Physica A 450 (2016) 657–669

Contents lists available at ScienceDirect

Physica A

journal homepage: www.elsevier.com/locate/physa

Evolutionary features of academic articles co-keyword network and keywords co-occurrence network: Based on two-mode affiliation network

Huajiao Li [a,b,c,d,*], Haizhong An [a,b,d], Yue Wang [a,e], Jiachen Huang [f], Xiangyun Gao [a,b,d]

[a] School of Humanities and Economic Management, China University of Geosciences, Beijing 100083, China

[b] Key Laboratory of Carrying Capacity Assessment for Resource and Environment, Ministry of Land and Resources, Beijing 100083, China

[c] Department of Energy and Mineral Engineering in the College of Earth and Mineral Sciences, The Pennsylvania State University, PA 16802, USA

[d] Lab of Resources and Environmental Management, China University of Geosciences, Beijing 100083, China

[e] Oil & Gas Strategy Research Center, Ministry of Land and Resources of the People's Republic of China, Beijing 100034, China

[f] School of Energy Resources, China University of Geosciences, Beijing 100083, China

HIGHLIGHTS

- A novel method to grasp articles' key points and relations from a holistic view.
- An empirical study based on the two-mode affiliation network theory.
- Integrates statistics, text mining, complex networks and visualization.
- Constructed the articles co-keyword networks and keywords co-occurrence networks.
- Defined innovation coefficient of the articles in annual level.

ARTICLE INFO

Article history:
Received 28 June 2015
Received in revised form 3 December 2015
Available online 28 January 2016

Keywords:
Articles co-keyword network
Keywords co-occurrence network
Two-mode affiliation network
Topological features evolution
Innovation coefficient

ABSTRACT

Keeping abreast of trends in the articles and rapidly grasping a body of article's key points and relationship from a holistic perspective is a new challenge in both literature research and text mining. As the important component, keywords can present the core idea of the academic article. Usually, articles on a single theme or area could share one or some same keywords, and we can analyze topological features and evolution of the articles co-keyword networks and keywords co-occurrence networks to realize the in-depth analysis of the articles. This paper seeks to integrate statistics, text mining, complex networks and visualization to analyze all of the academic articles on one given theme, complex network(s). All 5944 "complex networks" articles that were published between 1990 and 2013 and are available on the Web of Science are extracted. Based on the two-mode affiliation network theory, a new frontier of complex networks, we constructed two different networks, one taking the articles as nodes, the co-keyword relationships as edges and the quantity of co-keywords as the weight to construct articles co-keyword network, and another taking the articles' keywords as nodes, the co-occurrence relationships as edges and the quantity of simultaneous co-occurrences as the weight to construct keyword co-occurrence network. An integrated method for analyzing the topological features and

* Corresponding author at: School of Humanities and Economic Management, China University of Geosciences, Beijing 100083, China. Tel.: +8610 82322073; fax: +8610 82321783.
E-mail address: babyproud@126.com (H. Li).

http://dx.doi.org/10.1016/j.physa.2016.01.017

evolution of the articles co-keyword network and keywords co-occurrence networks is proposed, and we also defined a new function to measure the innovation coefficient of the articles in annual level. This paper provides a useful tool and process for successfully achieving in-depth analysis and rapid understanding of the trends and relationships of articles in a holistic perspective.

1. Introduction

Within the recent development and popularization of information and data analysis technology, computing large-scale data-intensive analysis of scientific data is a new trend of data-mining [1], and as one of the main aspects of data-mining, text mining has become a new method of knowledge discovery. Text mining is a useful tool for understanding the basic information provided by one or more texts through structured algorithms. However, it can also be used to determine the relationships among the textual elements and the texts themselves. The results can be used for knowledge discovery and other applications. As an important tool and method for knowledge discovery, text mining has been used in many fields, such as medicine [2], biochemistry [3], business [4], and so on. The objects of analysis in text mining include literatures [5], news [6], network information [7], long texts [8], etc. Various technologies [9–15] and tools [16–21] are used in this field. Such technologies and tools have been enhanced not only to conduct single-text analysis but also to analyze big data and complexity.

Existing literatures indicate that one of the most frequently use of text mining methods is to conduct a literature review, which allows researchers to determine the developing trends in the field. Literature reviews are also fundamental in academic research. There are currently two ways to conduct a literature review. The first is to identify important academic articles by their citation frequency and the impact factors of the journals in which they were published [22]. This method is used to identify recent developments in a field during a short period. However, due to the limited sample, it is difficult to achieve a holistic perspective using this method. The other frequently used method is content analysis [23], a research technique that involves the systematic, objective and quantitative description of a text's content. This method has recently received increased attention. Researchers can use content analysis to find multi-text statistics and clusters by delimiting a research object and establishing a quantification standard. However, it is difficult to maintain the consistency of the quantitative criteria because both the classification and coding rules are based on the knowledge and experience of the researchers, which undermines the objectivity of the analysis. Content analysis is also an inadequate method for mining the complex relationships among the texts.

There is an urgent need to develop a tool for tracking the trends in academic articles and rapidly understanding the key points and inner relationship of a collection of texts from a holistic perspective. Keywords, an important textual element, can provide a concise overview of the important content and key points of a body of articles. Keyword analysis can also expedite text mining [24,25]. Many scholars use tag clouds to analyze unstructured keywords because this method allows the user to highlight the most significant concepts, which facilitates navigation and visualization [26]. However, tag clouds only show the frequency of single words and do not show the relationships of the keywords and the relationships between the articles based on the keywords. Unlike tag clouds, complex network is a young but active method to discover the inner relationship between different entities from real or virtual system. It is well used in different areas, such as economic networks, biological networks, and so on. It can effectively model a network's topological features [27–29], mine its relationships [30], and analyze its evolution [28]. As a new frontier of complex network, multi-mode network has been shown to better represent reality according to its heterogeneous attributes, it has been successfully used in some other area, such as multi-mode societal ecological affiliation network [31–35], fibers transmission [36] and shareholding network of the listed companies [37].

In this paper, we study the patterns of relationships among academic articles on a given theme, complex network, from a holistic perspective by constructing and analyzing annual articles co-keyword equivalent networks (AENs for short) and annual keywords co-occurrence equivalent networks (KENs for short). The process of constructing the two different networks is the same as the one employed to construct equivalent networks [38] using the two-mode affiliation network. The topological features of the two networks in annual level and the evolution as well as the stability of the two networks are analyzed. Then, the innovation coefficient of the networks about the given theme, complex networks, is defined and analyzed.

