Discipline Hotspots Mining Based on Hierarchical Dirichlet Topic Clustering and Co-word Network

Ying Cai, Fang Huang*, Mengya Peng
School of Information Science and Engineering, Central South University, Changsha, China

* Corresponding author. Tel.: +86 138 7314 0019; email: hfang@csu.edu.cn

Manuscript submitted July 3, 2016; accepted August 30, 2016.
doi: 10.17706/jsw.11.11.1089-1101

Abstract: Discovering inherent correlations and hot research topics among various disciplines from massive scientific documents is very important to understand the scientific research tendency. The LDA (Latent Dirichlet Allocation) topic model can find topics from big data sets, but the number of topics must to be told before topic clustering. There is a lot of randomness to determine the number of topics for the unknown structure of data sets. Therefore, this paper introduces the Hierarchical Dirichlet Process (HDP) to achieve topic clustering with discipline division. Those clustering topics are composed by a discrete set of words, and these words do not have semantic relation. For this problem, this paper proposes a method to find out relationships between topic words so as to extract discipline hotspots. This method contains classifying topics with the co-occurrence of subject words, constructing co-word network and analyzing discipline hotspots with weak co-occurrence theory. The experiment results indicate that the Hierarchical Dirichlet Process can mine topic word-sets, and effectiveness better than the LDA topic model. The co-word network based on the weak tie theory can effectively find the discipline hotspots, which explicitly reflects the research hotspots and inherent connections of disciplines.

Key words: Co-word network, discipline research hotspots, hierarchical dirichlet Process (HDP), weak co-occurrence theory.

1. Introduction

Mutual penetration has been the main characteristic for the innovation of current modern science and technology. New theories, methods and technologies have been emerging from various subjects, and new research fields and topics have been arising. How to find out potential connections and hotspots among various subjects from the tremendous information is important for grasping the trend of science and technology and constructing a modern subject classification system. Most previous research work focuses on the discipline classification and correlation of science management. For example, Vessey, I. *et al.* [1], in 2005, designed a unified classification system for computing disciplines based on research-focused characteristics: topic, approach, method, unit of analysis, and reference discipline, according to the characteristics of knowledge overlapping and sharing in computer science. In the year of 2010, Istvan Z. K. *et al.* [2] applied the individual-based epidemic models to describe the topic diffusion of across disciplines, in order to find the knowledge flow among disciplines. In 2012, Baohong, W. *et al.* [3] compared the national standard disciplines classification system with the practical information classification system, and revealed the overall structure and development trend in natural science from the macro level. Jinzhu, Z. *et* *al.* [4] analyzed the cross disciplines of the library information field from the discipline quantity, distribution and differences with reference classification in the year of 2013, and visualized by superposition graph. To find out potential correlation and predict discipline classification, Lin, Z. et al. [5] proposed a hybrid mixed clustering algorithm based on "Citation–Text" fitting similarity to study the discipline classification and cross discipline structure in 2013. They divided all journals into "core journals" and "periphery journals" and used "periphery journals" to explain the structure characteristics of disciplinary decomposition and cross. In 2016, Julianite, F. et al. [6] have conducted a statistical analysis of the cited articles, keywords and authors of the articles recently published in major publications related to public service management. It produces the systematic knowledge from scientific publications, identifies the scientific gaps, and provides existing theoretical expansion and innovation model in the future researches. With the development of science, to find out the inner connection and trend is still an open research problem. In recent years, numerous scientific documents, such as fund applications, progress reports, concluding reports and so on, are submitted via web platforms, which contain important trends for research fields. It is the current important research subject to use method and technology involved in big data knowledge discovery to mine cross-fields hotspots from the massive project documents.

In the paper, we focus on the topic extraction and hotspot mining. The proposed scheme is composed of the topic extraction with hierarchical Dirichlet process, topic classify with subject-content words, and discipline hotspot analysis with co-word network with weak co-occurrence theory. We apply the aforementioned theory and method to the process of big text data knowledge discovery and design the solution for analyzing the co-word network and research topics.

The rest of the paper is organized as follows: Section 2 provides an extensive review on related work. Section 3 displays the procedure of topic extraction and hotspots analysis. Section 4 introduces the method of hierarchical Dirichlet process, which can extract topic from documents, and gives the evaluation. Section 5 introduces the process of mining hotspots and displays experiment results. Section 6 finally summarizes the concluding remarks.

2. Related Work

Generally, the authors of fund applications and project report will thoroughly investigate the relevant research activities. The applications usually reflect the most recent research problems, theories, methodologies and technologies. Topic model has been a very good application and development in the field of text mining, such as text categorization, topic search and topic evolution etc. Its fundamental ideal is to adjust the parameters of the topic model to find out topics, which mining text from semantic deeply and implicitly.

