He C, et al., J Altern Complement Integr Med 2015, 1: 006
DOI: 10.24966/ACIM-7562/100006

HSOA Journal of Alternative, Complementary & Integrative Medicine

Review Article

Hotspot Detection in Traditional Chinese Medicine Based on PubMed

Chong He[1], Changbo Zhao[1] and Guo-Zheng Li[1,2]*

[1]*Department of Control Science and Engineering, Tongji University, Shanghai 201804, China*

[2]*Data Center of Traditional Chinese Medicine, China Academy of Chinese Medical Science, Beijing 100700, China*

Abstract

Objective: As the development of Traditional Chinese medicine (TCM), more and more researchers engage in the study of TCM. Meanwhile, some aspects of TCM are paid more attention by TCM researchers, which are regarded as specific hotspot research directions on the field of TCM. These hotspot research directions can help the researchers understand the dynamic change of TCM researches and guide their future research plan. Therefore, it is meaningful to mine hotspots of TCM from the existing database.

Methods: Based on the E-utilities interface of PubMed, the source data can be obtained. They are 10291 abstracts on TCM including manually annotated MeSH terms. Moreover, two hotspot detection schemes are developed according to the characteristics of the source data, one is based on Medical Subject Headings (MeSH) terms co-occurrence while the other is based on text network.

Results: From the experiments results on different hotspot detection schemes, similar key words are obtained to describe major research directions of TCM. Then, according to domain knowledge and practical experience of TCM experts, four consistent hotspots are derived with two different hotspot detection schemes. As shown in the followings: (1) Research on Chinese medical formula and its mechanism of action, (2) Research on pharmacology and pharmacodynamics of antineoplastic agent in TCM, (3) Research on the therapy of chronic diseases in TCM, and (4) Research on traditional therapy methods in TCM.

Conclusion: We study the hotspots detection and dynamic change for TCM research directions on PubMed. The experimental results validate the effectiveness of the proposed approach. Although the proposed schemes are applied to TCM in our research, they also can be extended to other academic fields.

Keywords: Hotspots detection; PubMed; Traditional Chinese Medicine

***Corresponding author:** Guo-Zheng Li, Department of Control Science and Engineering, Tongji University 4800 Cao'an Road, Shanghai, 201804, China, Tel: +86 1064089668; E-mail: gzli@ndctcm.cn

Citation: He C, Zhao C, Li GZ (2015) Hotspot Detection in Traditional Chinese Medicine Based on PubMed. J Altern Complement Integr Med 1: 006.

Received: August 17, 2015; **Accepted:** November 18, 2015; **Published:** December 03, 2015

Introduction

The past decades have brought remarkable advances in our understanding of traditional Chinese medicine. It is not only popular in China, but also widely propagated in other areas all over the world such as America, Japan, German and other countries. It is necessary to recognize the hotspots in TCM as the number of researchers on TCM is larger than ever before. Detecting hotspots for traditional Chinese medicine is great helpful to researchers, who want to devote to TCM research fields.

As data resource on TCM is countless, it is not feasible to deal with these data just rely on expert statistics. Researchers develop several methods to detect hotspots in different fields such as mapping [1-3], clustering [4-6] and visualization [7,8]. Clustering is a popular technology always used in text mining domain. He et al., [9] recognize hot events with TF-IDF model and incremental cluster algorithm. Li et al., [10] adopt SVM and K-means to cluster texts and take the centers of clusters as the hotspots. In addition, various measures are applied such as term frequency analysis, citation analysis and co-occurrence of keywords or authors. Haribhakta et al., [11] represent documents with keywords and compare the co-occurrence of keywords to detect hot topic. Bu-Yeo Kim et al., [12] compare the co-occurrence of keywords to map the dementia research area at the micro-level. All these works focus on hot topic detection with singular scheme, such as the straightforward keywords statistics. However, it is hard to validate the detection confidence due to lack of quantitative ground truth assessment with singular scheme. Therefore, we propose two different schemes for different data subjects, which could mutually validate the hotspots detection results to each other. For specific application, we apply them into TCM domain to mine the research directions in certain years. Moreover, the research hotspots transfer issues of TCM are also studied through the proposed approaches.