2. Methods and data

2.1. Methods

2.1.1. Constructing the AENs and KENs

In this paper, affiliation relationships can be found between keywords and the articles in which they appear. Networks constructed according to affiliation relationships are a typical type of two-mode network called a member-network [39]

or hyper-network [40]. The two-mode affiliation network is composed of a set of actors (keywords) and a set of events (articles) [41]. According to Wasserman [42] and Li et al. [38], when there are two nodes, α and β, that have the same relationship with γ, then α and β are equivalent. In the keywords affiliation network, the articles that have the same keyword(s) are equivalent, and the keywords that belongs to a single paper are also equivalent. Therefore, we can construct both the derivative articles co-keyword equivalent networks and derivative keywords co-occurrence equivalent networks based on the equivalent relationships. Then, all of the articles equivalent networks and keywords equivalent networks are added respectively. We assumed that $\mathbf{A}$ is the primitive $N \times M$ matrix of the affiliation relationships between keywords and articles, while N is the set of keywords, and M is the set of articles. Then we can use Formula (1) to get the derivative $M \times M$ matrix ($\mathbf{X}$) of the co-keyword relationship between the articles, and we can also use Formula (2) to get the derivative $N \times N$ matrix ($\mathbf{Y}$) of the co-occurrence relationship between the keywords. Formula (1) and Formula (2) show the typical way to analyze two-mode networks and can be used to determine the number of the actors which the two events are co-containing as well as the number of the events at which the two actors are co-attendant [39].

$$\mathbf{X} = \mathbf{A}' \times \mathbf{A} \tag{1}$$

$$\mathbf{Y} = \mathbf{A} \times \mathbf{A}'. \tag{2}$$

2.1.2. Topological features of the network

There are dozens of indexes for quantitatively analyzing the topological features and evolution of complex networks. This paper primarily analyzes evolution of two networks, AENs and KENs in annual level using four indexes: average degree, average weighted degree, average clustering coefficient and average short path length. Meanwhile, the auto-correlation function is used to calculate the stability of the keywords co-occurrence networks in annual level, and we also defined a function to calculate the innovation coefficient of the keywords co-occurrence networks in annual level.

The node's degree indicates the number of the nodes in contact with it. The rule $d(b_{ij} \geq 1) = 1$ is used to change the weighted networks into un-weighted networks, resulting in formula (3). Then, using formula (4), the average degree of the network is calculated by calculating each node's degree.

$$\begin{cases} d_{ij} = 1 & \text{node } i \text{ and node } j \text{ have co-relationship} \\ d_{ij} = 0 & \text{node } i \text{ and node } j \text{ do not have co-relationship} \end{cases} \tag{3}$$

$$R = \langle R_i \rangle = \left\langle \sum_{j=1}^{n} d_{ij} \right\rangle \quad i \neq j. \tag{4}$$

In the weighted networks, it is typical to measure the strength of each node's connection to other nodes. In the articles co-keyword network and the keywords co-occurrence equivalent networks, it is important to determine the closeness of the relationships between the articles and the strength of the connection between the keywords. It can be measured by considering the number of nodes to which a given node is connected and by considering the nodes' weights. A higher weighted degree indicates better connectivity. The average weight ed degree of node is calculated as follows:

$$S = \langle S_i \rangle = \left\langle \sum_{j=1}^{n} b_{ij} \right\rangle \quad i \neq j \tag{5}$$

where b_{ij} is determined by Formula (1) and Formula (2).

Usually, we use clustering coefficient to measure the connectivity of the neighbors of a given node, and use average clustering coefficient to analyze the closeness and strength of the relationships between the nodes, which can be calculated by Formula (6) [43]

$$C = \langle C_i \rangle_i = \left\langle \frac{2Ei}{P_i(P_i - 1)} \right\rangle \tag{6}$$

where E_i is the number of existing edges between node i's first neighbors, and $(\frac{1}{2}P_i(P_i - 1))$ is the number of potential edges between node i's neighbors. Meanwhile, the average shortest path length of the network is the average value of the least quantity of edges between any of the two nodes in the network, which can imply the connectivity of network. and it can be calculated by Formula (7) [44].

$$D_{ij} = \frac{1}{N(N-1)} \sum_{i,j \in N,\ i \neq j} d_{ij}. \tag{7}$$

To judge the stability of the evolution of the keywords co-occurrence networks, an auto-correlation function (Jaccard-Coefficient) [45,46] was introduced. This function is widely used to determine group dynamics to measure the overlap between two networks from time $t-1$ to time t [42,47,48]. The function is presented as Formula (8).

$$\Re_f(t = K) = \frac{(\mathcal{N}_K \bigcap \mathcal{N}_{K-1})}{(\mathcal{N}_K \bigcup \mathcal{N}_{K-1})} \tag{8}$$

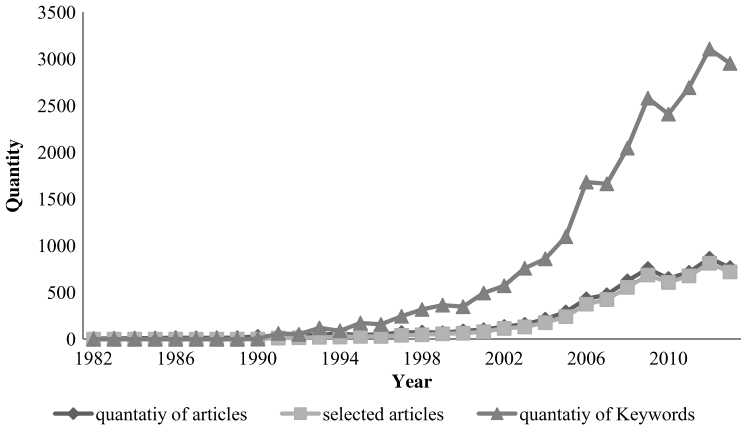

Fig. 1. The quantity of the articles, selected articles and non-repetitive keywords *s*.

where R_f $(t = K)$ is the network stability coefficient, $\mathcal{N}_K$ represents the nodes in the network at time K and $\mathcal{N}_{K-1}$ represents the nodes in the network at time $K-1$. $\mathcal{N}_K \bigcap \mathcal{N}_{K-1}$ is the common nodes at $\mathcal{N}_{K-1}$ and $\mathcal{N}_K$, and $\mathcal{N}_K \bigcup \mathcal{N}_{K-1}$ is the nodes at the union of $\mathcal{N}_{K-1}$ and $\mathcal{N}_K$.

As Formula (8) shows, the stability coefficient can only show the group dynamics between two different time, in order to estimate the trend of the articles year after year considering all the history data, we defined the innovation coefficient of the keywords co-occurrence networks, which determined by the ratio between the emergence of new keywords in present time $(t = K)$ and all the keywords from the first year $(t = 1)$ to present time $(t = K)$.

$$In_c\,(K) = \frac{\left\{\mathcal{N}_K - \left[\mathcal{N}_K \cap \left(\bigcup_{t=1}^{K-1} \mathcal{N}_t\right)\right]\right\}}{\left(\bigcup_{t=1}^{K} \mathcal{N}_t\right)} \tag{9}$$

where $\mathcal{N}_K$ represents the nodes in the network at time K, $\bigcup_{t=1}^{K} \mathcal{N}_t$ represents the nodes at the union of all the networks from $t = 1$ to $t = K$, and $\bigcup_{t=1}^{K-1} \mathcal{N}_t$ represents the nodes at the union of all the networks from $t = 1$ to $t = K-1$. $\{\mathcal{N}_K - (\bigcup_{t=1}^{K-1} \mathcal{N}_t)\}$ represents the new nodes appear in $\mathcal{N}_K$ and never appeared from $t = 1$ to $t = K-1$.