The first topic model is Latent Semantic Indexing (LSI) model, which was proposed by Deerwester, S. C. *et al*. [7] in 1990. This model is not a real sense of probabilistic topic model, but it lays the foundation for the development of topic model. Then, 1999, Hufmann, T. [8] improved the LSI model and proposed a new topic model called the probabilistic latent semantic indexing (pLSI), which is considered to be the first probabilistic model and promotes the development of the topic model. After 2003, David, M. B. et al. [9] proposed the latent Dirichlet allocation on the basis of pLSI and this model has been widely used. For example, Rosen-Zvi, M. et al. [10] applied the LDA to analyze technical documents in the year of 2004, and proposed an author-topic model. This model can find relationships between authors and topics and reveal researchers' interests and preferences. Teh, Y. W. et al. [11] in 2006, proposed a Hierarchical Dirichlet Process (HDP) to improve the LDA model need to manually determine the topic number shortcomings and applied it to discover the evolution of topics. HDP is a generation model based on Dirichlet process, and using construction method of Stick-breaking, Polya Um or Chinese Restaurant Process (CRP) to realize

Dirichlet process. Compared with traditional TF-IDF statistical model, HDP can excavate the deep semantic information, and get “document-topic-word” three-layer model by mapping the text vector to the topic feature space. Compared with the LDA, HDP is able to dynamically determine the topic number and get topics from documents. HDP and Dirichlet process are also widely used. For example, in 2008, Canini, K. R. *et al*. [12] used HDP and HP to analyze the role of topic sharing in human category learning and transfer learning problem in psychology research. Li, X. *et al*. [13] fully used the Dirichlet process mixture model as a priori distribution, try to achieve video surveillance based on trajectory information of video retrieval. In 2015, Di, W. *et al.* [14] designed an incremental learning method with partial supervision for HDP, which guide the topic model from partial knowledge to gradually adapt to the latest available information. Xianghua, F. *et al*. [15] added sentiment level to the HDP model and used the non-parametric hierarchical Dirichlet model to dynamically track the sentiment and topics in reviews. In 2016, Xianghua, F. *et al.* [16] proposed a dynamic online hierarchical Dirichlet process to discover the topic evolution of Chinese social data. With the extensive application of HDP in natural language processing, topic extraction and document mining, the learning model based on HDP is still a hot research issue.

As a method of content analysis, co-word analysis is based on statistic the number of two words appeared in a document. It reflects the relationships between words, and we can analyze topic structural changes [17]. Here, we use co-word analysis to build co-word network, display topic words relationships and get discipline hotspots.

Weak tie theory was first used to analyze the impact of human intelligence network [18]. Later, it was used for information sharing [19] and the technological innovation analyzing [20]. In social relation, weak tie has the following three features: universality, heterogeneous and unstructured [21]-[23]. Universality refers to the enormous number individual of weak tie, and anyone other than yourself can constitute a weak relationship with you. Heterogeneous means that there is a big difference, such as thoughts, education, social position etc., between individual and individual in a weak relationship. Unstructured refers to a node in weak tie can play the role of bridge to transfer information between different groups or subgroups. Applied it to text analysis, weak co-occurrence network can better discover the relationship between topic words which may be the topic keyword, and discover not prominent research topics which may evolve into hotspots.

3. Procedure of Discipline Hotspots Analysis

The procedure of analysis discipline hotspots mainly includes generating valid subsets of data, topic clustering and post-processing of topic clustering, namely, feature compression in the scientific documents, topic clustering of discipline and mining hotspots. The details of the process are shown in Fig. 1.

3.1. Document Feature Compression

Feature selection is necessary to form an effective science and technology documents subset. The document structured data should be extracted and segmented into separated words by using the Chinese word segmentation system ICTCLAS. With a prohibited word database provided in this system, word de-noising and filtering could be implemented to get rid of those erroneous segmentation and non-academic words. Feature vector sets can be generated from weight calculation and ranking selection.

3.2. Topic Clustering and Discipline Division

The procedure of topic clustering and discipline division contains two steps: one is the topic clustering, the other is the topic discipline division. The Hierarchical Dirichlet Process was introduced to determine the number of topics and get topics of documents. At the same time, we choose a group of words with discipline abstract characteristic as the subject content word, and the topic is divided into the subject by the high co-occurrence of subject words and topic words.

