchical relationship. Semantically, hyponym is included in hypernym, similar to the relationship between included and including, species and genus, part and whole. Once they have the subject heading, users can access the hypernym and hyponym headings by following the thesaurus tree numbers in Mesh or CMeSH, and thus understand the specialization and discipline attributes of the subject heading in a broader sense (hypernym) or narrower sense (hyponym).
- (2) *Check tags and broad terms filtering* – the interfering words with little value in helping discover target concepts are removed. The check tags are a group of concept identifiers that are commonly used by researchers in biomedical fields. Broad terms are from the first or second level in CMeSH or MeSH tree, and are considered less useful in helping further identify the target concept, therefore they are removed from the start and intermediate concept sets.
- (3) *UMLS semantic filtering* – the candidate of the start or intermediate concept set is mapped to the semantic types in UMLS semantic network for concept filtering. For English data, the system will look for the UMLS semantic types corresponding to the subject headings on MeSH; whereas for Chinese data, the system maps MeSH to CMeSH first, and then defines the UMLS semantic types of Chinese subject headings. The semantic types that either match the concept set or are close to matching the concept set (such as parent concept and child concept) are extracted and recommended to the users. Eventually, users select UMLS semantic types according to their own requirements and thus increase the specificity of knowledge discovery.

After the three filtering processes above, the start literature set A , the intermediate concept set B and the target concept set C are eventually obtained.

2.2.3. Concept sorting. The present system provides a variety of different strategies with regard to set B (intermediate) and set C (target) concept sorting. Users can sort the concepts of set B or set C according to their own requirements, and make use of the sorting results to choose meaningful intermediate concepts or target concepts. The sorting strategies designed in this study include word frequency sorting and mutual information sorting. Users can choose either method to sort the intermediate concept set and decide the scope of the threshold value, so as to reduce the number of subject headings in the intermediate concept set.

- (1) *Word frequency sorting* – the frequency of the subject headings in concept sets is calculated, and sorted in an inverted order, making it more convenient for users to choose parts of the concept set that represent the main contents of the literature and use them for concept set B or concept set C , in order to discover the pertinent target concept.
- (2) *Mutual information sorting* – mutual information is based on the notion that the similarity degree of two concepts depends on the similarity degree of their contexts. If the contexts of the two words in the tagged corpus are always very similar, they can be considered very similar to each other. Take the intermediate concept and the start concept as an example, the higher similarity B has with A , the higher B will be prioritized. The same principle applies to the mutual information analysis between the target concepts and the intermediate concepts. The computation formula of mutual information is as follows:

$$MI(B_i, A) = -\log_2 \frac{p(B_i)p(A)}{p(B_i, A)} = \log_2 \frac{\frac{C(B_i, A)}{N}}{\frac{C(B_i)}{N} * \frac{C(A)}{N}} = \log_2 \frac{C(B_i, A) * N}{C(B_i) * C(A)}$$

where $p(B_i)$ and $p(A)$ are the marginal probability distribution functions of subjects B_i and A , respectively, and $p(B_i, A)$ is the joint probability distribution function of subjects B_i and A . $C(B_i)$ and $C(A)$ are the frequencies of subjects B_i and A , respectively, and $C(B_i, A)$ is the joint frequency of B_i and A . N is the total number of subjects in the intermediate literature set B .

3. System implementation and experiment

Based on the general framework of an open disjointed literature-based knowledge discovery system and the results of research on the key technologies, we designed and completed the construction of a CmedLBKD. To validate the feasibility of the CmedLBKD system, we then conducted a knowledge discovery experiment on the relationship analysis of migraine and magnesium, which had been previously confirmed by Swanson. The experiment attempted to reproduce Swanson's findings by selecting the same search term, 'migraine', with the purpose of obtaining the target word 'magnesium' by means of the CBM database.

3.1. Construction of start concept set

After entering the search term 'migraine' into the CmedLBKD system, a list of CMeSH-based Chinese subject heading vocabularies that matched 'migraine' was generated (Figure 2).