This paper is organized as follows: we first introduce the flowchart of our hotspots detection schemes step by step. Then, articles of PubMed database on the field of TCM researches are taken as an instance to perform our experiment. Furthermore, we analyze the hotspots detection results and discuss the advantages and disadvantages of our approaches. Finally, we make a conclusion of our work.

Materials and Methods

This hotspots detection study includes three main parts: data collection and preprocessing feature set selection and similarity computation, MeSH terms/text clustering and hotspots detection (Figure 1).

Data collection and preprocessing

In this paper, all the source data are downloaded from PubMed due to its professional and comprehensive database in biomedical field. Through E-utilities, an interface to access the data used for information retrieval. The retrieved keyword to access data is "Traditional Chinese Medicine". So we mainly download 10201 passages which include abstracts on TCM in recent five years

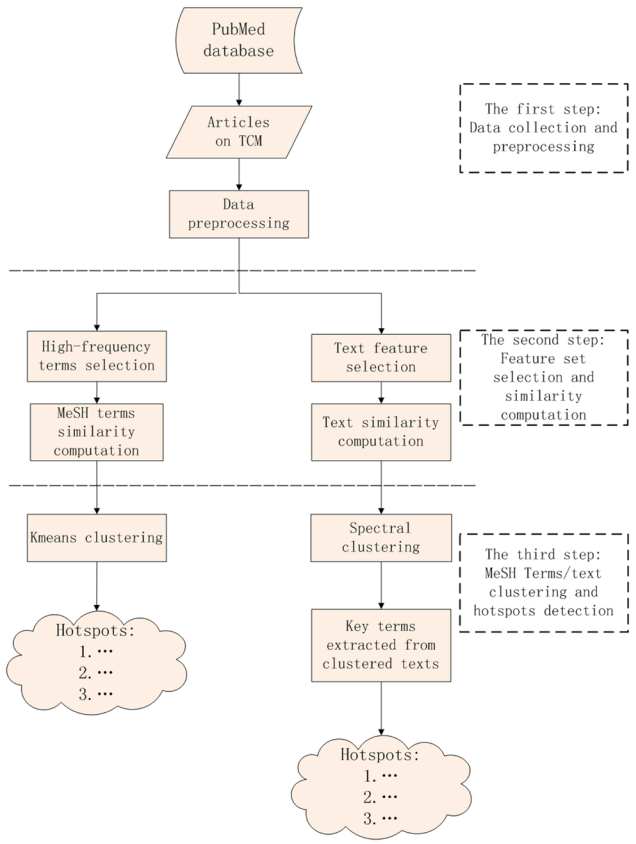

Figure 1: Hotspots detection flowchart. The flowchart is divided into three stages: data collection and preprocessing, feature set selection and similarity computation, and MeSH terms/text clustering and hotspots detection.

(2010-2014). Generally, the topics of articles provide important information to discover the research hotspots. Based on the source data, we find that two data subjects (MeSH terms and abstracts) have highly summarized the topics of articles. MeSH is the NLM (US National Library of Medicine) controlled vocabulary thesaurus used for indexing articles for PubMed. And abstracts generally reflect the main ideas of passages. Besides, as the source data include other redundant subjects which are not related to the topic, data preprocessing step is necessary to be done. To be detailed, we manually filter those redundant data such as publishers of articles and preserve the structured data which contain abstracts, MeSH terms.

Feature set selection and similarity computation

According to the characteristics of data subjects, different subjects are inclined to be analyzed by different feature representations and similarity measurements, so we need to adopt two ways to obtain optimal feature sets and similarity matrix for abstracts and MeSH terms respectively. The first way is based on high-frequency MeSH terms selection and MeSH terms co-occurrence [13], and the second is based on VSM (Vector Space Model) and TF-IDF (Term Frequency-Inverse Document Frequency) feature.

