Topic Discovery in Scientific Literature

Yujian Huang, Qiang Liu, Jia Liu, and Yanmei Hu(✉)

Chengdu University of Technology, Chengdu, China
{huangyj,huyanmei}@cdut.edu.cn, liujia0833@stu.cdut.edu.cn

Abstract. With the progress of society and science, various fields have achieved unprecedented development. Various research directions and problems have blossomed, and a huge amount of scientific literature has emerged. Scientific literature contains rich "knowledge", e.g., research hotspots and topics. If those "knowledge" can be obtained from scientific literature, it would be of great practical significance to both government and researchers. The existing methods generally obtain "knowledge" by analyzing the semantics of scientific literature, which is complex and time-consuming. In this paper, we aim to explore new methods of research hot word extraction and research topic discovery from the perspective of network. Firstly, the word network is constructed based on the text of scientific literature. Next, a research hot word extraction method based on node centrality and a structural topic discovery method are proposed on the word network. Then, the consistency between structural topics and semantic topics is explored. Finally, the proposed methods are experimentally verified on a real dataset. The experimental results show that the proposed centrality based hot word extraction method can effectively extract research hot words, and the topics obtained by the structural topic discovery method are consistent with the semantic topics in some cases, providing a new way to textual knowledge discovery.

Keywords: topic discovery · network structure · centrality · clustering · community discovery

1 Introduction

Scientific literature is the main manifestation of scientific research carried out by scientific and technical workers, and it condenses the highest wisdom of human beings, gathers the concerns of various research fields and even the whole human society, and contains the intricate relationships between research problems and key technologies. Therefore, by observing and analyzing scientific literature, we can understand the concerns and key technologies of different research fields and capture the correlation between them, and even find interesting patterns which would give us a deeper insight on the development of science and technology. For instance, mining research hotspots and topics hidden in scientific literature is significantly important to researchers and governments, since research hotspots and topics can tell what the concerns are in different fields. In recent years, the scientific literature has shown a rapid growth and a large number of scientific articles have been appearing, since many countries pay more and more attention to scientific research. Taking China as an example, it has ushered in the peak period of rapid

© The Author(s), under exclusive license to Springer Nature Singapore Pte Ltd. 2023
Y. Sun et al. (Eds.): ChineseCSCW 2022, CCIS 1681, pp. 481–491, 2023.
https://doi.org/10.1007/978-981-99-2356-4_38

growth of scientific articles since the reform and open policy was executed. According to reports from China Science and Technology Network, the total number of scientific papers indexed by SCI, SSCI and the Humanities and Arts Citation Index (A&HCI) from China exceeded 10,000 in 1995; the number of scientific articles from China was nearly 140,000 in 2010; and the annual output of scientific articles is nearly 290,000 in 2015. Obviously, to observe and analyze "knowledge" from such a large amount of scientific literature, it is extremely unrealistic to only rely on manual labour.

Fortunately, data mining has been developed widely and the related technologies are more and more mature, making it possible to automatically accomplish a lot of tasks using machines. Data mining refers to the process of extracting hidden rules and valuable information from a large amount of data by algorithms, involving mathematical statistics, machine learning, pattern recognition, etc. A natural way to apply data mining to exploring "knowledge" from scientific literature is performing text mining on a collection of articles, since each article in the scientific literature is essentially a document. For example, we can use clustering to categorize articles into different groups, and many methods, e.g., traditional clustering algorithms such as K-means [1] and its variants, and ontology-based clustering [2], are available to text clustering. We can also use topic models such as Probabilistic Latent Semantic Analysis (PLSA) [3], Latent Dirichlet Allocation (LDA) [4], and their variants to detect research topics. We can also simply count word and take the words with high frequency as research hot words. However, regarding to the traditional clustering methods, each article (or the used text of each article) is required to be represented as numerical vector using techniques such as coding or Doc2vec [5]. The former one must use an element to represent each word appearing in the corpus, i.e., the length of the numerical vector is equal to the number of words, consuming a huge amount of memory and computation time; the later one is developed based on Word2vec and is much more complicated, and it also requires a corpus large enough to obtain good representation. Regarding to topic models, they can automatically extract topics as well as the belongingness of each article to each topic based on text information, but the quality of the model is not guaranteed and the major problem is that the topic result is not visual and may not reflect the difference between topics. In addition, word frequency indicates the number of times that a word appears in the used articles and reflects word's popularity, but it does not consider the interaction and similarity between words.