2.2. Data

A set of sample data was used to demonstrate the method in detail. The sample data were extracted from the Web of Science database, one of the largest paper citation indexes in the world. The attributes of the data that were considered salient included the title, the keywords and the remaining keywords, and the time of publication. All of the data, including the four above attributes, which had been indexed using the keyword "complex network (s)" were downloaded. The results yielded 6900 records (articles) published between 1982 and 2013; the data were downloaded on February 12, 2014, and all of the records without keywords were eliminated. Fig. 1 shows the quantity of articles published each year and the quantity of articles which have keywords. The total number of articles being analyzed was 5944, or 86.14% of the original articles. Because 101 of the articles published between 1982 and 1989 did not have keywords, the annual articles co-keyword network and keywords co-occurrence network were constructed using data from articles published from 1990 to 2013. Meanwhile, Fig. 1 also shows the annual quantity of non-repetitive keywords of the articles which have keywords. In order to analyze the inner relations between the quantity of keywords and the quantity of articles which have keywords, we calculated the Pearson correlation coefficient according to Formula (10) [49], we can get the correlation efficient between the non-repetitive keywords and the quantity of articles which have keywords is 0.9972. It means they are strong related with each other. To analyze the data more effectively, duplicate items were deleted, and the titles and keywords were coded. "*T*" indicates the beginning of a title code, and "*W*" indicates the beginning of a keyword code. Five digits follow this initial letter, and each code represents a unique title or keyword.

$$\text{Correlation}(x, y) = \frac{\text{Cov}(x, y)}{\sigma\,(x) * \sigma\,(y)} \tag{10}$$

where $\sigma\,(x)$ is the standard deviation of the annual change of keywords, and $\sigma\,(y)$ is the standard deviation of the annual change of articles which have keywords. $\text{Cov}(x, y)$ is the covariance of annual change of keywords and articles.

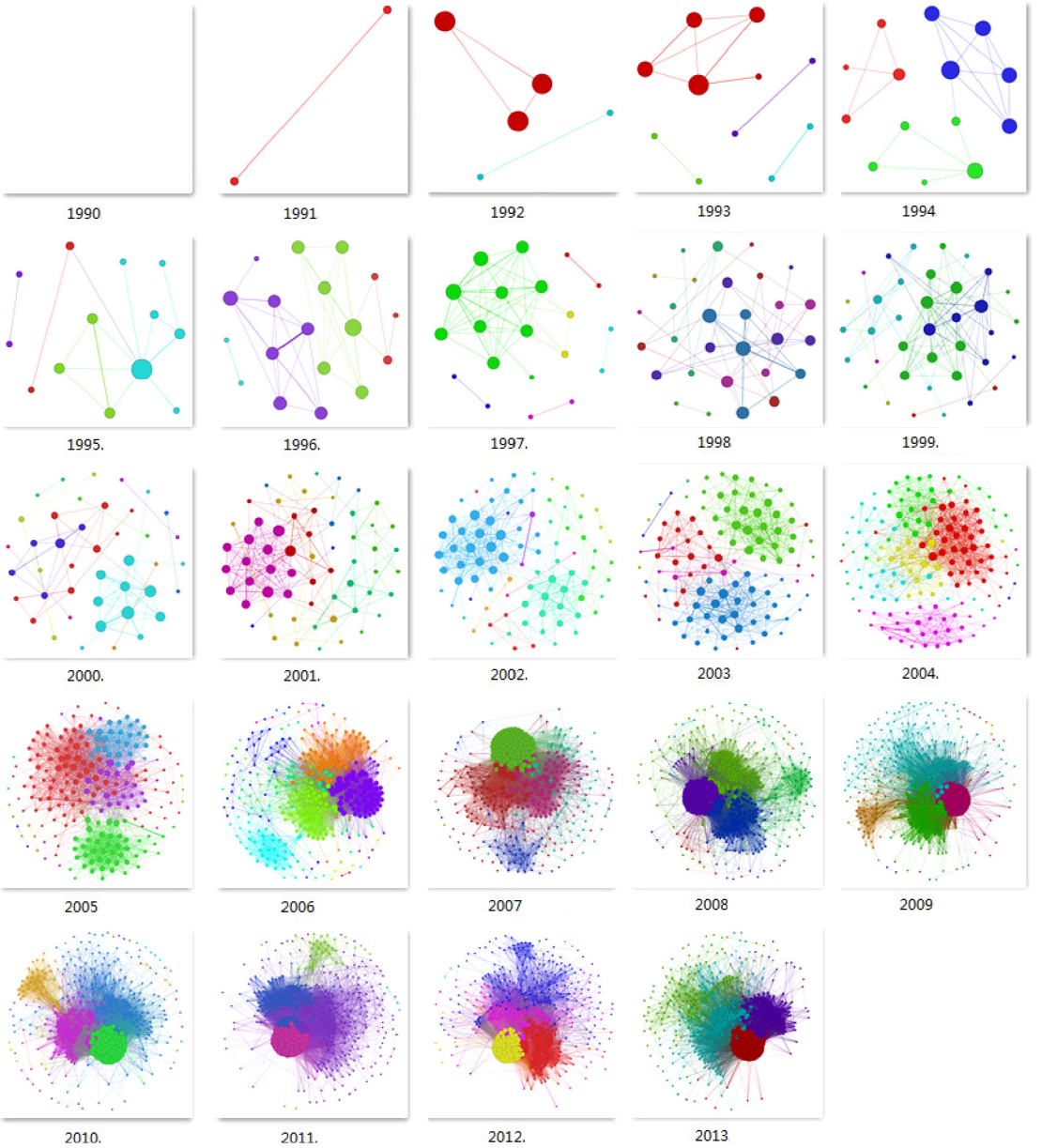

Fig. 2. The visualization results of the AENs (the nodes with the same color represent they belong to the same community get by heuristic method in Gephi). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3. Calculation and analysis

3.1. The visualization of the two different networks

The equivalent networks of each keyword and each paper were constructed based on co-keywords and keywords co-occurrence at the same period (year). The equivalent networks were then superimposed to form the AENs and KENs. Figs. 2–4 present the visualization results and the quantity of nodes and edges of AENs and KENs.