In the first scheme, high-frequency MeSH terms are chose as feature set because the number of MeSH terms is large and low-frequency terms always can not reflect the hot topics. In order to distinct high-frequency terms and low-frequency terms, we adopt the following formula proposed by Donohue:

$$N = \frac{1}{2}(-1+\sqrt{1+8I_1}). \quad (1)$$

I_1 denotes the number of terms whose frequency is 1 and N represents the boundary number to distinct the high-frequency and low-frequency terms. If a term's frequency is higher than N, it can be taken as a high-frequency term. The next step is to compute the similarities of MeSH terms. Once a couple of terms both appear in an article, we may speculate they have some relationships with each other. Moreover, if they co-occur in several articles, it is confident to confirm the terms have potential relationships with each other. And higher the co-occurrence frequency is, more relevant the relationship is. The similarity values are just the degree of the relationship relevance. As mentioned above, a coefficient called Ochiia is applied to represent the similarity between two MeSH terms [14]. The computing formula is as follows:

$$O_{A,B} = \frac{M_{A,B}}{\sqrt{M_A} \times \sqrt{M_B}}. \quad (2)$$

Here $O_{A,B}$ denotes the similarity between MeSH term A and MeSH term B, whose value is in the range 0-1, and $M_{A,B}$ represents the co-occurrence frequency of MeSH term A and MeSH term B. M_A and M_B are corresponding to the term frequency of MeSH term A and MeSH term B, respectively.

In the second scheme, a document is represented as a vector in which each component indicates the value of the corresponding feature in the document [15]. We take the important words inside abstracts which can reflect topic as features. TF-IDF values are computed as feature values which help to emphasize the unique and important words having relative relationships with topics. Now a text can be represented as a vector in which each component indicates the TF-IDF value of the corresponding feature words in the text.

$$D_j = (t_{1,j}, w(t_{1,j})\ t_{2,j}, w(t_{2,j}),...,t_{N,j}, w(t_{N,j})), \quad (3)$$

where $(t_{1,j},t_{2,j},...t_{N,j})$ represents features in feature set. $w_{i,j}$ denotes the TF-IDF value of feature word t_i in text D_j. Taking $(t_{1,j},t_{2,j},...t_{N,j})$ as n-dimensional system of coordinate where $(w_{1,j},w_{2,j},w_{3,j},...w_{N,j})$ is a feature vector. The corpus can be represented as follows:

$$\begin{aligned} &D_1 : w_{11}, w_{12}, ..., w_{1N} \\ &D_2 : w_{21}, w_{22}, ..., w_{2N} \\ &\quad ..... \\ &\quad :: \\ &D_m : w_{m1}, w_{m2}, ..., w_{mN} \\ &\quad ..... \\ &\quad :: \\ &D_n : w_{n1}, w_{n2}, ..., w_{nN} \end{aligned} \quad (4)$$

In the method mentioned above, the content of abstract is mapped as a point in n-dimensional space and the complexity of problem is reduced. It should be noted that it only considers the statistical property but ignores semantic information. In order to improve the accuracy of text similarity, we bring in a method called Latent Semantic Analysis (LSA) [16]. As the corpus is represented as a data matrix, LSA takes a mathematical technique called Singular Value Decomposition (SVD) to reduce the dimensions of the feature vectors while preserving the semantic similarities of abstracts [17]. Besides, cosine similarity is applied to compute the similarities of abstracts.

MeSH terms/text clustering and hotspots detection

With the similarity matrix of MeSH terms and abstracts construction, clustering algorithms are performed to discovery similar topics. According to experts' domain knowledge and practical experience, the TCM research hotpots can be induced in respective manners.

In the first scheme, a classical clustering algorithm k-means is utilized to group similar MeSH terms. MeSH terms are clustered into different categories, which are regarded as hotspot key words information. Then final highly summarized hotspots are concluded by experienced experts in detail.

Similarly in the second scheme, we adopt spectral clustering algorithm to group similar articles into same categories. According to clustering results, the MeSH terms of each article are taken as keywords, all high-frequency keywords are extracted for experts to make final hotspots deduced.