In this paper, we mine "knowledge", specifically hot words and topics, in scientific literature from the perspective of network, and study the consistency between structural topics and semantic topics for the first time, with the aim of exploring a more visual and effective method to mine the main research contents and concerns in scientific literature. First, a word network is constructed to represent the collection of scientific literature. Second, a method of hot word extraction based on node centrality is proposed to identify research hot words from the word network. Then, a structural topic discovery method is proposed to detect research topics according to the topological structure of the word network. The consistency between structural topics and semantic topics is also explored. Finally, experiments are conducted on a collection of scientific literature in the field of computer science to test the proposed methods. Experimental results show that the research hot words can be effectively extracted by the proposed hot word

extraction method, and the obtained structural topics can be consistent with semantic topics. This result implies that it is feasible to discover the main contents and concerns in scientific literature from the perspective of network, providing a new way to "knowledge" discovery in text.

The organization of rest part is as follows. The most related work is described in Sect. 2. Section 3 presents the hot word extraction method based on node centrality. Section 4 presents the structural topic discovery method and analyzes the consistency between structural topics and semantic topics. Experiments are presented in Sect. 5, and Sect. 6 concludes the work.

2 Related Work

To detect research hot words in scientific literature, one can simply count each word appears in articles. However, for an article not each word is informative to the main content. It is thus more realistic to only count the keywords since they condense the most concerned content of article. There are commonly keywords explicitly provided for an article, but one may perform keyword extraction to obtain more objective and suitable keywords. As a fundamental problem in text mining, keyword extraction has been researched widely and there are many ways to extract keywords. For example, one can train a machine learning model on a set of documents where the keywords are known and then used the resulting model to obtain keywords for documents where the keywords are not known [6]. One can also apply statistical methods to obtain keywords. The n-gram statistics [7], word frequency, TF-IDF [8], word co-occurrence, and PAT tree [9] can all be used as statistics of words. Particularly, Biswas et al. proposed KECNW, which is based on node edge rank centrality with node weight depending on various parameters, to extract keywords [10].

To detect research topics in scientific literature, one can use topic models. LDA [11] is a classical model for topics mining in a set of documents. It applies statistics to obtain the topics and the distribution of each document on the topics. LDA can efficiently infer topics, but the number of topics is artificially preset. The HDP model overcomes this limitation by automatically determining the number of topics, but the number of hidden parameters in HDP increases with the data size [12]. The related topic model (CTM) represents another extension of LDA, and it uses a logistic normal distribution to model the variability in the topic proportions of each document to discover related topics [13]. Linstead et al. first used LDA to extract topics in source code and visualize software similarity [14]. Zhao et al. proposed a personalized topic recommendation method based on LDA, called hashtag-LDA, to discover latent topics in microblog [15]. Yin et al. propose a topic model named as LGTA, which is a combination of topic modeling and geographic clustering, to detect topics from geographic information and GPS related documents [16]. Link-LDA is extended from LDA to discover latent topics in a collection of articles by combining citation structures and textual information [17]. Some researchers also integrate author information into LDA, PLSA or HDP to solve the problem of mining author-topic distribution [18]. Cuietal utilizes TextFlow to show the split and fusion of themes [19]. Liu developed a method for mapping technological evolutionary paths using a novel nonparametric topic model named as CIHDP, which

adds citation information to the topic model to determine the number of topics for better dynamic topic detection and track scientific literature [20]. BALILI et al. proposed the TermBall framework, which can simulate the knowledge structure of research topics and track or predict the evolution of research topics [21]. TermBall represents research topics as communities of keywords in a dynamic co-occurrence network.

3 Research Hot Word Extraction Based on Node Centrality

Two problems need to be solved to extract research hot words from the perspective of network. One is how to construct a network based on text data of scientific literature. The other one is how to use the structure of word network to determine hot words. To solve these two problems, we propose a research hot word extraction method based on node centrality. The method consists of two steps: 1) construct a word network according to the adjacent positions of words; 2) apply a centrality metric to calculate each node's centrality value in the word network, and then select the words corresponding to nodes with high centrality value as hot words. Next, we will describe the two steps in detail.