款目词	主题词	命中文献数	提交
典型偏头痛 见	先兆偏头痛	43	<input type="checkbox"/>
家族性偏瘫型偏头痛 见	先兆偏头痛	43	<input type="checkbox"/>
基底动脉型偏头痛 见	先兆偏头痛	43	<input type="checkbox"/>
偏瘫性偏头痛, 家族性 见	先兆偏头痛	43	<input type="checkbox"/>
偏头痛△	偏头痛	6108	<input type="checkbox"/>
偏头痛 见	偏头痛	6108	<input type="checkbox"/>
偏头痛	偏头痛	6108	<input type="checkbox"/>
偏头痛持续状态 见	偏头痛	6108	<input type="checkbox"/>
偏头痛, 典型 见	先兆偏头痛	43	<input type="checkbox"/>
偏头痛, 复杂性 见	先兆偏头痛	43	<input type="checkbox"/>
偏头痛, 基底动脉型 见	先兆偏头痛	43	<input type="checkbox"/>
偏头痛, 急性精神错乱性 见	偏头痛	6108	<input type="checkbox"/>

Figure 2. Chinese subject headings of search term 'migraine'.

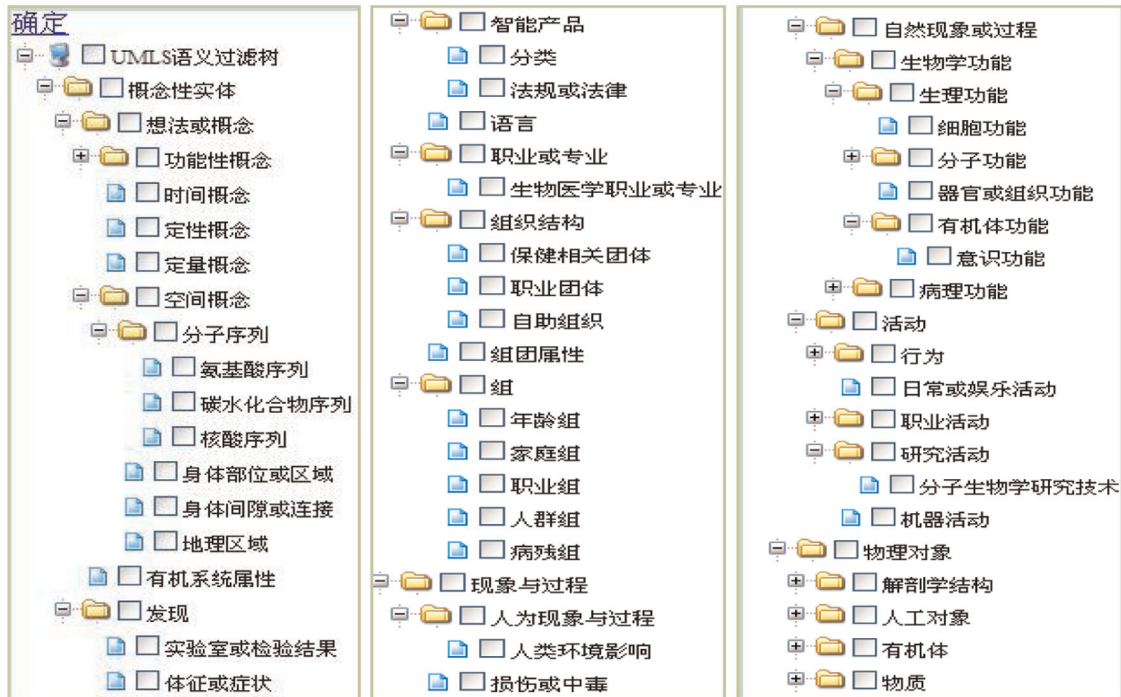


Figure 3. UMLS semantic types.

3.2. Construction of intermediate concept set

By searching the key word ‘migraine’ in CBM, the earliest literature containing both ‘migraine’ and ‘magnesium’ (as an endogenous substance) was published in 1994, indicating that researchers had discovered the relationship between the two in 1994. As a result, the experiment set the time range from 1979 to 1993, and tried to verify the potential relationship between the two words prior to 1994 using the CmedLBKD.