As the Figs. 2 and 4 show, the relationships between the words become increasingly complex over time (the nodes with the same color mean they are more strongly connected). Figs. 3 and 5 indicate that both the co-keywords relationships and the co-occurrence relationships of keywords also increase rapidly, especially after 2004. According to Formula (10), we can get the correlation coefficients between the number of nodes and the number of edges of the AENs and KENs, which are 0.9557 and 0.9979, respectively, which indicates that the number of nodes and edges has highly positive relative

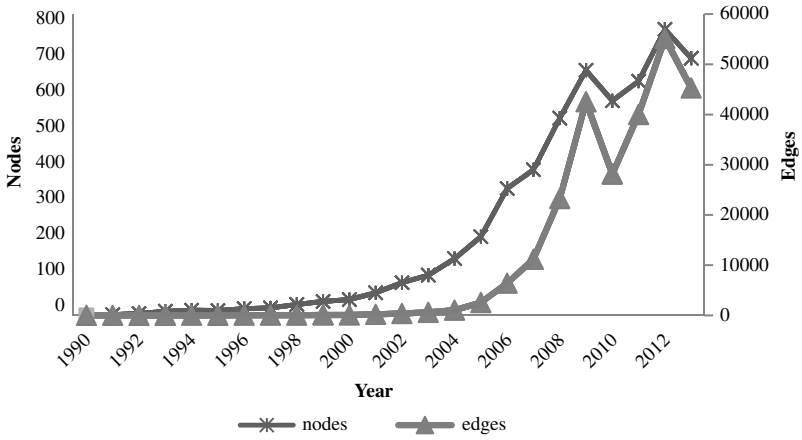

Fig. 3. The quantity of the nodes and edges of the AENs.

relationships. Since both AENs and KENs become more and more complex since 2004, we mainly analyzed the topological features evolution of the two different networks since 2004 in the following sections of this chapter.

3.2. The topological features evolution of AENs

3.2.1. The evolution of degree and weighted degree

The degree of the node in AENs indicates the number of articles which the given paper shares one or more keyword with. Meanwhile, the weighted degree of the node indicates both the number of articles which the given paper shares one or more keyword with and the number of keywords they share. According to Formula (4) and Formula (5), we can get the values of average degree and average weighted-degree of the AENs (see Fig. 6). It indicates that more and more articles about "complex network" share one or more keywords as time goes by. By Formula (10), we got the correlation coefficients between average degree and average weighted-degree of the AENs is 0.9997, which means both the two topological features of AENs evolved consistently.

3.2.2. The evolution of average clustering coefficient and shortest path length

Average clustering coefficient and average shortest path length are two topological features which are popularly used to measure the connectivity of the network [50]. According to Formula (6) and Formula (7), we calculated the two topological features of the AENs (see Fig. 7). According to Fig. 7, the average clustering coefficient increased since 2004, while the average shortest path length decreased gradually in the last decade. It indicates that the connectivity of AENs became better and better as time goes by, and also means that the articles about "complex networks" share more and more similar topics. Meanwhile, according to the result of average shortest path length, we can discover that any articles only need go through around two articles to arrive the other articles in AENs.

3.3. The topological features evolution of KENs

3.3.1. The topological features evolution of nodes and edges

As mentioned above, the weighted degree of the node of KENs represents the closeness between the given keyword and the other keywords with which it has contact. The larger the weighted degree is, the closer the nodes are, indicating that the given keyword's frequency of co-occurrence is relatively high. According to Fig. 4, the keywords' relationships became complex noticeably after 2004, so we focused on the analysis between 2004 and 2013. Database matching was used to retrieve the keywords represented by the codes. Here we use weighted degree to analyze the hot keywords between 2003 and 2013. Table 1 shows the top 20 keywords with the highest weighted degrees during the same period.

According to Table 1, the hot keywords had obvious time distribution differences. Some keywords only appeared in a single year, such as "SELF-ASSEMBLY", "EXPRESSION", "COUPLING DELAYS" et al. Several keywords appeared in phases, such as "CRYSTAL STRUCTURE(S)", "CRITERIA" et al. Other keywords were hot for a long period, such as "COMPLEX NETWORK(S)", "DYNAMICS", "MODEL", "SYSTEM(S)", and so on, which means they have good connectivity diversity.

The weights of the edges were used to measure the strength of the co-occurrence of any two keywords; higher weights indicated that the two keywords co-appeared more frequently in the same articles. The hot topics were identified by analyzing the co-occurring keywords. As Fig. 8 shows, the top 10 pairs of co-occurring keywords varied significantly between 2004 and 2013. It helps us to further analyze the topics of the articles about "complex networks". For example, in 2004, many

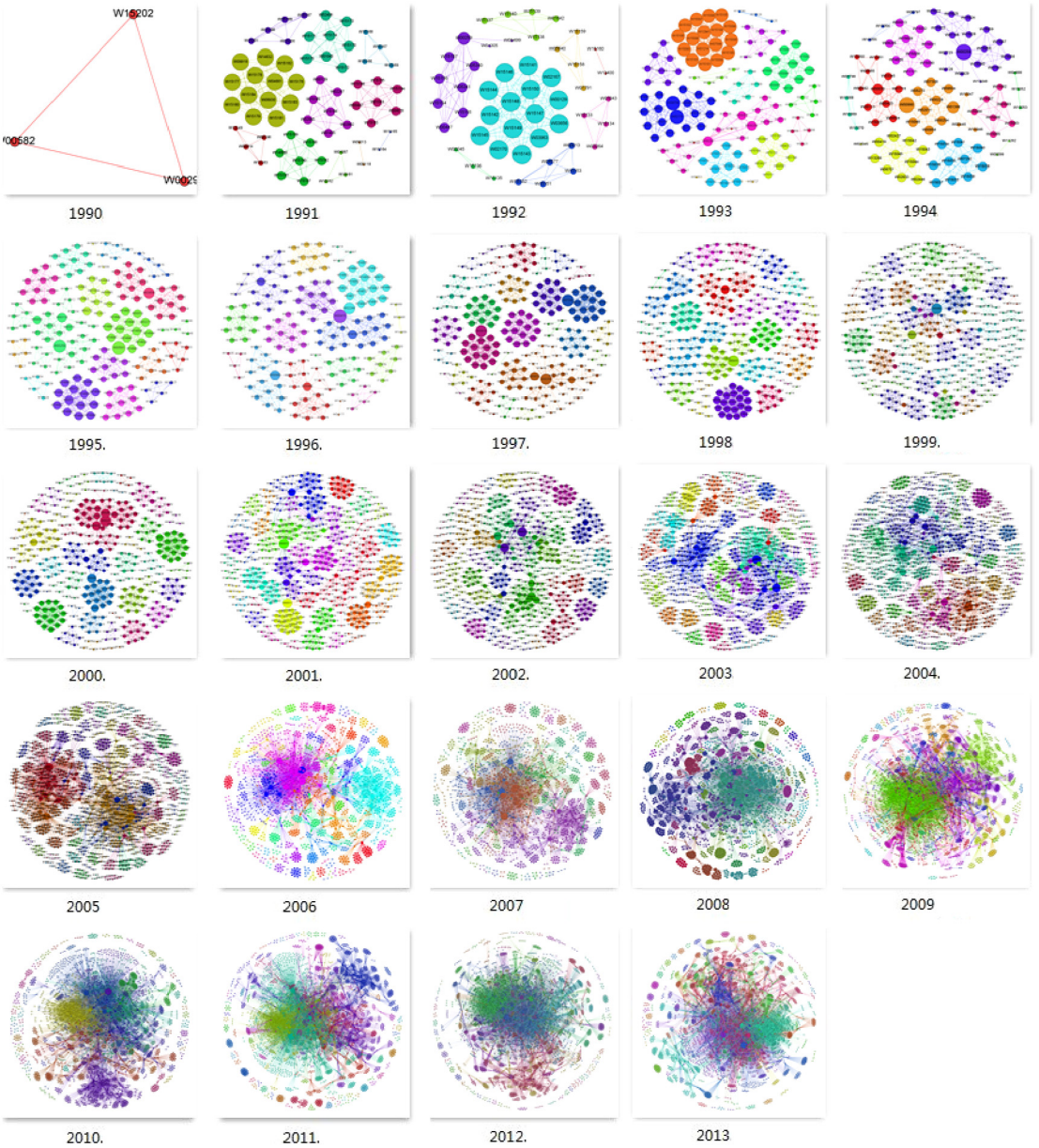

Fig. 4. The visualization results of the KENs (the nodes with the same color represent they belong to the same community get by heuristic method in Gephi). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

articles focused on synchronization and stability as well as small-world networks and internet, but in 2013, the hottest topics were systems and neutral networks as well as complex networks and dynamics.