3.1 The Construction of Word Network

For a collection of scientific literature, we first extract each article's abstract and concatenate all the abstracts to form a text. Here we only consider abstract for each article because abstract condenses the research content of a article. Then, we preprocess the text by removing meaningless words and irrelevant words. Meaningless words refer to words without specific meanings such as conjunctions, prepositions and modal verbs. Irrelevant words refer to unimportant words. For a word, its importance is measured as.

$$R(id) = atf_{id} * log(\frac{N}{n_{id}}) \tag{1}$$

where atf_{id} is the times of word id (each word is assigned a unique id) appearing in the text, N is the number of articles, n_{id} is the number of articles where word id appears. After the importance of each word is evaluated by Eq. 1, the words with small importance are considered as unimportant nodes and are eliminated. Actually, we find that the distribution of word importance follows a long tail distribution, i.e., a lot of words have small importance while a few words have extremely large importance.

After the preprocessing above, we construct the word network as follows: each word in the text is represented as a node, and if two words are adjacent to each other in the same sentence, a directed edge is established between the corresponding two nodes. For example, given a text with only one sentence "w1 w2 w1 w3." where w1, w2, and w3 are three different words, the word network constructed from this text contains 3 nodes with each one representing one word; and because w1 and w2 are adjacent to each other in this sentence, there is a directed edge from w1 to w2. Similarly, there are directed edges from w2 to w1 and from w1 to w3.

3.2 Node Centrality Calculation and Hot Words Selection

The centrality of a node is used to evaluate the importance or influence of the node, generally according to the network structure [22]. Thus, we can utilize the centrality of nodes in word network to extract the hot words in scientific literature. There are many centrality metrics in the literature, e.g., degree centrality, betweenness centrality [23, 24], closeness centrality [25], PageRank centrality [26], eigenvector centrality [27]. Some metrics based on local structure such as local clustering coefficient [28] and neighborhood conductance [29] can be used as local centrality. Among these centralities, which one is more suitable for our case? First, hot words should appear frequently; second, hot words should span multiple domains. Moreover, we find that the word network is very dense, and most nodes have relatively high degree and all nodes have a degree more than 50 (see experimental part for details). It means that most nodes meet the first condition. For a node v_i, its betweenness centrality is calculated as:

$$(v_i) = \sum_{s \neq t \neq v_i} \frac{\sigma_{st(v_i)}}{\sigma_{st}} \tag{2}$$

where σ_{st} is the number of shortest paths from node s to node t, and $\sigma_{st(v_i)}$ is the number of shortest paths through v_i. It can be seen from Eq. 2 that if a node frequently appears on the shortest paths between other nodes, its betweenness centrality is high, implying that nodes with high betweenness centrality are important bridges connecting other nodes. This meets the second condition. Therefore, we calculate the betweenness centrality of each node in the word network, and then select as hot words the words corresponding to nodes with high centrality value. Specifically, nodes are sorted in descending order of centrality value, and the top k nodes are selected to obtain hot words.

4 Structural Topic Discovery and Consistence Analysis

4.1 Structural Topic Discovery

The word network reflects the contextual relationship between words, and the nodes that are densely connected in the word network are context-dependent on each other. We call the context-dependent words as a structural topic, since they are grouped together through the topology of word network. Discovering structural topics is essentially the task of community detection, since each structural topic corresponds to a group of densely connected nodes in the word network. Thus, we can perform community detection on the word network to complete the discovery of structural topics.

Community detection is to find out the subsets of densely connected nodes and take each of these subsets as a community [30], and is a fundamental problem in network science. There have been a large number of community detection algorithms proposed in the literature [31], we can choose one of them to perform on the word network, and take each obtained community as a structural topic.

4.2 Consistence Analysis

Further, it is interesting to explore whether the structural topic is consistent with the semantic topic obtained through semantics (e.g., the topic obtained by LDA). If the structural topic is consistent with the semantic topic, then we can discover topics through word network, without using more complex methods such as topic models. Then, how to analyze the consistence between structural topics and sematic topics? After obtaining the structural topics as described in the previous subsection, we follow three steps to fulfil this task: 1) cluster words into different groups to obtain sematic topics; 2) analyze the connectivity within each semantic topic; 3) analyze the distribution of each cohesive sematic topic on structural topics.

Sematic Topics. To obtain sematic topics, we first apply the technique of Word2vec to represent each word as a vector, then use K-means to cluster the word vectors into different groups. Each group is taken as a sematic topic.

Connectivity within a Semantic Topic. For a semantic topic ST_i, we randomly choose a word in it, and take the node corresponding to the chosen word as starting node to perform the breadth-first search algorithm on the word network, with the constrain of ST_i. In particular, for a node encountered by the breadth-first search algorithm, if the corresponding word belongs to ST_i, then the breadth-first search will continue on this node; otherwise, the breadth-first search is truncated at this node. If the corresponding node of each word in ST_i has been visited after the breadth-first search terminates, ST_i is cohesive; otherwise, it is not.