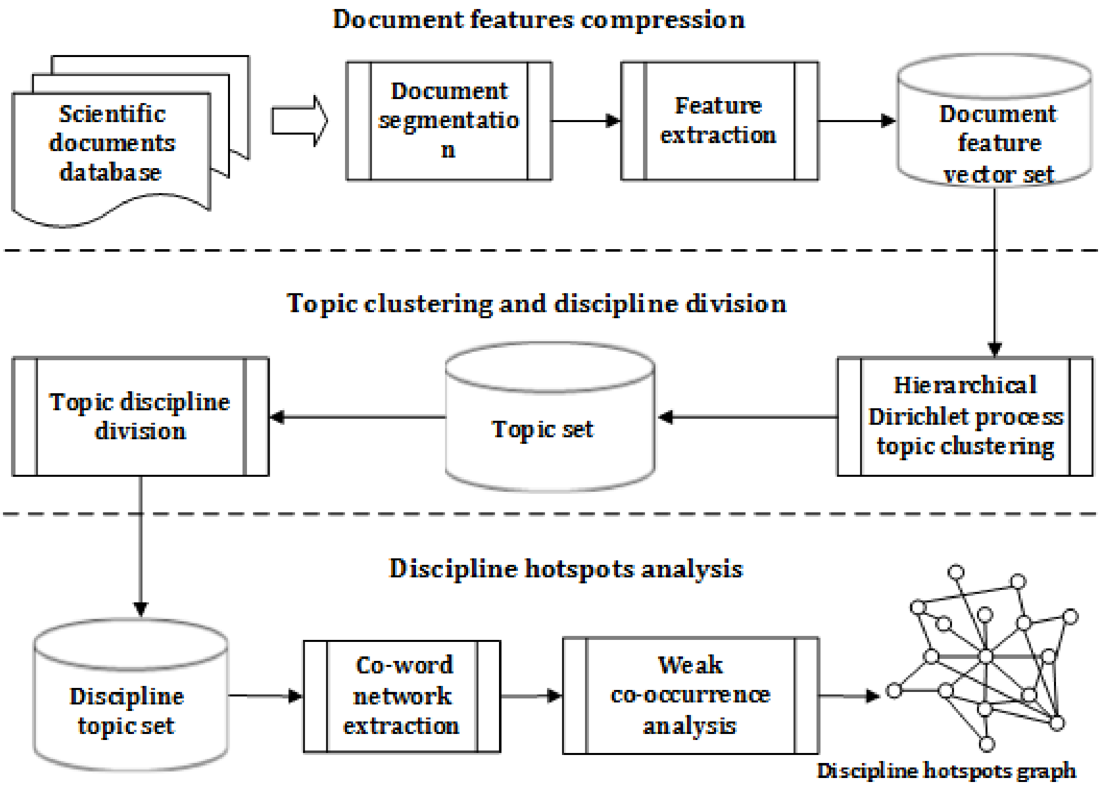

Document features compression
Scientific documents database
Document segmentation
Feature extraction
Document feature vector set
Topic clustering and discipline division
Topic discipline division
Topic set
Hierarchical Dirichlet process topic clustering
Discipline hotspots analysis
Discipline topic set
Co-word network extraction
Weak co-occurrence analysis
Discipline hotspots graph

Fig. 1. Procedure of discipline hotspots analysis.

3.3 Discipline Hotspots Analysis

In the process of hotspot mining, the co-word network is established to find out the co-occurrence relationship between topic words. Then through the heuristic threshold filtering to simplify the co-word network, using the weak co-occurrence theory to extract the discipline hotspots and form discipline hotspots map.

4. Topic Discipline Division Based on Hierarchical Dirichlet Process

Currently, LDA topic model is a widely used method of topic extraction in text analysis, but it needs to give cluster number in advance. For the large amount and variable data sets, it consumes a lot of effort to determine the number of clusters in LDA topic model. In this paper, Hierarchical Dirichlet Process [11] is used for topic modeling, which can automatically calculate the number of clusters according to the data, lead to simplify the parameters of modeling and to increase the usefulness of the topic model. However, these topics do not have the characteristics of the subject in order to analyze research hotspots from discipline level. A method of topic discipline division based on co-occurrence is proposed to divide topic into discipline.

4.1. Hierarchical Dirichlet Process and Construction

Hierarchical Dirichlet Process is essentially a multi-Dirichlet process, which is an important tool for text mining. One Dirichlet process is a Dirichlet distribution, which is based on an assumption that all documents share the same topic sets, and the number of clusters is unlimited, which can be automatically inferred according to the document set. Compared with the parametric Bayesian model, HDP has good robustness and flexibility.

In this paper, the topic model is the two layers Dirichlet processes. Firstly, the sampling distribution of the whole document set G_0 in the Dirichlet process is composed of the base distribution H and the Concentration parameter γ. The base distribution H is a Dirichlet distribution with parameter η. Then, the distribution of topics for each document G_j is composed of the distribution G_0 and the Concentration parameter α_0. The specific formula is as follows.

$$\begin{aligned} &G_0 \sim DP(\gamma, H) \\ &G_j \mid G_0 \sim DP(\alpha_0, G_0) \end{aligned} \tag{1}$$

From the formula (1), the topics of each document are following the base distribution H, which ensure that all documents share the same topic sets. Each topic ϕ_k is an independent sample of H, which is in essence the probability distribution of words in topics. In a document x_j: $(\theta_{ji})_{i=1}^{N_j}$ is an Independent and Identically Distributed sequence of random variables following G_j, and θ_{ji} represents the probability that word x_{ji} belongs to topic ϕ_k. The generation process of the document x_j is as follows:

$$\begin{aligned} &\theta_{ji} \sim G_j \\ &x_{ji} \mid \theta_{ji} \sim F(\theta_{ji}) \end{aligned} \tag{2}$$

In the formula, $F(\theta_{ji})$ represents the word x_{ji} distribution with the parameter θ_{ji}. The parameter θ_{ji} is following the distribution G_j independently, and x_{ji} is condition independently obeying the distribution $F(\theta_{ji})$.

Suppose the document is composed of topics, and the topic is composed of words [24]. Assuming J documents share with the same topic sets: $\phi = (\varphi_k)_{k-1}^{K}$, K represents the topic number. The document j contains m_j topics: $(\varphi_{jt})_{t=1}^{m_j}$, and each topic contains N_j words. The word x_{ji} has a probability of $\frac{n_{jt}}{i-1+\alpha_0}$ to choose an existing topic φ_{jt}, distribution parameter θ_{ji}, and a probability of $\frac{\alpha_0}{i-1+\alpha_0}$ to choose a new one $\varphi_{jt_{new}}$. The word is chosen by the following formula (3).

$$\theta_{ji} \mid \theta_{j1}, \theta_{j2}, \ldots, \theta_{ji-1}, \alpha_0, G_0 \sim \sum_{t=1}^{m_j} \frac{n_{jt}}{i-1+\alpha_0} \delta_{\varphi_{jt}} + \frac{\alpha_0}{i-1+\alpha_0} G_0 \tag{3}$$

In this formula, n_{jt} is the word number of the topic t in the document j, i indicates the number of words that have been selected for topics, α_0 is the parameter to control the generation of new topics, and $\delta_{\varphi_{jt}}$ is the word distribution of the topic φ_{jt}. The probability distribution of a new topic is represented as G_0. The whole process of topic generation is a Dirichlet process.

