Automatic Hotspots Recognition and Trends Prediction in Traditional Medicine

Wei Lu[1], Xiyan Qin[1], Wu Chen[1], Shuting Chen[2]
1.Center for studies of information resources, Wuhan University,430072, China;
2.Hubei Academy of Scientific & Technical Information,430072, China
reedwhu@gmail.com; qinxiyan2003@yahoo.com.cn; geminiwhu@gmail.com; c.shuting@126.com

Abstract

Exemplified by the field of traditional medicine in this paper, we classify articles extracted from SCIE (2004—2008) medical databases into categories from different perspectives and apply co-word analysis, multidimensional scaling methods to each category. Eventually，two maps are produced for each category, which are the current hotspots map and the trends map. Taking "Drugs, Chinese, Herbal" as an example, its current hotspots map reveals three hot topics: plant extraction, antioxide and cancer cell in 2008. We also find topics concerning cancer in its 2006 trends map, which proved our method regarding high growth rate words as an indicator of the future trends is valid.

1. Introduction

With the development of science and technology, information is growing at exponential rate. So domain hotspots recognition and trends prediction becomes more and more important. However, the task is complex. It is a time-consuming, laborious process, which requires extensive knowledge in terms of data mining, data analysis, and etc. Therefore, it is desirable to assist various users for hotspots recognition and trends prediction using the measures in information science.

Nowadays, the international traditional medicine information platform–APTMNET is playing an active role in the development of traditional medicine cooperation in the Asia-Pacific region. However, how to select the reliable partners and primarily research subjects went into dilemma. So in this paper, we tried to use various measures, such as co-word analysis, multidimensional scaling methods, to recognize and predict internal and foreign hotspots of traditional medicine through massive scientific literatures, seeking to support the scientific research, partner strategies and policy decision making.

The paper is organized as follows: we start with a review of related work about hotspots recognition and trends prediction that is rooted in fields such as scientometrics, bibliometrics, citation analysis, and information visualization. Then exemplified by the field of traditional medicine, we described the implementation of our experiment, such as data source, different methods we applied in different phases, and evaluated the findings of preliminary results. Finally, we made a conclusion of our work.

2. Related work

In general, there are three kinds of techniques that can be used in domain hotspots recognition—mapping, clustering and visualization. Mapping [1] is useful for the subject matter expert and non-expert alike. It provides a means to quickly investigate trends and new information for the expert, and provides an entry point into a domain for the non-expert. Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects [2]. Visualization [3] is any technique for creating images, diagrams, or animations to communicate a message. It has ever-expanding applications in science, education, interactive multimedia, medicine, etc.

Noyons and van Raan [4] tried to map neural network research for Monitoring Scientific Developments from a Dynamic Perspective; Losiewicz, Oard, and Kostoff [5] tried to apply mapping techniques in science and technology management. Hook and Börner[6] regarded domain maps as an alternative means to organize, navigate, and internalize scholarly knowledge and projected a potential future of educational knowledge domain visualizations. Nowell

978-1-4244-2511-2/08/$25.00 ©2008 IEEE

et al. [7] provided scientific guidance for use of graphical encoding to convey information in an information visualization display and gave out several empirical studies. White and McCain [8] provided a kind of visual dynamic view of how real-time cluster presentations might look by means of periodic maps of Information Science.

In addition, various measures are used in this area, such as citation analysis, word frequency analysis, co-occurrence (or co-classification) of words or authors, and Latent semantic analysis etc Moed [9] proposed criteria for proper use of citation analysis as a research evaluation tool. Liming Liang and Caixia Xie [10] made a survey in China's nanotechnology on the base of word frequency analysis. Lei Cui [11] applied co-word analysis in the area of Bronchialveolar Lavage Fluid. Spasser [12] applied co-classification of the International Pharmaceutical abstracts to identify social networks. Boyack et al. [13] used citation and classification based techniques recently to map technology domains based on U.S. patents. Wenlan Li et al. [14] divided the literatures into several categories according to Chinese Library Classification, and those had the most articles were regarded as hotspot.