Distribution of a Cohesive Semantic Topic on Structural Topics. For a cohesive semantic topic ST_i, , find the structural topics that overlap with it (two topics are overlapped if they have at least one common word), and evaluate the degree of overlap between ST_i and each of these structural topics.

5 Experimental Results and Analysis

In this section, the proposed methods are tested, and the consistency between structural topics and semantic topics is analyzed. The used dataset is obtained by crawling articles published in ACM and IEEE (only the articles related to computer science are considered in this publisher) from 2015 to 2019, and a total of 11,592 articles are contained.

Word Network. Following the method described in Sect. 3.1, we construct the word network corresponding to the dataset, and its structural information is as follows: the word network contains 849 nodes and more than 100,000 edges, and the average degree is 135.11, the average length of the shortest paths is 1.84, the density is 0.159, and the average clustering coefficient is 0.333, implying that the word network is relatively dense and nodes can easily reach each other. In addition, the degree distribution of the word network is shown in Fig. 1. It can be seen that this distribution is quite different from the well-known power-law distribution, and even the minimum degree is larger than 50.

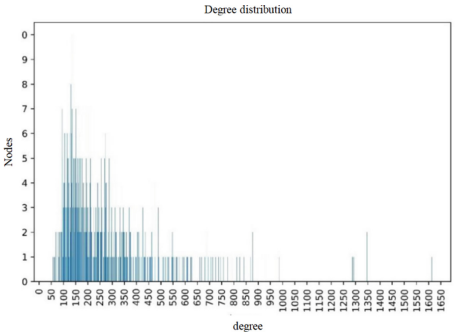

Fig. 1. Degree distribution of word network

Hot Words. The betweenness centrality distribution is shown in the left figure in Fig. 2. It can be seen that the range of the centrality values is [12.8, 30654.5] which is a relatively large span; this distribution follows a power-law distribution, which implies that only a few nodes have a very high centrality value and most nodes have a very low centrality value. We empirically take the top 30 words as hot words and show them separately in the right figure of Fig. 2. It can be seen that the hot words are common technical terms such as model, algorithm and graph, and there are also some words related to research topics such as recommendation, classification and cluster.

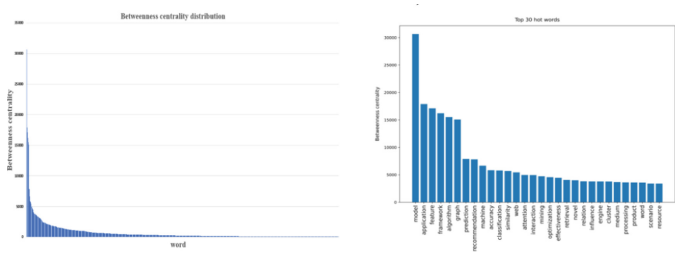

Fig. 2. Betweenness centrality distribution (left) and the top 30 hot words (right).

In addition, we analyze the neighborhood of hot words. Specifically, for each hot word, we get the neighbor nodes of its corresponding node, and sort these neighbor nodes in descending order of centrality value. We find that all the hot words have very similar neighbors with high centrality, which are further very similar to the hot word set. Taking "graph" as an example, the first 30 neighbor nodes are exactly the same to the top 30 hot words (the ones shown in the right figure in Fig. 2). This implies that all the hot words are highly connected. Further, it seems that the word network is composed of highly connected hot words and marginal words closely surrounding the hot words.

Structural Topics and Consistence with Semantic Topics. To obtain the structural topics on the word network, we apply a very popular algorithm of community detection,

named as Louvain [32], to perform community discovery on the word network, and each community is taken as a structural topic. In order to evaluate the quality of the obtained structure topics from network structure (i.e., the quality of the discovered communities), we apply two metrics: conductance and density. The conductance of a community is the proportion of out-going edges to the total edges induced by the nodes in this community, and density is the ratio of edges in the community to the ones in the complete graph containing the same nodes. Lower conductance indicates better community while higher density indicates better one. The details of the two metrics are referred to [33]. The left figure in Fig. 3 shows the evaluation result (isolated nodes are not shown); each node represents a community and the node size is proportional to the community size, the value on the left of "-" is conductance and the one on the right is density. It can be seen that the connections within communities are relatively dense (high density), but there are also many connections between communities (high conductance). It can be inferred that the boundary between communities is blurred, which is reasonable in dense networks.