If the document is made up of topics, the process of generating a document is similar to the generation of the topic, it is also a Dirichlet process. The document has the probability $\frac{m_k}{\sum_k m_k + \gamma}$ to choose an existing topic ϕ_k, and according to the probability $\frac{\gamma}{\sum_k m_k + \gamma}$ to choose a new topic $\phi_{k_{new}}$.

$$\varphi_{jt} \mid \varphi_{11}, \varphi_{12}, \ldots, \varphi_{21}, \ldots, \varphi_{jt-1}, \gamma, H \sim \sum_{k=1}^{K} \frac{m_k}{\sum_k m_k + \gamma} \delta_{\phi_k} + \frac{\gamma}{\sum_k m_k + \gamma} H \tag{4}$$

In formula (4), m_k is the number of documents that contain topic ϕ_k, m is the topic number of selected topics.

After Hierarchical Dirichlet Process, we get topic-word distribution matrix and document-topic distribution matrix, and the document is represented by a collection of topics and the topic by a collection of words. The topic can be formalized as:

$$t_i = (w_1, p_1; w_2, p_2; \ldots; w_l, p_l) \tag{5}$$

Here, t_i is taken as the i topic, w_l is as the l word in topic t_i, and p_l represent the probability that the word w_l belongs to the topic t_i.

4.2. Topic Clustering Experiment Analysis

We randomly choose 530 project applications to test the performance. First of all, segmenting documents into separated words by using the Chinese word segmentation system ICTCLAS [25]. With a prohibited word database provided in this system, word de-noising and filtering could be implemented to get rid of those erroneous segmentation and non-academic words. Then, generating document feature vector sets with weight calculation of TF-IDF algorithm and sort selection. The experimental results show that 142 topics are generated from the HDP model.

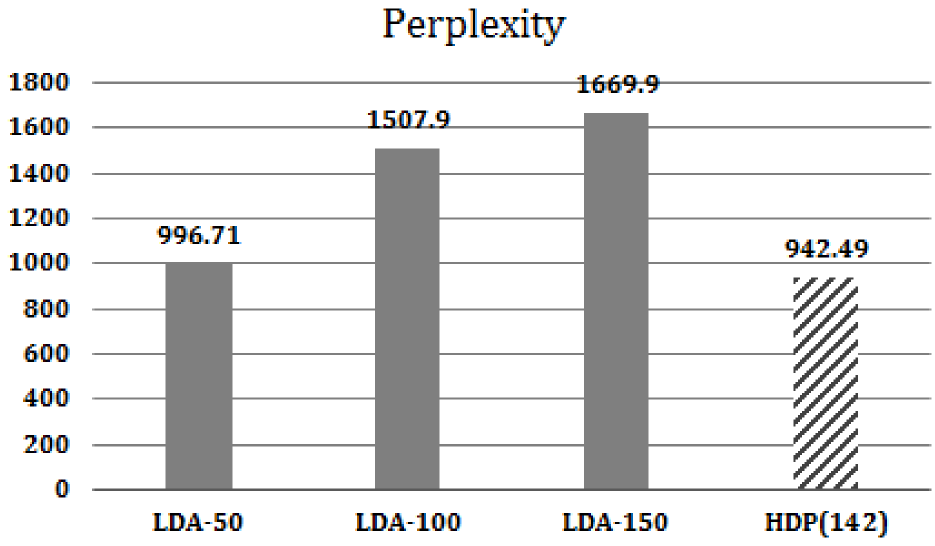


Fig. 2. Comparison of clustering effects of different topic models.

4.2.1. Topic model assessment

In order to evaluate the clustering effect of the HDP topic model, we use the perplexity value to analyze the clustering results. Perplexity is one of the commonly used evaluation methods of the topic model, which is essentially a kind of information theory. A low perplexity indicates the topic model is good at clustering topic. In the experiment, we use the HDP topic model to compare with the LDA model of different cluster numbers. We take the k value of 50, 100, 150 to calculate the perplexity value, which are expressed as LDA-50, LDA-100, and LDA-150. As shown in Fig. 2, the HDP model cluster 142 topics and perplexity value is 942.9, it is smaller than the perplexity value of LDA-150 1669.9, which indicates that the HDP model performs better than LDA model. For the same algorithm LDA, perplexity value increases with the increasing in the number of topics. The smaller the number of topics means the performance better. But for large-scale data, the smaller number of topics may not cover all topics and ignore some similarity topics. Therefore, the HDP has a good performance in topic clustering, for it can automatically learn the number of clusters according to the data sets, and lower perplexity than the same cluster number of LDA model.