3. Experiment and evaluation

3.1. Data source

In this paper we take the field of traditional medicine as an example. Firstly, 187 terms were selected from MeSH by domain experts, owing to the rich and controlled vocabulary of MeSH thesaurus may represent traditional medical domain. Secondly, the terms selected are seen as keyword to retrieve in SCIE medical literature databases for a series of successive years (2004-2008), which resulted in 95652 articles. Finally, we extracted the titles, abstracts plus keywords of all retrieved articles, and divided them into five separate parts according to their published year. By comparing them with the results from different parts we could discover the trends of traditional medicine.

3.2. Methods

The methods discussed below were applied to the five parts separately (2004,2005,2006,2007,2008).

Our work consisted of two phases and we adopted different methods in different phases. At phase one, we mainly focused on the classification of those articles from perspectives of contents and countries (or regions). We counted the number of literature in all categories, and the category contains the largest number of literature can be seen as hot spot. At phase two, we applied word frequency analysis, co-word analysis and multidimensional scaling analysis to each category. Detailed steps are as follows.

3.2.1. Phase one. On the one hand, we classified all the articles into categories according to their content. MeSH tree is a very good classification system for medical literature classification. The general structure of MeSH tree is shown in Figure 1. Every node of the tree has a subject heading and several entry terms(from which the 187 terms are extracted by several domain experts), so articles retrieved using a particular term should belong to the node containing this term.

Firstly, according to the seven nodes (B, C, D etc.) on MeSH tree, the literatures extracted are divided into seven categories. Then, we analyzed them respectively. From the results, we can get the categories with the most articles which could be regarded as the hotspots of traditional medicine research.

Secondly, according to the leaf-nodes of MeSH tree, we put all the literature into 57 categories. Though the nodes are not on the same depth of MeSH tree (such as B06.560 and E02.190.488.660.800), they were treated equally. Then we also obtained a ranked list of results, the greater the ranked position, the less attention it attracted.

On the other hand, we followed countries (or regions) classification perspective, so that we can identify hotspots and trends in particular countries.

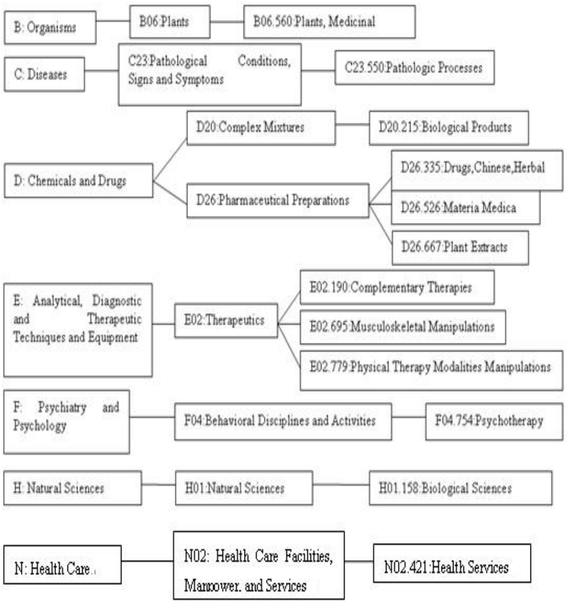
B: Organisms
B06:Plants
B06.560:Plants, Medicinal
C: Diseases
C23:Pathological Conditions, Signs and Symptoms
C23.550:Pathologic Processes
D: Chemicals and Drugs
D20:Complex Mixtures
D20.215:Biological Products
D26:Pharmaceutical Preparations
D26.335:Drugs,Chinese,Herbal
D26.526:Materia Medica
D26.667:Plant Extracts
E: Analytical, Diagnostic and Therapeutic Techniques and Equipment
E02:Therapeutics
E02.190:Complementary Therapies
E02.695:Musculoskeletal Manipulations
E02.779:Physical Therapy Modalities Manipulations
F: Psychiatry and Psychology
F04:Behavioral Disciplines and Activities
F04.754:Psychotherapy
H: Natural Sciences
H01:Natural Sciences
H01.158:Biological Sciences
N: Health Care
N02: Health Care Facilities, Manoower, and Services
N02.421:Health Services