3.3.2. The stability evolution of KENs

The stability of the KENs indicates that some influence causes the keywords to evolve over time. A larger stability coefficient indicates that the keywords in the two periods compared are more similar, demonstrating the similarity of the authors' interests. However, a small stability coefficient indicates that the articles tend to explore new research areas. Fig. 9 shows the stability coefficients of the keyword co-occurrence equivalent networks in the last decade, which are found using Formula (8). According to Fig. 9, the stability coefficient increases over time, and it can be divided into 3 different periods, the slow increase period (A before 2007), the rapid increasing period (B 2007–2009), the steady period (C after 2009), for example, between 2009 and 2013, the stability coefficient remained relatively stable (between 0.12 and 0.14), indicating that the articles that used the keyword "complex network" had a coincidence rate of 12%–14%. More than 80% of the remaining

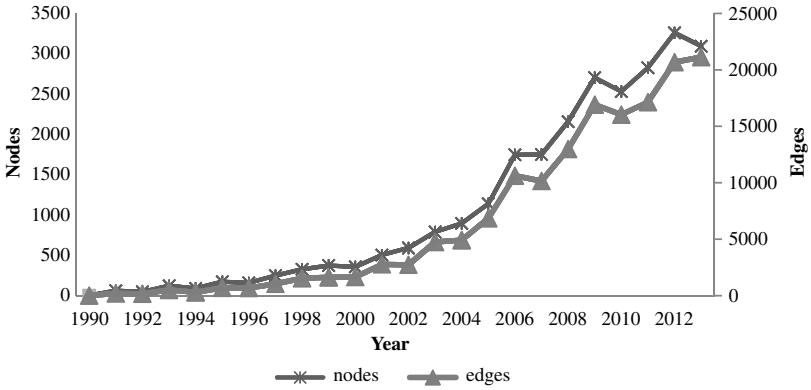

Fig. 5. The quantity of the nodes and edges of the KENs.

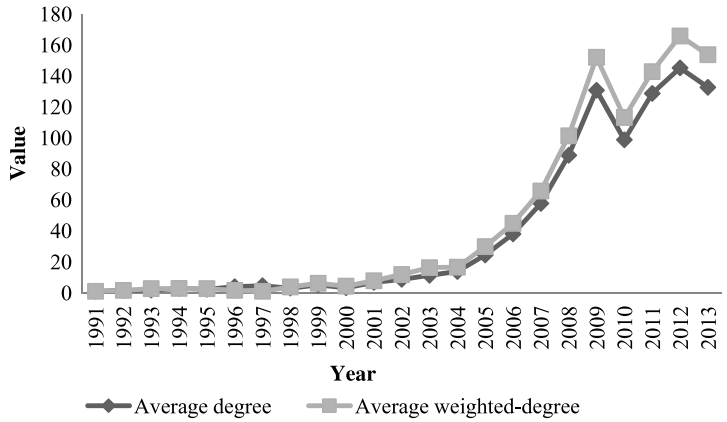

Fig. 6. Average degree and average weighted-degree of the AENs.

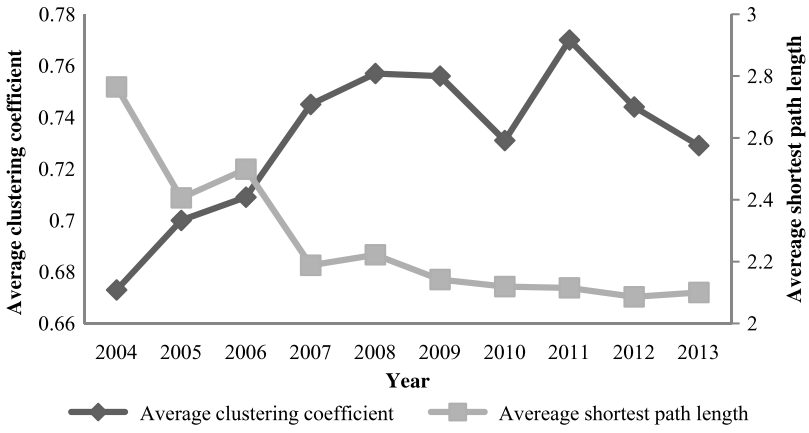

Fig. 7. Average clustering coefficient and average shortest path length of the AENs.

keywords were different. These results indicate that research on complex networks is still experiencing periods of rapid development.

3.3.3. The innovation coefficient evolution of KENs

In order to further analyze the innovation of KENs, as well as the whole trend of keywords in articles about "complex networks", we used Formula (9) to calculate the innovation coefficients of KENs in last 10 years (see Fig. 10) while

Table 1
Annual evolution of the 20 keywords with the highest weighted degrees between 2003 and 2013.

Keywords	2004	2005	2006	2007	2008	2009	2010	2011	2012	2013
COMPLEX NETWORK(S)	T	T	T	T	T	T	T	T	T	T
DYNAMICS	T	T	T	T	T	T	T	T	T	T
MODEL	T	T	T	T	T	T	T	T	T	T
SYSTEM(S)	T	T	T	T	T	T	T	T	T	T
TOPOLOGY		T	T	T	T	T	T	T	T	T
SYNCHRONIZATION	T		T	T	T	T	T	T	T	T
STABILITY	T		T	T	T	T	T	T	T	
SCALE-FREE NETWORKS	T	T	T	T	T	T	T		T	
SMALL-WORLD NETWORKS	T	T	T		T	T		T	T	T
ORGANIZATION	T	T	T		T	T	T	T		T
COMMUNITY STRUCTURE				T	T	T	T	T	T	T
INTERNET	T	T	T	T	T	T				
DYNAMICAL NETWORKS					T	T	T	T	T	T
NEURAL-NETWORKS					T	T	T	T	T	T
GLOBAL SYNCHRONIZATION					T	T	T	T	T	T
CRYSTAL STRUCTURE(S)	T	T	T	T	T					
CRITERIA						T	T	T	T	T
NETWORK(S)	T	T	T	T						
LIGANDS	T	T	T				T			
EVOLUTION		T	T	T		T				
DESIGN			T	T	T			T		
PINNING CONTROL							T	T	T	T
ADAPTIVE SYNCHRONIZATION							T	T	T	T
CHEMISTRY	T	T	T							
IDENTIFICATION	T							T		T
COORDINATION POLYMERS		T		T			T			
COMPLEX SYSTEMS	T		T							
METABOLIC NETWORKS	T			T						
CHAOS					T	T				
MODULARITY							T		T	
EXPONENTIAL SYNCHRONIZATION								T		T
HYDROTHERMAL SYNTHESIS	T									
SMALL-WORLD	T									
HYDROGEN-BOND		T								
MAGNETIC-PROPERTIES		T								
SELF-ASSEMBLY		T								
EXPRESSION			T							
WEB				T						
ESCHERICHIA-COLI				T						
COUPLING DELAYS						T				
ALGORITHM									T	
STABILIZATION									T	
SOCIAL NETWORKS										T
TIME-VARYING DELAYS										T