Results

Results based on MeSH terms co-occurrence

At the first step, we extract the annotated MeSH terms of each article. Then MeSH terms are ranked according to the value of their term frequency. Based on the discriminant formula of high-frequency and low-frequency terms we mentioned above, all terms whose frequency values lower than 75 are filtered. Finally, 283 high-frequency MeSH terms can be obtained. Moreover, a co-occurrence network is built to illustrate the relationships among MeSH terms, as figure 2.

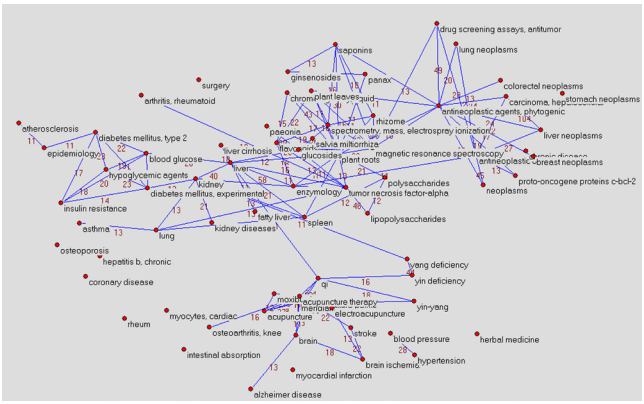

Figure 2: MeSH terms co-occurrence network. MeSH terms co-occurrence network reflects the relationships of MeSH terms.

In figure 2, the blue line values represent the frequency of two related terms appearing in the same article. Seen from the visual relationships of terms, we can easily divide the terms into four parts, circled with four different colors. Terms in the same part are similar with each other, and can reflect hot topics. This is also a considerable method to discovery hot topics. But this method is subjective and coarse which might not accurately reflect the research hotspots. Thus, it is more feasible to apply this network to calculate the similarities of all extracted terms appearing in the same article. Then, the k-means clustering algorithm can be performed to divide the data into 10 clusters. Final hotspots deduction is carried out by TCM experts.

Here, we list the first four primary hotspots discovered by our approach. The terms in parentheses are key words which can reflect the hotspots information as shown in figure 3.

Figure 3: TCM hotspots keywords. Keywords in figure reflect the hotspots of TCM. Larger the word size is, higher the word frequency is.

- Research on Chinese medical formula and its mechanism of action. (plants, medicinal; technology, pharmaceutical; seeds; fruit; rhizome; plant leaves; alkaloids; lignans; plant stems; saponins).
- Research on traditional therapy methods in TCM (mainly acupuncture).(acupuncture; acupuncture therapy; moxibustion; electroacupuncture; stroke; acupuncture points; meridians).
- Research on pharmacology and pharmacodynamics of antineoplastic agent in TCM. (antineoplastic agents; antineoplastic agents, phytogenic; drug screening assays, antitumor; neoplasms; liver neoplasms; carcinoma, hepatocellular; gene expression regulation, neoplastic; xenograft model antitumor assays; lung neoplasms).
- Research on the therapy of chronic diseases in TCM (mainly diabetes). (hypoglycemic agents; body weight; diabetes mellitus, type 2; blood glucose; diabetes mellitus, experimental; hypertension; brain ischemia; myocardium; blood pressure).

In order to illustrate the quantity distribution of hotspot articles, we count all available articles with respect to each hotspot according to MeSH terms, as shown in figure 4. It is obvious that the first four hotspots are far more on articles quantity than other research directions in TCM field. Besides, the articles quantity of the first hotspot articles are close to 1600, which is the dominant research direction against to the other three research hotspots.

Results based on text network

For this text network based scheme, the article abstracts are our research subjects. Hence, we need to process the text such as word segmentation, stop words filtering, words stemming and so on at first. Then through TFIDF text feature extraction technique, each abstract is represented as a numerical vector to build a vector space model. To reduce complexity and discovery the most relevant features, LSA is applied to reduce feature dimension and compute the articles similarities. The LSA parameter is set to 300. Furthermore, with applying spectral clustering algorithm, text network construction and