Figure 1. MeSH tree structure

3.2.2. Phase two. Firstly, we parsed and extracted the "Author Keyword" and "Keyword Plus" of the 95652 articles and applied Porter stemming algorithm on them. Secondly, we created a frequency list as well as a growth rate list for each category. And only the top 300 words in the year's lists are taken into account. It is proved that words ranked higher than 300 have little influence on the results. Thirdly, we selected the top 40 keywords from the frequency list and growth list respectively, and constructed two matrixes by co-occurrences of the keywords in the list. We normalized these "raw co-occurrence" matrixes by using Ochiia, which resulted into a similarity matrix; and then we converted it to distance matrixes. Finally, we applied multidimensional scaling analysis on the two matrixes and produced two maps. One result derived from high frequency words is called "current hotspots map". It summarizes the current hot topics in particular sub-areas; another one derived from high growth rate words is called "trends map". It gives an outlook of the possible research hotspots in the near future.

3.3. Results and evaluation

At phase one, Table 1 presents the classification and statistic results based on 7 categories. From it we can see that "E: Analytical, Diagnostic and Therapeutic Techniques and Experiment" is always the top one during the five years (2004-2008). So it is regarded as a steady hotspot in Traditional Medicine.

Table 2 shows the top 5 categories in the 57 nodes. From the results list of 2008, we can see that most researchers focus on " D26.667: Plant Extracts ". It didn't appear in the top five until 2007, which implied that it had been growing rapidly since 2006. The second one is "H01.158.273.118.598: Pharmacognosy", which didn't enter into the top five until last year, which is worth being noticed.

Table 1.Results of classification into 7

	2004	2005	2006	2007	2008
B	0	0	462	4492	1581
C	14	70	57	68	23
D	144	1199	2000	5094	3070
E	1158	4029	10286	20913	9358
F	66	85	100	110	31
H	30	162	385	5926	2487
N	0	97	227	221	61

The results of country-based classification are showed in Table 3. From it, we know USA is the most active place, it has been producing most of the articles for the five years; In addition, China has been playing an important role since 2005.

At phase two, we only illustrated the results of "D26.335: Drugs, Chinese, Herbal" in this paper. And Figure 2 is its trends map in 2006. From it we can see two sub-fields are growing rapidly. District A is about 'nephropathy' and B about "cancer". Figure 3 is its current hotspots map in 2008. There are three research topics in the figure. C is about 'plant extraction'; D focus on "antioxide" and E concerns "cancer cell". Actually, we found district B and E are quite close to each other, which proved that the measure using high growth rate words as the indicator of the future trends is reasonable.

Table 2.Results of classification into 57

2004	2005	2006	2007	2008
E02.190.901.249 229	D26.335 603	D26.335 934	H01.158.273.118.598 4511	D26.667 2376
E02.190.388 229	D20.215.784.500.350 603	D20.215.784.500.350 934	B06.560 4492	H01.158.273.118.598 2077
E02.695.466.500 121	E02.190.321 457	E02.190.321 740	D26.667 2760	E02.190.901.433 1675
E02.695.421 121	E02.190.901.249 300	B06.560 462	H01.158.703.060.500 1470	E02.190.488.505 1675
E02.190.599.374.500 121	E02.190.388 300	D20.215.784.500.492 449	D26.335 1346	B06.560 1581

Table 3.Results of country based classification

2004	2005	2006	2007	2008
USA 425	USA 1395	USA 2873	USA 6731	USA 3879
England 102	China 754	China 1760	China 3207	China 1334
Japan 78	England 439	England 951	India 1801	England 875
Germany 73	Japan 251	Japan 566	Japan 1418	Japan 608