Note: The singular and plural forms of keywords are combined to one keyword.
The gray cells with "T" are the keywords which appeared in the Top 20 keywords.

considering all the appeared keywords from 1990 to the given year, and we also calculated the percentage of new keywords of each year. According to Fig. 10, we can discover that the innovation coefficient experienced an increase trend before 2005, and then decreased steadily after 2005. The innovation coefficients were between 0.1 and 0.25, which means there were 10%–23% new keywords appeared each year while considering the historical appeared keywords, and the percentage of

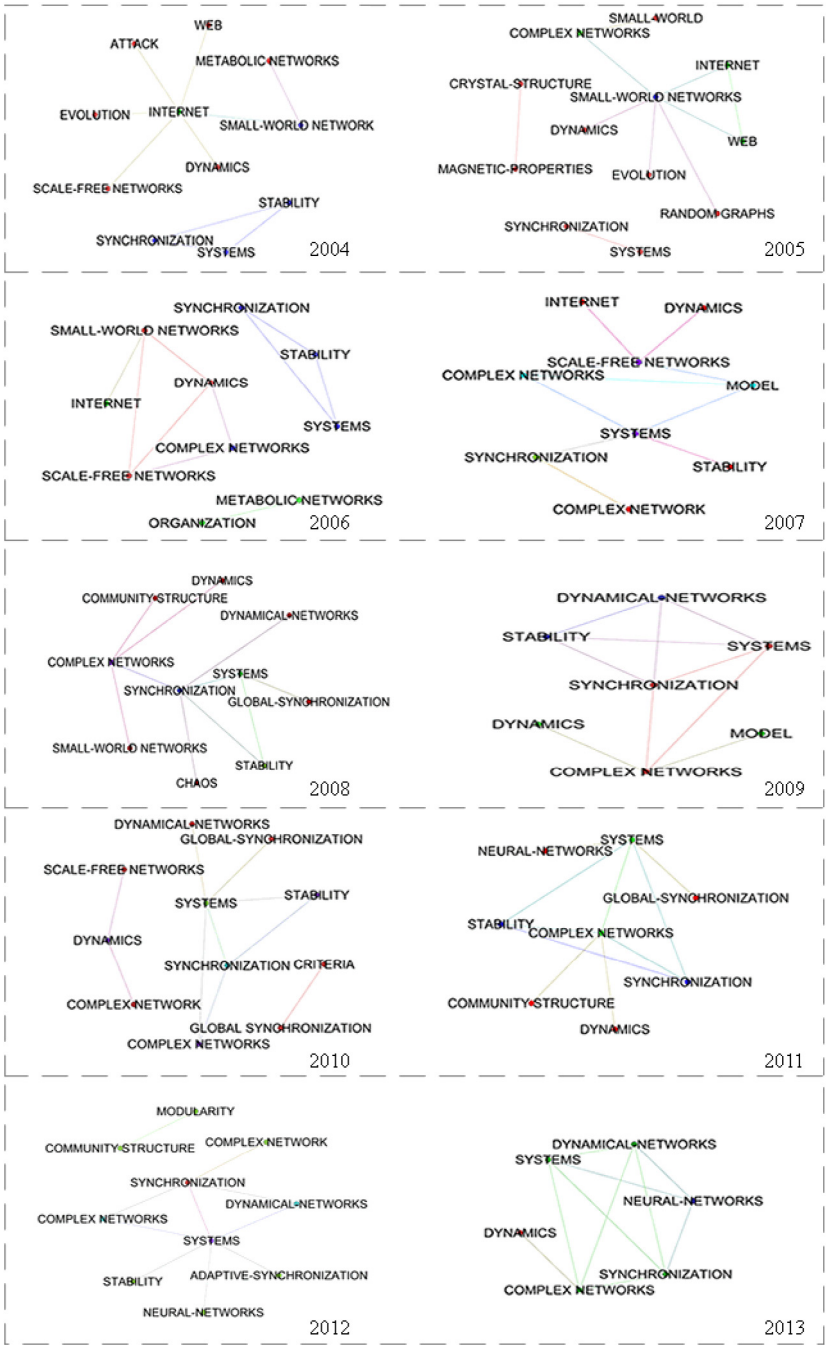

Fig. 8. The 10 edges with the highest weights from 2004 to 2013.

the new keywords each year was around 50%–70%, which was calculated by the number of new keywords divided by the number of nodes in the given year.

4. Discussion and conclusion

In order to gain the evolutionary features of a body of articles and their relations, in this paper, we used 5944 "complex networks" articles that were published between 1990 and 2013 as the sample. Based on the two-mode affiliation network theory, we constructed the AENs by taking the articles as nodes, the co-keyword relationships as edges and the quantity of co-keywords as weights and the KENs by taking the articles' keywords as nodes, the co-occurrence relationships as edges and the quantity of simultaneous co-occurrences as weights. The integrated method demonstrated in the paper visualizes

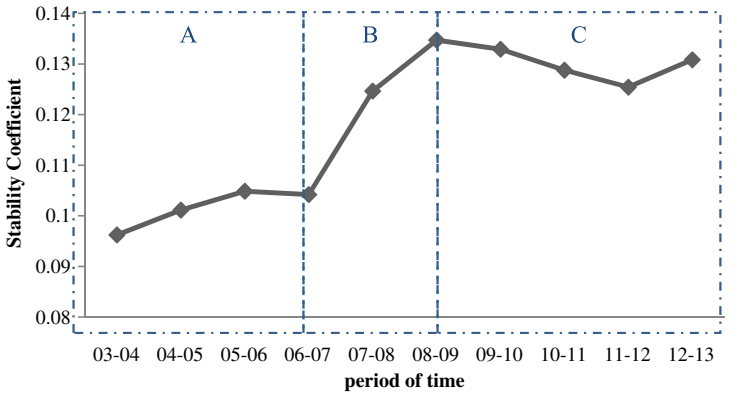

Fig. 9. The stability evolution of the KENs.