After submitting the subject heading ‘migraine’ as start concept A , it generated a set of 212 subject headings as intermediate concept set candidates B' . After concept pruning and filtering based on the UMLS semantic types, the number of concepts was narrowed down to a user-specific range (Figure 3).

After selecting the semantic type of ‘physiological function’, the number of subject headings of set B' was narrowed down from 212 to 11 (Figure 4).

The CmedLBKD system also supports various sorting strategies and provides concept pruning for different users. Users may select intermediate concepts they believe to be important and construct intermediate concept sets to find the potential target concept. For example, Figure 4(a) was sorted using a word frequency strategy and 4(b) was sorted using a mutual information strategy in which the closer the subject headings are to the start concept A in term of context, the more those headings are prioritized.

The three mesh words, hemodynamics, blood circulation and cerebrovascular circulation, were then selected from the set B' as intermediate concept set B list.

3.3. Generation of target concept set

The intermediate concept set B was submitted to the system for further searching and consequent acquisition of the intermediate literature set that had joint subject headings with B . In total, 990 subject headings were obtained as target concept set C' candidates.

Finally, concepts of the target concept set C' with a co-occurrence relationship with the concepts of the start concept set A , and the concepts of the inclusion relationship between A and C' were deleted. Then the target concept set C was generated. By searching the CBM database, a potential connection between ‘migraine’ and ‘magnesium deficiency’ was found, although apparently in the Chinese medical literature database the start concept was much less associated with the target concept.

主题词	命中数	主题词	互信息参数
血小板聚集	5	微循环	1003
血液粘度	4	血流动力学	128
月经	1	血液循环	52
月经周期	1	月经周期	8
脑血管循环	1	脑血管循环	5
血管舒张	1	血液粘度	5
诱发电位, 躯体感觉	1	细胞凋亡	4
微循环	1	血管舒张	4
细胞凋亡	1	月经	3
血流动力学	1	血小板聚集	3
血液循环	1	诱发电位, 躯体感觉	2

(a)

(b)

Figure 4. Intermediate concept set B' candidate with semantic type of 'physiological function'.

4. Conclusion

Key technology studies on Chinese disjointed literature-based knowledge discovery and the development of the CmedLBKD system were conducted from the standpoint of expanding the application of literature-based knowledge discovery in Chinese medical science. While there is some encouraging progress, further efforts should be made to overcome the following problems:

- (1) The results of knowledge discovery are complicated by human factors to a certain extent. In the case of UMLS semantic filtering, it is necessary to have a design that allows users to select different semantic types according to their specific needs in order to reduce the workload. However, while excessive UMLS semantic types will increase the number of noise words, using fewer UMLS semantic types may reduce the probability of discovery of new knowledge.
- (2) How to best determine the threshold in the sorting algorithm remains to be further studied. In the current sort algorithms, the threshold is still empirically determined as a consequence of lacking experimental supports data to accurately decide the best threshold value that can appropriately recommend the intermediate and target concepts. Further study is needed to exploit better strategies or threshold values that better define the selection range of intermediate concepts and target concepts to reduce the workload and simultaneously increase the accuracy of word selection and concomitantly the efficiency of knowledge discovery.
- (3) The concept representation needs to be improved in both breadth and depth. On the breadth, this system relies on the non-ideal word segmentation resulting from the imperfect Chinese medical terminology database, thus causing the loss of some concepts owing to its inability to find new and unknown words. Although it adopts subject heading recommendations in the start concept selection, it has not fully solved this problem. On the depth, according to the indexing rules, the combination of either 'subject headings + subheadings' or 'subject headings + subject headings' may be used to reveal the theme concept. In the present study, however, only a single subject heading was used.
- (4) The application of the CmedLBKD system was also limited by the lack of innovation in Chinese medical knowledge. On one hand, the majority of the researchers prefer to publish their results in a foreign language journal in order to share their achievements with more people. On the other hand, the politics involved in academic promotion in China also drives researchers to publish their papers in Science Citation Index journals and ignore most Chinese journals that currently are not cited in the Science Citation Index. As a result, many research achievements accomplished by Chinese scientists are not known in China for a long time, which delays work on domestic new knowledge discovery.