4.2.2. Topic clustering analysis

Table 1 shows some results of topic clustering using HDP model. From Table 1, the HDP model can cluster documents into clear topics. This topic model can identify potential information in large scale documents, rather than simple statistical analysis of original documents. It can reflect the document content by the knowledge level.

Table 1. Topic Clustering Based on Hierarchical Dirichlet Process

Topic1	Topic2	Topic3	Topic4	Topic5	Topic6	...
decision	strategy	soil	optimize	Cell	schistose	...
worm	credit	pollute	soil	Apoptosis	structure	...
method	model	Cd	method	interface	tombarthite	...
criterion	industry	Pb	cover	Cr	material	...
structure	bank	As	ecology	mitochondria	market	...
network	bloc	heavy metal	inequality	Guide	kentanium	...
information	method	vegetables	area	Organic	preparation	...
system	extract	polymer	model	Part	cell	...
confrontation	efficiency	polycarbonate	environment	relation	gene	...
macrograph	risk	fat	remote sense	dependence	vole	...
model	structure	content	ninja	influence	audit	...
Magnetize	risk management	medicine	algorithm	Jump	express	...
policy	commerce	chemistry	system	Height	influence	...
...	...	...	...	...	...	...

In the Table 1, topic3 and topic4, topic5 and topic6 are similar topics separately. Topic3 and topic4 are mainly about the ecological environment, and topic5 and topic6 are in the biological sciences. While topic1 is in information science, and topic2 is in management science.

4.3. Topic Discipline Division

The HDP model puts 530 documents into 142 topics, which reflect the document by the knowledge level, but these topics do not have discipline characteristics, unable to analysis hotspots of discipline. To achieve topics discipline division, the co-occurrence matching method is proposed to divide topics into the disciplines. First of all, we have chosen a group of words with typical discipline characteristics as subject context words, which are classified according to the discipline classification code of the application project, and adding article key words of each category to enrich the subject context word. Then, calculating the percentage of one topic contains subject context words in each category, choosing the top one as its topic category. The part of the results of topic discipline division is shown in Table 2.

5. Hotspots Analysis Based on Co-word Network and Weak Co-occurrence Theory

Topic clustering through the hierarchical Dirichlet process from documents is a collection of discrete topic words, and there is no determinate connection between topic words. In order to find out hotspots by knowledge semantic level from subject topic sets, we use the co-occurrence relationship between two words in a document to construct relationships between the topic words. Meanwhile, we use the heuristic threshold to filter out the low degree co-occurrence relationship, and establish a co-word network based on the topic words of the discipline feature. The co-word network is a mapping of the subject hotspot knowledge graph. However, there are also some academic common words, such as "ratio", "method", "identification", "model", "compute" etc., which cannot accurately analyze the subject hotspots in the co-word network. These common words are often high-frequency words, which lead to a high degree of co-occurrence. So, the co-word network contains many high co-occurrence relationships, which do not have obvious discipline characteristics. To solve this problem, we use the weak tie theory in the co-word network to filtering out high co-occurrence network, and form the weak co-occurrence network. This weak co-occurrence network is composed of subject topic words and co-occurrence relations which have the obvious characteristic of discipline. Finally, discipline hotspots are extracted by visual analysis.

5.1. Co-word Network

Co-word network is formed by the frequent contacts of two words appeared in the same article, which is named as DF (Document Frequency). The formula for calculating the average number of DF is as follows:

$$Avg\_df(t_k) = \frac{\sum_{i=1}^{N}\sum_{j=i+1}^{N} df(w_i, w_j)}{C_N^2} \tag{6}$$

t_k represents the topic k, w_i is the i word in topic k, C_N^2 as the number of word pairs in the topic of N words, $df(w_i, w_j)$ indicates the number of articles which contain word w_i and word w_j, and $Avg\_df(t_k)$ is the average number of all word pairs co-occurrence in the network.

Hotspots can be found in the co-word network by the heuristic threshold filtering. The threshold is calculated by the expression (7), and r is a heuristic factor. Usually, the threshold thr is 1.1-1.3 times in the average value empirically.

$$thr = r * Avg\_df(t_k) \tag{7}$$

5.2. Weak Co-occurrence Network Extraction

In relationship networks, the strong tie usually represents a stable relationship between actors with high level of interaction. Consequently, the information contained in the strong tie is usually with redundant and high repetitive and it is easy to become a closed system. On the contrary, weak ties are able to transmit non repetitive information among different groups, which implies new information with low redundancy and high distinguishability. We introduce the weak tie theory into the text analysis to find the subject hotspots by means of the weak connection between the topic words.

In the topic co-word network of the discipline, there are also some words in strong tie relationship with high degree of co-occurrence, such as "target", "subject", "parameter", "model", "compute" etc., which have obvious academic universality, that is, the information implied in the relations has a weak discriminability for the disciplines. To reduce the influence of these common words with strong tie, and more accurately extract the subject hotspots with obvious characteristics. We use the weak tie theory to filter out the high degree co-occurrence academic common words in co-word network. Specific metric is shown in (8).

$$thr_{weak} = \lambda * Avg\_df(t_k) \quad (8)$$

The general value of the weak contribution heuristic factor λ is 10-13 usually, which means the topic word co-occurrence degree is not more than 10-13 times of the average value in the co-word network, consequently, the high degree co-occurrence words are filtered through the threshold thr_{weak}. The remains is called as weak co-occurrence network.