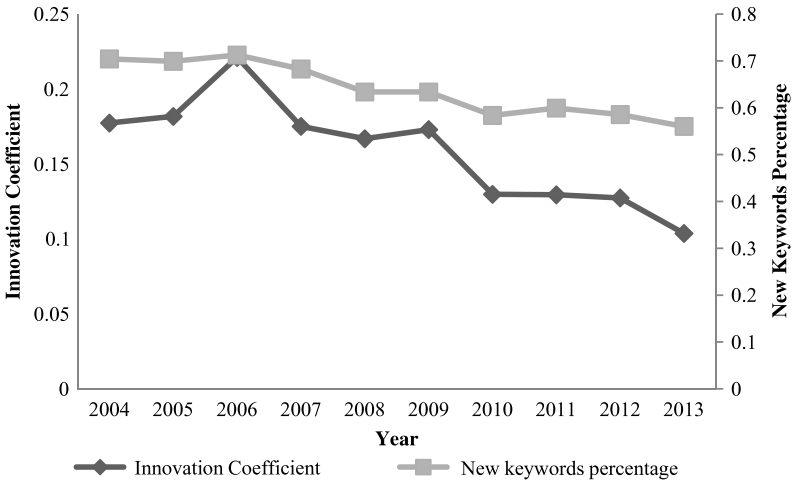

Fig. 10. The innovation coefficient evolution of the KENs.

the evolution of the AENs and KENs and facilitates a rapid understanding of the hot words and hot topics using a quantitative network evaluation index. The principle content and conclusions are presented below:

First, the visualization results of AENs and KENs indicate that both the relations between articles and relations between keywords become more and more complicated and interactively, which means that, in the "Complex network" field, more and more articles focus on the similar sub-topic.

Second, according to the evolution results of degree and weighted degree of AENs, we can conclude that, both the connectivity diversity and the strength of connectivity evolved more and more strongly from 2004 to 2013, which means more and more articles shared the same keywords and talked about the same topics. Meanwhile, according to the evolution of average clustering coefficient and shortest path length of AENs, the connectivity of AENs became better and better as time goes by, and the articles can reach each other by no more than 2 articles on average, which provides further proof that more and more articles shared the similar topics.

Third, according to the hot-keywords and the top pairs of keywords analysis, the hot keywords and the hot topics of the articles about "complex networks" had obvious differences in time distribution.

Forth, according to the result of the stability coefficient, more and more articles in different years share the same keywords. Meanwhile, according to the innovation coefficient we defined, we find that, each year, around 50%–70% of the keywords are new, if considering the historical appeared keywords, the percentage of the new keywords is 10%–23%, so we can conclude that "complex network" research is still in a period of innovation, but the innovation rate decreased steadily from 2006 to 2013.

This paper proposed a method for determining trends in the literature and rapidly understanding the key ideas presented in a body of literature from a holistic perspective by examining the articles co-keyword relationships and keywords co-occurrence relationships. However, some problems still remain unresolved. For example, semantic similarities continue to pose a challenge to keyword analysis, as does keyword accuracy. Although keyword co-occurrence equivalent networks provide a method that decreases researchers' subjectivity and limitations, the authors of the articles may differ in ability. Therefore, different authors may assign different keywords to the same paper, leading to a biased understanding of the

paper. Future research should explore additional ways to extend keyword identification methods, such as performing word segmentation of the titles, abstracts and bodies of articles. Meanwhile, as the derivative one-mode network, both AENs and KENs mentioned in this paper have their own community structure features, so it is necessary to improve the traditional methods in order to divide and analyze the communities more accurately.

Acknowledgments

This research is supported by grants from the National Natural Science Foundation of China (Grant No. 71173199), the China Scholarship Council (File No. 201406400004), the Humanities and Social Sciences Planning Funds Project under the Ministry of Education of the PRC (Grant No. 10YJA630001), and the Fundamental Research Funds for the Central Universities (Grant No. 2-9-2014-104). The authors would like to express their gratitude to the reviewers and Xuan Huang, Xiaoqing Hao, Xiaoliang Jia, Shupei Huang as well as the scholars in PACIS 2014, Chengdu who provided valuable suggestions. Meanwhile, the authors would like to thank AJE—American Journal Expert (www.aje.com) for their professional suggestions about language usage, spelling, and grammar of this paper.