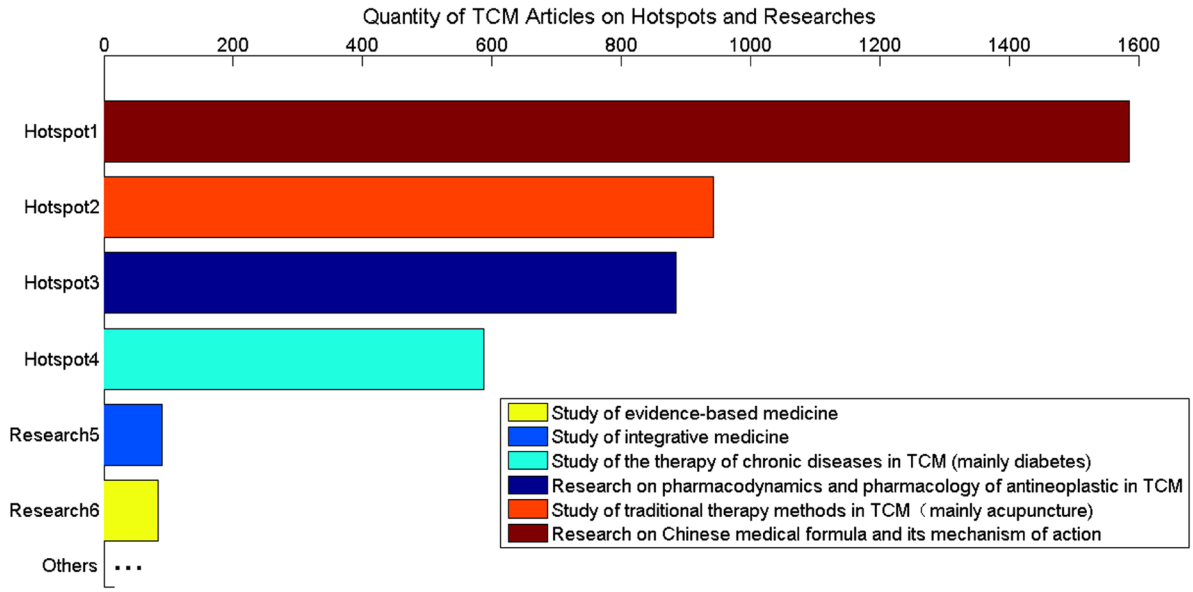

Figure 4: Quantity distribution of TCM articles on hotspots and researches based on MeSH terms co-occurrence.

community detection for the same topic can be carried out. It is hard for us to conclude hotspots from large chunks of article abstracts. Thus MeSH terms of each article are chosen as keywords to precede hotspots detection. In a community, if the frequency of a keyword is larger than 10, we preserve it. At last, according to the preserved keywords, we can obtain the conclusions what are the hotspot researches on TCM. The terms in parentheses are key words which can reflect the hotspots information.

- Research on Chinese medical formula and its mechanism of action. (plant extracts; *cordyceps*; plants, medicinal; rhizome; *coptis*; *salvia miltiorrhiza*; ginsenosides; reishi)
- Research on pharmacology and pharmacodynamics of antineoplastic agent in TCM. (antineoplastic agents; antineoplastic agents, phytogenic; cell line, tumor; apoptosis; gene expression regulation, neoplastic; carcinoma, hepatocellular; liver neoplasms; breast neoplasms; neoplasm invasiveness; lung neoplasms; carcinoma, non-small-cell lung; drug screening assays, antitumor; stomach neoplasms; tumor necrosis factor-alpha; neoplasm transplantation)
- Research on the therapy of chronic diseases in TCM (mainly diabetes). (kidney diseases; kidney; kidney failure, chronic; parkinson disease; blood; blood glucose; insulin resistance; liver cirrhosis, experimental; liver cirrhosis; alzheimer disease; cognition disorders; atherosclerosis; coronary disease; hepatitis b virus; hepatitis b, chronic; pulmonary disease, chronic obstructive; arthritis, rheumatoid; osteoporosis; diabetes mellitus, type 2; diabetes mellitus, experimental; insulin)
- Research on traditional therapy methods in TCM (mainly acupuncture).(acupuncture therapy; acupuncture points; meridians; pain measurement; pain management; acupuncture; electroacupuncture; moxibustion)

Similar to MeSH terms based hotspot detection scheme, quantity distribution of TCM articles on hotspots via text network is also presented, as shown in figure 5. These results are consistent with the previous results with respect to the first four hotspots, which validates the hotspots detection results are credible. In addition, the quantity distribution is a little bit different to the previous scheme. This denotes different article subjects might contain different topics information, which may lead to articles assigned to different hotspot researches.