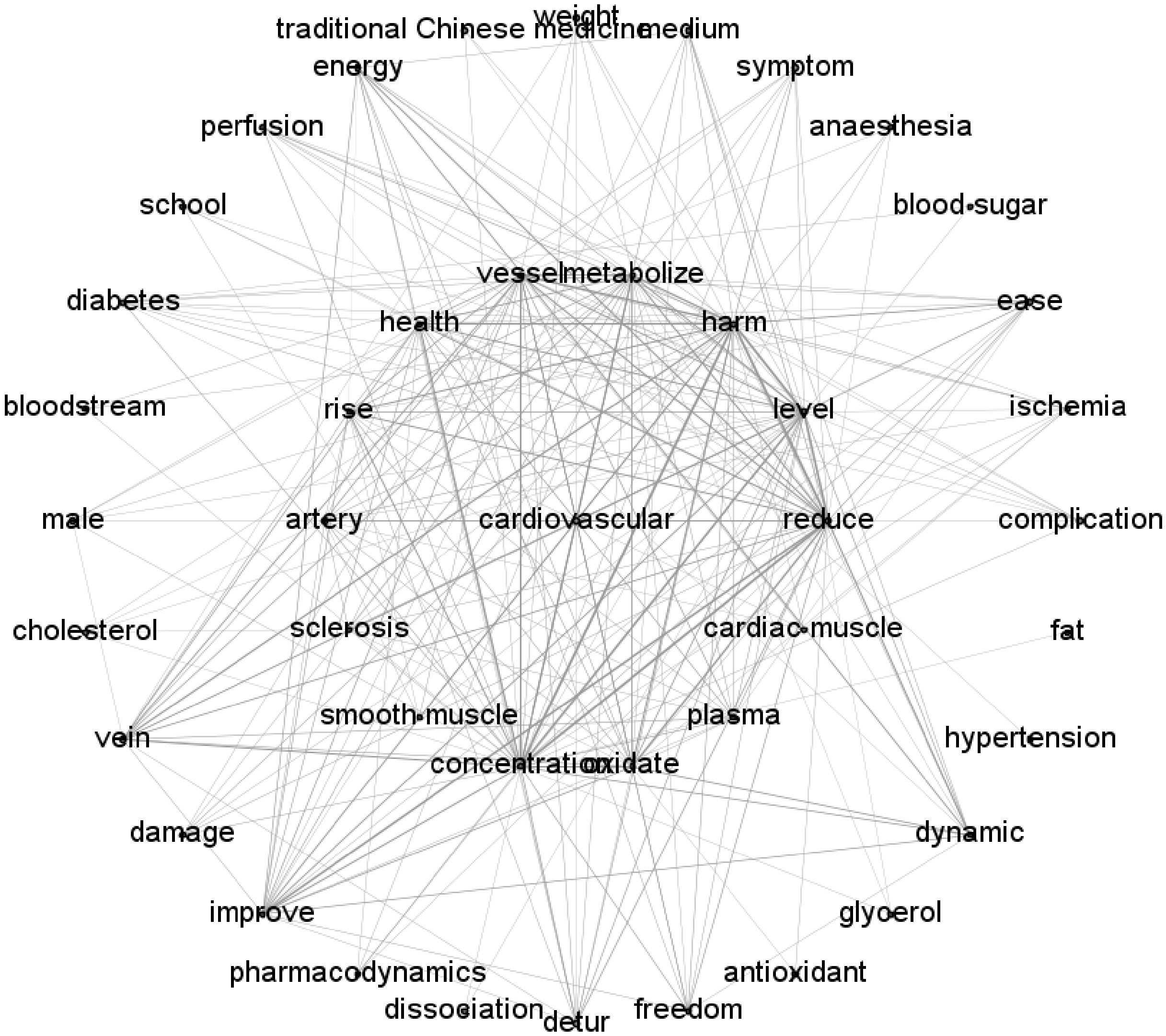
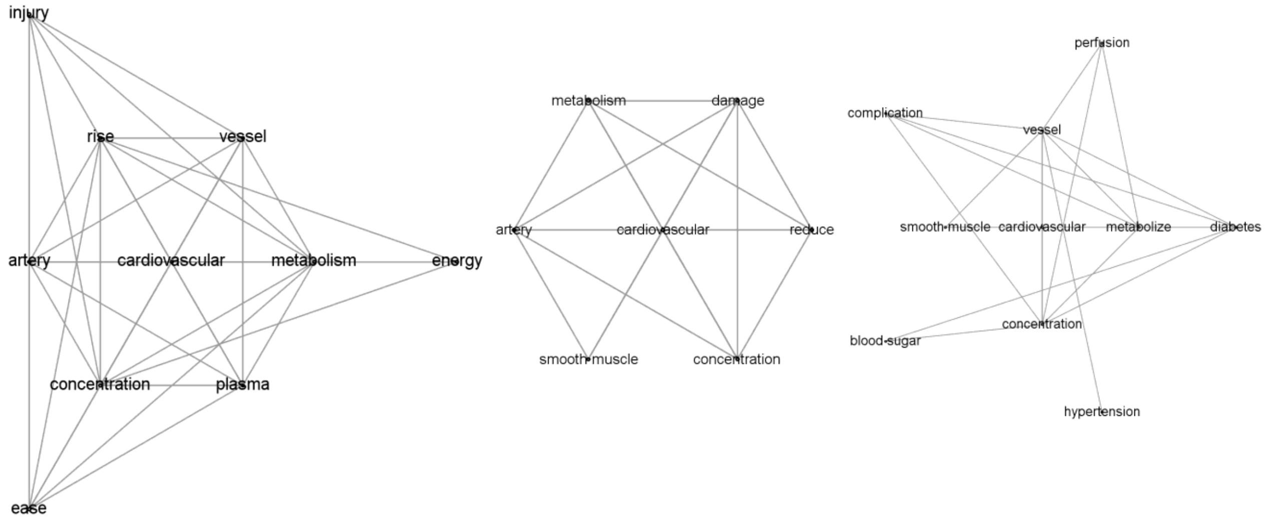