References

[1] L. Wang, J. Tao, R. Ranjan, H. Marten, A. Streit, J. Chen, D. Chen, G-Hadoop: MapReduce across distributed data centers for data-intensive computing, Future Gener. Comput. Syst. 29 (2013) 739–750.
[2] P. Warrer, E.H. Hansen, L. Juhl-Jensen, L. Aagaard, Using textmining techniques in electronic patient records to identify ADRs from medicine use, Br. J. Clin. Pharmacol. 73 (2012) 674–684.
[3] M. Miwa, T. Ohta, R. Rak, A. Rowley, D.B. Kell, S. Pyysalo, S. Ananiadou, A method for integrating and ranking the evidence for biochemical pathways by mining reactions from text, Bioinformatics 29 (2013) i44–i52.
[4] X. Gao, S. Murugesan, B. Lo, Extraction of keyterms by simple text mining for business information retrieval, in: e-Business Engineering 2005 of IEEE International Conference, Beijing, China, 2005, pp. 332–339.
[5] A. Korhonen, I. Silins, L. Sun, U. Stenius, The first step in the development of text mining technology for cancer risk assessment: Identifying and organizing scientific evidence in risk assessment literature, BMC Bioinformatics 10 (2009) 303.
[6] P. Kroha, R. Baeza-Yates, B. Krellner, Text mining of business news for forecasting, in: Database and Expert Systems Applications: 17th International Workshop, Krakow, Poland, 2006, pp. 171–175.
[7] J. Švec, J. Hoidekr, D. Soutner, J. Vavruška, Web text data mining for building large scale language modelling corpus, in: Text, Speech and Dialogue, Springer, Berlin, Heidelberg, 2011, pp. 356–363.
[8] H. Takeuchi, S. Ogino, H. Watanabe, Y. Shirata, Context-based text mining for insights in long documents, in: Practical Aspects of Knowledge Management, Springer, Berlin, Heidelberg, 2008, pp. 123–134.
[9] A.G. Skarmeta, A. Bensaid, N. Tazi, Data mining for text categorization with semisupervised agglomerative hierarchical clustering, Int. J. Intell. Syst. 15 (2000) 633–646.
[10] W. Claster, S. Shanmuganathan, N. Ghotbi, Text mining in radiological data records: An unsupervised neural network approach, in: First Asia International Conference of Modelling & Simulation, 2007, pp. 329–333.
[11] C.H. Lee, C.H. Wu, A self-adaptive clustering scheme with a time-decay function for microblogging text mining, in: Future Information Technology, Springer, Berlin, Heidelberg, 2011, pp. 62–71.
[12] Z. Xu, X. Wei, X. Luo, Y. Liu, L. Mei, C. Hu, L. Chen, Knowle: a semantic link network based system for organizing large scale online news events, Future Gener. Comput. Syst. 43 (2015) 40–50.
[13] Z. Xu, X. Luo, S. Zhang, X. Wei, L. Mei, C. Hu, Mining temporal explicit and implicit semantic relations between entities using web search engines, Future Gener. Comput. Syst. 37 (2014) 468–477.
[14] X. Luo, Z. Xu, J. Yu, X. Chen, Building association link network for semantic link on web resources, IEEE Trans. Autom. Sci. Eng. 8 (2011) 482–494.
[15] Y. Liu, L. Chen, X. Luo, L. Mei, C. Hu, Z. Xu, Semantic link network based model for organizing multimedia big data, IEEE Trans. Emerging Top. Comput. 1 (2014).
[16] H. Lee, G.S. Yi, J.C. Park, E3Miner: a text mining tool for ubiquitin-protein ligases, Nucleic Acids Res. 36 (2008) W416–W422.
[17] R.E. Saunders, S.J. Perkins, CoagMDB: a database analysis of missense mutations within four conserved domains in five vitamin K-dependent coagulation serine proteases using a textmining tool, Hum. Mutat. 29 (2008) 333–344.
[18] F. Yuan, Y. Zhou, CDGMiner: A new tool for the identification of disease genes by text mining and functional similarity analysis, in: Advanced Intelligent Computing Theories and Applications, With Aspects of Artificial Intelligence, Springer, Berlin, Heidelberg, 2008, pp. 982–989.
[19] R. Jelier, M.J. Schuemie, A. Veldhoven, L.C. Dorssers, G. Jenster, J.A. Kors, Anni 2.0: a multipurpose text-mining tool for the life sciences, Genome Biol. 9 (2008) R96.
[20] D.S. DeLuca, E. Beisswanger, J. Wermter, P.A. Horn, U. Hahn, R. Blasczyk, MaHCO: an ontology of the major histocompatibility complex for immunoinformatic applications and text mining, Bioinformatics 25 (2009) 2064–2070.
[21] M. Gerner, F. Sarafraz, C.M. Bergman, G. Nenadic, BioContext: an integrated text mining system for large-scale extraction and contextualization of biomolecular events, Bioinformatics 28 (2012) 2154–2161.
[22] Y.X. Wang, J. Zhang, X.Y. Chu, Y. Liu, F. Li, Z.B. Wang, L.X. Wei, Diagnosis and multi-modality treatment of adult pulmonary plastoma: Analysis of 18 cases and review of literature, Asian Pac. J. Trop. Med. 7 (2014) 164–168.
[23] V.J. Duriau, R.K. Reger, M.D. Pfarrer, A content analysis of the content analysis literature in organization studies: Research themes, data sources, and methodological refinements, Organ. Res. Methods 10 (2007) 5–34.
[24] M. Conway, Mining a corpus of biographical texts using keywords, Lit. Linguist. Comput. 25 (2010) 23–35.
[25] R. Feldman, I. Dagan, H. Hirsh, Mining text using keyword distributions, J. Intell. Inf. Syst. 10 (1998) 281–300.
[26] G. Koutrika, Z.M. Zadeh, H. Garcia-Molina, Data clouds: summarizing keyword search results over structured data, in: Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, Saint-Petersburg, Russia, 2009, pp. 391–402.
[27] H. An, X. Gao, W. Fang, X. Huang, Y. Ding, The role of fluctuating modes of autocorrelation in crude oil prices, Physica A 393 (2014) 382–390.
[28] W. Zhong, H. An, X. Gao, X. Sun, The evolution of communities in the international oil trade network, Physica A 413 (2014) 42–52.
[29] H. Li, H. An, X. Gao, J. Huang, Q. Xu, On the topological properties of the cross-shareholding networks of listed companies in China: Taking shareholders' cross-shareholding relationships into account, Physica A 406 (2014) 80–88.
[30] X.Y. Gao, H.Z. An, W.Q. Zhong, Features of the correlation structure of price indices, PLoS One 8 (2013) e61091.
[31] X.H. Xia, Y.B. Chen, J.S. Li, H. Tasawar, A. Alsaedi, G.Q. Chen, Energy regulation in China: Objective selection, potential assessment and responsibility sharing by partial frontier analysis, Energy Policy 66 (2014) 292–302.
[32] Z.M. Chen, G.Q. Chen, Demand-driven energy requirement of world economy 2007: A multi-region input–output network simulation, Commun. Nonlinear Sci. Numer. Simul. 18 (2013) 1757–1774.

[33] G.Q. Chen, S. Guo, L. Shao, J.S. Li, Z.M. Chen, Three-scale input–output modeling for urban economy: Carbon emission by Beijing 2007, Commun. Nonlinear Sci. Numer. Simul. 18 (2013) 2493–2506.
[34] Z.M. Chen, G.Q. Chen, X.H. Xia, S.Y. Xu, Global network of embodied water flow by systems input–output simulation, Front. Earth Sci. 6 (2012) 331–344.
[35] X.H. Xia, G.T. Huang, G.Q. Chen, B. Zhang, Z.M. Chen, Q. Yang, Energy security, efficiency and carbon emission of Chinese industry, Energy Policy 39 (2011) 3520–3528.
[36] Z.H. Khan, I.Y.H. Gu, A.G. Backhouse, Robust visual object tracking using multi-mode anisotropic mean shift and particle filters, IEEE Trans. Circuits Syst. Video Technol. 21 (2011) 74–87.
[37] H. Li, W. Fang, H. An, L. Yan, The shareholding similarity of the shareholders of the worldwide listed energy companies based on a two-mode primitive network and a one-mode derivative holding-based network, Physica A 415 (2014) 525–532.
[38] H.J. Li, H.Z. An, J.C. Huang, X.Y. Gao, Y.L. Shi, Correlation of the holding behaviour of the holding-based network of Chinese fund management companies based on the node topological characteristics, Acta Phys. Sinica 63 (2014) 48901. 048901.
[39] R.L. Breiger, The duality of persons and groups, Soc. Forces 53 (1974) 181–190.
[40] J.M. McPherson, Hypernetwork sampling: Duality and differentiation among voluntary organizations, Soc. Networks 3 (1982) 225–249.
[41] H. Li, W. Fang, H. An, L.L. Yan, The shareholding similarity of the shareholders of the worldwide listed energy companies based on a two-mode primitive network and a one-mode derivative holding-based network, Physica A 415 (2014) 525–532.
[42] S. Wasserman, Social Network Analysis: Methods and Applications, Vol. 8, Cambridge university press, 1994.
[43] H. Ebel, L.I. Mielsch, S. Bornholdt, Scale-free topology of e-mail networks, Phys. Rev. E 66 (2002) 035103.
[44] U. Brandes, A faster algorithm for betweenness centrality, J. Math. Sociol. 25 (2001) 163–177.
[45] G. Palla, A.L. Barabási, T. Vicsek, Quantifying social group evolution, Nature 446 (2007) 664–667.
[46] P. Jaccard, The distribution of the flora in the alpine zone, New Phytol. 11 (1912) 37–50.
[47] C.A. Hidalgo, C. Rodriguez-Sickert, The dynamics of a mobile phone network, Physica A 387 (2008) 3017–3024.
[48] P. Nonacs, Measuring and using skew in the study of social behavior and evolution, Am. Nat. 156 (2000) 577–589.
[49] K. Pearson, Contributions to the mathematical theory of evolution. II. Skew variation in homogeneous material, Philos. Trans. R. Soc. Lond. (1895) 343–414.
[50] H. Li, W. Fang, H. An, X. Huang, Words analysis of online Chinese news headlines about trending events: A complex network perspective, PLoS One 10 (2015) e0122174.