Acknowledgements

This paper was funded by the National Key Technology R&D Program (grant no. 2011BAH10B06-02), the Youth Program of National Social Sciences Fund (grant no. 11CTQ016) and the Youth Fund Program of MOE Humanities and Social Sciences Research (grant no. 11YJC870001).

References

- [1] Smalheiser NR and Swanson DR. Linking estrogen to Alzheimer's disease: An informatics approach. *Neurology* 1996; 47: 809–810.
- [2] Weeber M and Klein H. Text-based discovery in biomedicine: The architecture of the DAD-system. In: *Proceedings of the AMLA annual fall symposium, 2000*, pp. 903–907.
- [3] Yetisgen-Yildiz M and Pratt W. Using statistical and knowledge-based approaches for literature-based discovery. *J Biomed Informat* 2006; 39: 600–611.
- [4] Srinivasan P. Generating hypotheses from MEDLINE. *J Am Soc Inform Sci Technol* 2004; 55: 396–413.
- [5] Narayanasamy V and Mukhopadhyay S. TransMiner: Mining transitive associations among biological objects from Medline. *J Biomed Sci* 2004; 6: 864–873.
- [6] Hristovski D, Peterlin B, Mitchell JA et al. Using literature-based discovery to identify disease candidate genes. *Int J Med Informat* 2005; 74: 289–298.
- [7] Xu JY, Ma M and Wang MK. Enlightenment of development of traditional Chinese medical science from Dr. Swanson's non-interactive literature-based knowledge discovery. *Modern Tradit Chin Med Mater Med Wld Sci Technol* 2005; 7: 48–52.
- [8] Shao YF and Wang J. The application of non-interactive literature-based knowledge discovery method in research of traditional Chinese medicine. *J Jiangxi Univ Tradit Chin Med* 2005; 3: 8–9.
- [9] Gao HJ, Zhao YK and andCui M. Research on knowledge discovery based on non-interactive literatures analysis of diabetic nephropathy diagnose and therapy. *Int J Tradit Chin Med* 2011; 8: 710–712.
- [10] Zhang LW and Shan P. Analysis on application of Swanson in border control and immigration intelligence work. *Inform Res* 2007; 2: 121–123.
- [11] Liu Y, Duan HM and Shui ZF. Research on data foundation for non-interactive literature knowledge discovery. *J Inform* 2006; 25: 104–107.
- [12] Zhang J. *Research based on non-interactive literature knowledge discovery of Chinese medicine*. Hefei: University of Science and Technology of China, 2007.
- [13] Huang SQ, Xiong J and Li ZY. Validation of close non-interactive literature knowledge discovery method in the Chinese literature. *J Library Sci China* 2007; 33: 83–87.
- [14] Huang SQ, Cheng C and Li ZY. Validation of open non-interactive literature knowledge discovery method in the Chinese literature. *Inform Stud Theory Applic* 2008; 31: 246–250.
- [15] Huang SQ, Zhang T and Yang DQ. Construction of data sets applied in agricultural economics non-interactive literature-based knowledge discovery. *Jiansu J Agricult Sci* 2010; 16: 192–196.
- [16] Huang SQ and Ma JL. Empirical research of noninteractive literature-based knowledge discovery in Chinese social science literature. *J Library Sci China* 2009; 4: 31–38.
- [17] Cao ZJ and Leng FH. Research on covert-linking literature-based knowledge discovery. *J China Soc Sci Tech Inform* 2010; 4: 605–613.
- [18] Leng FH, Wang L and Li Y. Research on covert-linking literature-based knowledge discovery. *J China Soc Sci Tech Inform* 2011; 4: 1072–1077.
- [19] Zhang YQ. A study on key techniques for disjoint literature-based discovery. *J China Soc Sci Tech Inform* 2008; 4: 521–527.
- [20] Li DY, Hu TJ and Zhu WY. Retrieval system for the Chinese medical subject headings. *Chin J Med Library* 2001; 4: 1–2, 9.