Fig. 3. Life Science topic map.

5.3. Visual Analysis of Discipline Hotspots for the Weak Co-occurrence Network

By constructing co-word network and weak co-occurrence extraction, we obtain the knowledge map with subject characteristics. The knowledge network graph is built by the co-occurrence relationship between topic words. In the co-word network the sub graph formed by topological relations presents a research hotspot in a subject topic. In order to facilitate the people according to their own knowledge background to intuitively observe and discover hotspots in the subject topic graph, we employee the visualization tool Gephi to display the co-word network of the subject topic. In Gephi, the concentric layout is a commonly used method of network visualization. The concentric layout is based on the hops between nodes and the center to determine the node's location. The choice of the center can be the max degree node or a specially appointed node. All nodes are distributed around the center in a ring form, and the N distance of nodes from the center is laid in the first N ring, which means that the closer the center node is, the more relevant to the center node. In the experiment, the r value is 1.2 and the λ value is 10.



(a) (b) (c)

Fig. 4. Hotspots sub-graphs.

We selected a topic co-word network of Life Science as an example. The topic map shown in Fig. 3, it is the result of network visualization through Gephi concentric layout. We choose the topical node "cardiovascular" as the center, the other words distributed around it in a ring. The nodes on the first ring are directly associated with cardiovascular, such as "artery", "plasma", "oxidation", "damage", "cardiac muscle" and so on. Nodes on the second ring are indirectly associated with cardiovascular, such as "cholesterol", "complication", "hypertension", "diabetes" and so on, which are related to diseases or medicine. Through the graphical presentation, it is easily to draw a number of cardiovascular related research topics, which form research hotspots of Life Science.

Fig. 4 contains three hotspot sub-graphs starting from the center node "cardiovascular". Fig. 4 (a) shows the sub-graph contains "cardiovascular", "plasma", "vessel", "concentration" etc., it reflects the research focuses on the influence of blood plasma concentration on the cardiovascular system; Fig. 4 (b) includes the topic words "cardiovascular", "cardiac muscle", "smooth muscle", "artery, metabolize" etc., which indicates the effect of arterial metabolism and heart muscle on cardiovascular diseases. Fig. 4 (c), "cardiovascular", "diabetes", "blood sugar", "concentration", "metabolism", "complication", "vessel" etc., reveals the influence of diabetes on cardiovascular diseases. In addition, Fig. 3 can extract other hotspots, such as sub-graph "cholesterol", "fat", "glycerin", "atherosclerosis", "cardiovascular", "arterial", and presents the link between cholesterol and cardiovascular disease. Sub-graph "Pharmacology", "index", "health", "Chinese medicine", and "injury" is another research focus on the influence of traditional Chinese medicine on human health. Through the visual analysis of the subject hotspots words network, it can be seen that cardiovascular disease is a focus of attention in Life Science. This research focuses not only on the pathology of cardiovascular disease, but also combine with other disease and methods to carry out comprehensive research. However, some irrelevant words appear in the weak co-occurrence network, such as “schools”, “freedom”, “male”, “improve”, “damage” etc., these words are common words without the academic characteristics, which are the background words in the academic document. The appearance of these words is because of the inaccurate extraction of the document features, which likely cause confusion on the understanding hotspot sub-graph.

6. Conclusion

For the growing accumulation of research documents, we employ the process of big text data knowledge discovery to mine the discipline hotspots implied in these documents. For the science and technology project applications, the topic clustering based on Dirichlet process is introduced to realize the effective extraction of discipline research topic in documents. On the basis of this, we propose a co-word network based on the weak tie theory to build weak co-occurrence network by threshold filtering, and extract subject hotspots through the visual analysis for the co-word network, and reached the goal of revealing the current research issues from the documents of scientific project applications based on the knowledge semantic level. The experiment results show that the proposed scheme is feasible and effective for the hotspots extraction from the project documents. The research problem and method involved in this paper can discover the internal relations of disciplines in a deeper level, which lay the foundation for understanding the trend of the modern discipline development. The hotspots extraction is through the visualization of the subject topic words network, which still need human knowledge structure and observation. Therefore, automatic recognition hotspots from the co-word network will be our next work in the future.

Acknowledgment

This work was supported by Project 61073105 of National Natural Science Foundation of China and Project 2016JC2011 of Science and Technology Plan of Hunan Province.