From another perspective, TCM researches not only interest in the research hotspot during a certain period, but also concern the hotspots transfer in the next few years. So we also perform the experiments to explore the changes and development trends of hotspots on TCM field. There are 6076 articles from 2009 to 2011 and 6469 articles from 2012 to 2014. We take the methods mentioned above and compare the hotspots of the former 3 years and the latter 3 years. We find the main hotspots of the former three years and the latter three years are unchanged, but some slight transfer can be discovered. According to the percentages of articles on different hotspots in whole dataset of every three years, we illustrate the articles distribution on different hotspots from 2009 to 2011 and 2012 to 2014.

In figure 6, the bars denote the percentage of articles on different hotspots in whole dataset of every three years. Hotspot1 represents research on Chinese medical formula and its mechanism of action. Hotspot2 denotes research on pharmacology and pharmacodynamics of antineoplastic agent in TCM. Hotspot3 is research on the therapy of chronic diseases in TCM. Hotspot4 refers to research on traditional therapy methods in TCM. As time passed by, it is obvious that researchers tend to pay more attention on study of Chinese medical formula and its mechanism of action, and research on traditional therapy methods in TCM. Moreover, research on traditional therapy methods in TCM is increasing rapidly, it catches up with and surpasses the research on the therapy of chronic diseases in TCM. In summary, the dynamic development of TCM in past six years can be easily gained from figure 6.

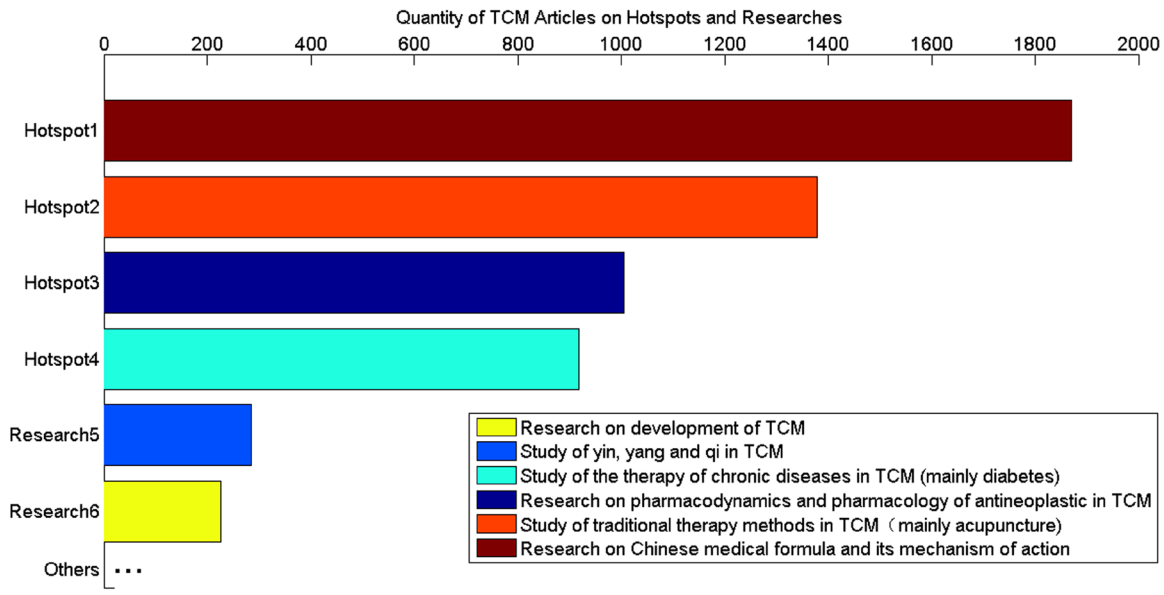

Figure 5: Quantity of TCM articles on hotspots and researches based on text network. The distribution of articles shows the hotspots of TCM.