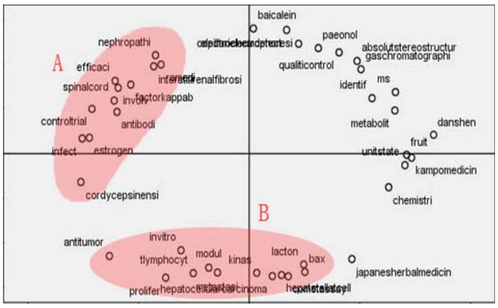

Figure 2. D26.335 2007 trends map

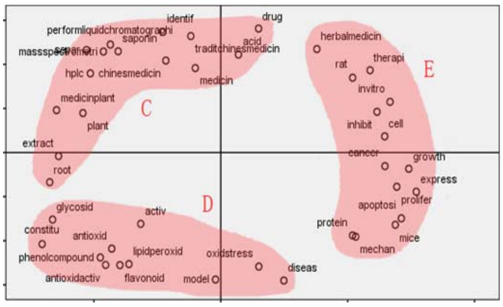

Figure 3. D26.335 2008 current hotspot

4. Conclusion

The paper seeks to automatic recognize hotspots, and ultimately make trends prediction about traditional medicine. It provides means of exploring and evaluating scientific and technical information with enterprises, governments and researchers. From the illustration about “Drugs, Chinese, Herbal” of 2008 above, we found three hot topics: plant extraction, antioxide and cancer cell. We also found topics concerning cancer in its 2006 trends map. It means that our methods are effective. Moreover, similar analysis could be performed for other or large fields of science to support human knowledge analysis and decision-making.

But the work presented here has several limitations: Firstly, our work basically takes place at the level of words and short phrases, and our experiments only use word frequency and word co-occurrence analysis to recognize hotspots in the medical literature, without reference to its deeper linguistic significance. Secondly, we only extracted five years articles from SCIE, whether the time span is enough needs further discussion.

Therefore, there is much work we should do in the future. For example, we will pay efforts to utilize larger datasets, including ISTP(or NSTL), DII, WPI, PCI and web; and aside from the measures discussed above, we will incorporate other methods, such as citation analysis , in it.

5. References

[1] Boyack, K., “Mapping knowledge domains: Characterizing PNAS”, *Proceedings of the National Academy of Sciences*, 2004: 5192-5199.

[2] Jiawei, H., and Micheline K., “Data mining: concepts and techniques”, *China Machine Press*, 2006:383-401.

[3]Visualization:http://en.wikipedia.org/wiki/Scientific_visuali-zation# Fields_of_visualization.

[4] Noyons E., and Raan A..Monitoring “scientific developments from a dynamic perspective: self-organized structuring to map neural network research”, *Journal of the American Society for Information Science*, 1998,49(1):68–81.

[5] Losiewicz, P., Oard, D.W., and Kostoff, R.N., “Textual data mining to support science and technology management”, *Journal of Intelligent Information Systems*, 2000:99–119.

[6] Hook, A. and Börner, K., “Educational knowledge domain visualizations: tools to navigate, understand and internalize the structure of scholarly knowledge and expertise”, *Dordrecht:Springer,* 2005,187-208.

[7] Nowell, L., Schulman,R., Hix, D., “Graphical encoding for information visualization: An Empirical Study”, IEEE Symposium on Information Visualization, 2002, 43- 50.

[8] White, H.D., and McCain, K.W, “Visualizing a discipline: An author co-citation analysis of information science, 1972–1995”, *Journal of the American Society for Information Science*, 1998, 49(4):327 – 355.

[9] Moed, H. F., “Citation analysis in research evaluation”, Dordrecht, *The Netherlands: Springer*, 2005,1-60.

[10] Liming Liang, and Caixia Xie, “Investigation of China’s nanotechnology study based on frequency analysis of key words”, *Studies in Science of Science*.2003, 21(2):138-142.

[11] Lei Cui, “A co-word cluster analysis of the high frequency subject headings of the document set of a relative narrow subject coverage”, *Information Studies:Theory & Application*.1996, 19(4):49-51.

[12] Spasser, and M.A., “Mapping the terrain of pharmacy: Co-classification analysis of the International Pharmaceutical Abstracts database”, *Scientometrics*,1997: 77–97.

[13] Boyack, K.W., et al. “Analysis of patent databases using VxInsight”, *Proceedings of New Paradigms in Information Visualization and Manipulation*, 2000.

[14] Wenlan Li, and Zuguo Yang, “Bibliometrics analysis on information science research subjects distribution”, *Information Science*.2005, 23(3):396-40.