References

[1] Vessey, I., Ramesh, V., & Rovert, L. G. (2005). A unified classification system for research in the computing disciplines. *Information and Software Technology*, 245-255.

[2] Istvan, Z. K., Broom, M., Paul, G. C., & Rafols, I. (2010). Can epidemic models describe the diffusion of topic across disciplines? *Journal of Informetrics*, 74-82.

[3] Baohong, W., & Yidong, W. (2012). Development trends of nature science discipline according to classification systems of science disciplines. *Information Science*, *30(6)*, 930-936.

[4] Jinzhu, Z., Tao, H., & Xiaomei, W. (2013). An analysis on interdisciplinary of library and information science based on references’ subject category. *Library and Information Service*, 57(1), 108-111.

[5] Lin, Z. (2013). Analysis of subject classification and interdisciplinary structure based on journal hybrid clustering. *Library and Information Service*, *57(3)*, 78-84.

[6] Juliani, F., & Otávio, J. O. (2016). State of research on public service management: Identifying scientific gaps from a bibliometric study. *International Journal of Information Management*, 1033-1041.

[7] Deerwester, S. C., Dumais, S. T., & Landauer, T. K. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science* (JASIS), *41(6)*, 391-407.

[8] Hofmann, T. (1999). Probabilistic latent semantic indexing. *Proceedings of the 22nd annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 50-57).

[9] David, M. B., Andrew, Y. N., & Michael, I. J. (2003). Latent dirichlet allocation. *The Journal of Machine*

Learning Research, *3*, 993-1022.

[10] Rosen-Zvi, M., Griffiths, T., & Steyvers, M. (2004). The author topic model for authors and documents. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (pp. 487-494).

[11] The, Y. W., Jordan, M. I., & Beal, M. J. (2006). Hierarchical dirichlet processes. *Journal of the American Statistical Association*, *101(476)*, 1566-1581.

[12] Canini, K. R., & Griffithes, T. L. (2008). The hierarchical Dirichlet process as a model of human category learning. *Proceedings of the Workshop on Machine Learning Meets Human Learning*, Vancouver, Canada: The MIT Press.

[13] Li, X., Hu, W., Zhang, X., & Luo, G. (2008). Trajectory-based video retrieval using Dirichlet process mixture models. *Proceedings of British Machine Vision Conference*. Leeds, UK: The British Machine Vision Association (pp. 1-10).

[14] Di, W., & Ahmad, A. R. (2015). Incremental learning with partial-supervision based on hierarchical Dirichlet process and the application for document classification. *Applied Soft Computing*, 250-262.

[15] Xianghua, F., Kun, Y., Joshua, Zhexue, H., & Laizhong, C. (2015). Dynamic non-parametric joint sentiment topic mixture model. *Knowledge-Based Systems*, 102-114.

[16] Xianghua, F., Jianqiang, L., Kun, Y., Laizhong, C., & Lei, Y. (2015). Dynamic online HDP model for discovering evolutionary topics from Chinese social texts. *Neurocomputing*, 412-424.

[17] Qin, H. (1999). Knowledge discovery through co-word analysis. *Library Trend*, *48(1)*, 133-159.

[18] Granovetter, M. S. (1973). The Strength of Weak Ties. *American Journal of Sociology*, *78(6)*, 1360-1380.

[19] Chunmei, W., & Xiaoping, S. (2014). A research review of theory of weak relation and strong relation and the applications in information sharing. *Library*, *4*, 18-21.

[20] Ning, C., & Song-ting, P. (2008). The coupling relationship and synchrony evolving between strength of network tie and mode of innovation. *China Industrial Economics*, *4*, 137-144.

[21] Friedkin, N. E. (1982). Information flow through strong and weak ties in intraorganizational social networks. *Social Networks*, *3(4)*, 273-285.

[22] Liu, W. T., & Duff, R. W. (1972). The strength in weak ties. *The Public Opinion Quarterly*, *36(3)*, 361-366.

[23] Ryberg, T., & Larsen, M. C. (2008). Networked identities: Understanding relationships between strong and weak ties in networked environments. *Journal of Computer Assisted Learning*, *24(2)*, 1031-115.

[24] Zhimin, Z., & Shenyong, G. (2014). A survery on hierarchical dirichlet process principle and its applications. *Computer Applications and Software, 31(8)*, 1-5.

[25] Hua-Ping, Z., Hong-Kui, Y., De-Yi, X., & Qun, L. (2003). HHMM-based Chinese lexical analyzer ICTCLAS. *Second SIGHAN Workshop Affiliated with 41th ACL*.

Ying Cai was born in Hubei, China, in 1991, and she received the BSc degree in software engineering from Hubei University of Economics, China, in 2013. Currently, she is a master student with the School of Information Science and Engineering, Central South University, China. Her research interest is text topic modeling, visual analysis and text mining.

Fang Huang was born in Changsha, China, in 1963, and she received the PhD degree from Central South University, China, 2007. She is currently a professor in School of Information Science and Engineering of Central South University, China. Her research interests include social network mining and analysis, data mining and knowledge discovery, and big data analysis.

Mengya Peng was born in Hebei, China, in 1992, and she received the BSc degree in Software Engineering from Zhejiang University of Science and Technology, China, in 2013. Currently, she is a master student with the School of Software, Central South University, China. Her research interest is social network mining and analysis, text information extraction.