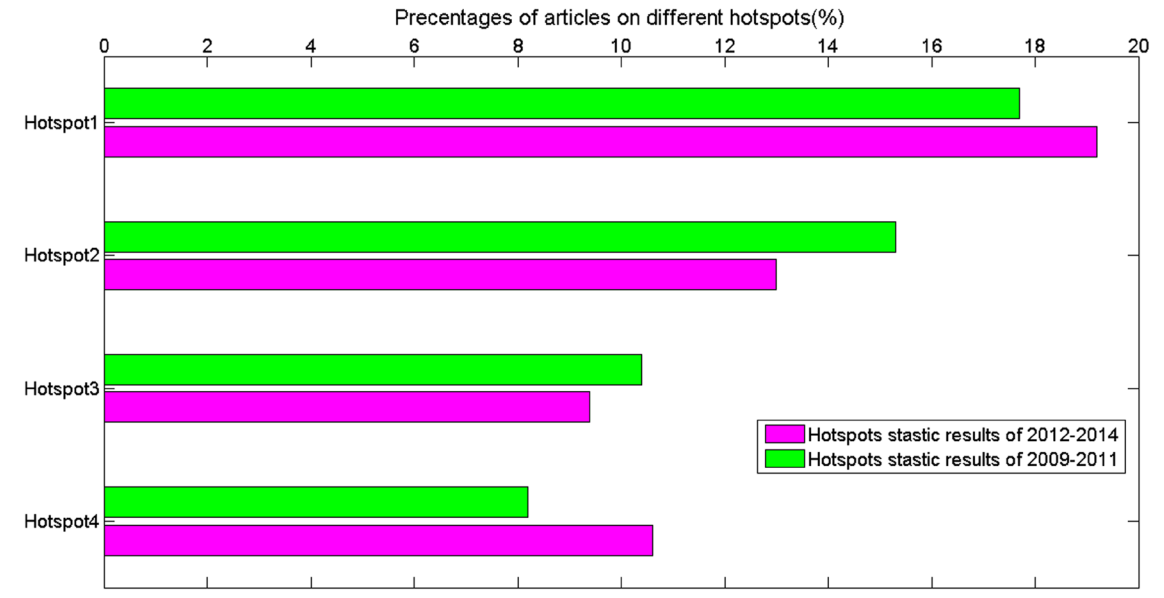

Figure 6: Hotspots comparison of 2009-2011 and 2012-2014. The percentage distribution of articles shows the dynamic development of hotspots of TCM.

Discussion

In this study, we adopt two schemes to detect the hotspot researches on TCM. To our satisfaction, these schemes are both effective to detect the hotspots. Although the schemes we adopt focus on different research data, both of them derive similar results, so it is reasonable to believe that the schemes are useful to detect the hotspots. Based on different schemes, we can find the quantities of articles on hotspots are different. The main reason is that different data subjects contain different topics information and different keywords selection approaches would extract different keywords representation. All these factors would slightly change article topics representation, which may finally result in the same article distributing to different hotspots with two proposed schemes. Moreover, the detected hotspots are also useful for us to make a preliminary understanding of the current research status and the hot research directions. For fresh researchers who want to devote to TCM researches, it is helpful for them to determine research direction according to the summarized hotspots. For experts who focus on TCM, the hotspots transferring over the recent years have a remarkable reference signification for them to master the trend of TCM developments. In our study, some drugs and diseases are clustered to the same category. The experts may discover

potential relationships between them and prescribe novel programs of treatment. We also visualize the data to improve the readability of the hotspots information. In addition, our methods performed on hotspots detection of TCM can also be used to other fields. Furthermore, it is considerable to construct a hotspots detection system to detect hotspots on academic field.

But there are still some limitations of our research. Firstly, the results we derived are almost words or phrases. There are still some difficulties to conclude sentences which can reflect the hotspots information. In most cases, our hotspots detection schemes need the expert domain knowledge and practical experience to summarize final results. Thus our research still cannot detect hotspots completely automatically without any manual intervention in some ways. In the future research, prior knowledge based hotspot detection should be developed to improve the current system. Meanwhile, it should further discuss that whether the five year span of the experiment data is appropriate or not. However, it is really meaningful for us to study the variation trends with data in different time spans.

Conclusion

In this study, we explore to detect the hotspots of researches on TCM, our research finds out several hotspots of TCM. Two schemes are developed in our research to detect the hotspots. One is based on MeSH terms and the other is based on text network. These two methods have several similar steps and the main distinct of them is that they focus on different data subjects with different feature representation and similarity computation. The two methods are effective to detect hotspots and helpful for researchers to acquire the dynamic development of TCM research directions. As what we mentioned above, there are still some limitations in our research. We prepare to improve the performance of our current schemes and design an automatic hotspots detection system in our further study.

Acknowledgment

This work was supported by the Natural Science Foundation of China (Grant No. 61273305).

References

1. Thornton PK, Jones PG, Owiyo T, Kruska RL, Herrero M, et al. (2008) Climate change and poverty in Africa: Mapping hotspots of vulnerability. African Journal of Agricultural and Resource Economics 2: 24-44.
2. Gimona A, Dan van der Horst (2007) Mapping hotspots of multiple landscape functions: a case study on farmland afforestation in Scotland. Landscape Ecology 22: 1255-1264.
3. Boyack KW (2004) Mapping knowledge domains: Characterizing PNAS. Proc Natl Acad Sci, 101: 5192-5199.
4. Steinbach M, Karypis G, Kumar V (2000) A comparison of document clustering techniques. KDD workshop on text mining 400: 525-526.
5. Congnan Luo, Yanjun Li, Soon M Chung (2009) Text document clustering based on neighbors. Data & Knowledge Engineering 68: 1271-1288.
6. Wang C, Zhang M, Ma S, Ru Liyun (2008) Automatic online news issue construction in web environment. Proceedings of the 17th international conference on World Wide Web, USA. ACM: 457-466.
7. Nowell L, Schulman R, Hix D (2002) Graphical encoding for information visualization: An Empirical Study. Proceedings of the IEEE Symposium on Information Visualization: 43- 50.
8. Daniel A Keim (2002) Information Visualization and Visual Data Mining. IEEE Transactions on Visualization and Computer Graphics 8: 1-8.
9. TT He, GZ Qu, SW Li, XH Tu, Y Zhang, et al. (2006) Semi-automatic hot event detection. Advanced Data Mining and Applications: 1008-1016.
10. N Li, DD Wu (2010) Using text mining and sentiment analysis for online forums hotspot detection and forecast. Decision Support Systems 48: 354-368.
11. Haribhakta Y, Malgaonkar A, Kulkarni P (2012) Unsupervised topic detection model and its application in text categorization. Proceedings of the CUBE International Information Technology Conference. ACM: 314-319.
12. Kim BY, Kang JS, Han JS, Jeon WK (2014) Mapping the dementia research area at the micro-level using co-terms analysis and positioning for traditional herbal medicine. Chin J Integr Med 20: 706-711.
13. Matsuo Y, Ishizuka M (2004) Keyword extraction from a single document using word co-occurrence statistical information. International Journal on Artificial Intelligence Tools 13: 157-169.
14. Peng C, Wei Z (2012) Co-word Analysis of Domestic Network Virtual Society Research Hotspots and Evolution. Management of e-Commerce and e-Government (ICMeCG), 2012 International Conference on. IEEE: 327-331.
15. Yung-Shen Lin, Jung-Yi Jiang, Shie-Jue Lee (2014) A similarity measure for text classification and clustering. IEEE Transactions on Knowledge & Data Engineering 26: 1575-1590.
16. Bradford RB (2008) An Empirical Study of Required Dimensionality for Large-scale Latent Semantic Indexing Applications. Proceedings of the 17th ACM conference on Information and knowledge management. ACM: 153-162.
17. Ma J, Xu W, Sun Y, Turban E, Wang S, et al. (2012) An Ontology-Based Text-Mining Method to Cluster Proposals for Research Project Selection. IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans 42: 784-790.