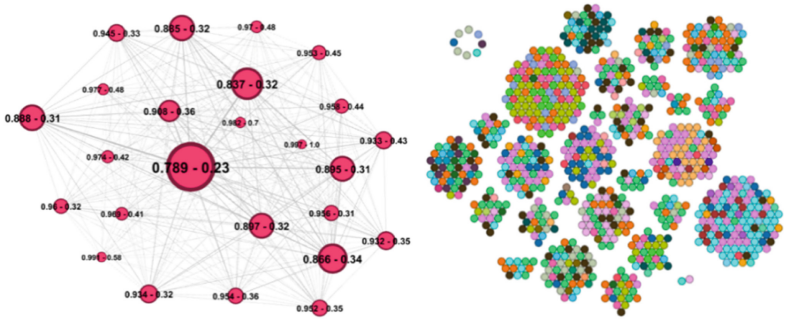

Fig. 3. The quality of the structural topics from the perspective of network structure (left) and the obtained structural and semantic topics and their distribution (right)

To obtain the sematic topics, we apply Word2vec to convert each word into a vector of 100 dimensions, and then use K-means to cluster the word vectors into different groups. Consequently, 35 semantic topics are obtained. After checking the connectivity of semantic topics, we find that the semantic topics to which the hot words belong are all cohesive. This means that the hot words have good connectivity in the word network, which is consistent with the definition of betweenness centrality. Besides, there are three semantic topics that are not cohesive, but they are also semantically unimportant. To analyze the consistence between structural topics and semantic topics, we compare structural topics and semantic topics from a macro perspective, and the results are shown in the right figure of Fig. 3, where each separated part indicates a structural topic, and each color indicates a semantic topic. It can be seen that there are three types of semantic topics that are relatively tight in structure, and most words of them are within the same structural topic. The first type is related to proper nouns (such as GPU, disk, file, etc.), including three semantic topics. Most words in them are distributed within one structural topic (the largest structural topic in the upper left corner), and they are the main members of that structural topic. The second type mainly involves business and market media (such

as amazon, commerce, sale, profit, market, twitter, etc.), including three semantic topics. Most words in them are also within the same structural topic (the largest structural topic in the lower right corner). The third type also includes three sematic topics, mainly involving expressions related to daily life (such as traffic, taxi, bus, GPS, home, car, vehicle, weather, etc.). To these three types of sematic topics, each of them concentrates on one structural topic, indicating that these sematic topics are similar to structural topics. Furthermore, the words belonging to semantic topics related to graph theory are basically distributed in the leftmost structural topic, but this structural topic also contains other words such as bridge, connectivity, neighbor, and motif, which are all related to graph theory. This indicates that in this case the structural topic is better than the semantic topic. In addition, there are eight isolated nodes (upper left corner) corresponding to words of model, framework, application, machine, feature, recommendation, graph and algorithm, respectively. There are also 4 semantic topics scattered in multiple structural topics, which are pink-purple, green, orange, and brown, and we find that the words of these sematic topics are unimportant in terms of centrality. From the discussion above, it can be inferred that in some cases the structural topics and the semantic topics are consistent.

6 Conclusion

From the perspective of network, this paper explores the methods of extracting research hot words and discovering research topics in scientific literature. Specifically, a word network is constructed to represent the contextual relationship of words in abstracts of articles. Based on the word network, we propose to extract hot words by node centrality and discover structural topics by community detection. Moreover, we analyze the consistency between structural topics and sematic topics. Experimental results on a collection of articles show that the proposed hot word extraction method can effectively extract research hot words, and the structural topics are consistent with semantic topics in some cases. This provides a potential way to mine research hot words and topics. However, this work is preliminary and requires more study in the future. For example, testing the methods on large datasets to obtain pervasive conclusion, trying or designing more suitable methods to discover structural topics, and designing suitable metrics to improve consistence analysis.

References

1. Yildirim, M.E., Kaya, M., Ince, L.F.: A case study: unsupervised approach for tourist profile analysis by k-means clustering in turkey. Internet Comput. Serv. **23**(1), 11–17 (2022)
2. Dou, D.J., Wang, H., Liu, H.S.: Semantic data mining: a survey of ontology-based approaches. In: IEEE ICSC (2015) 978–1–4799–7935–6
3. Bassiou, N.K., Kotropoulos, C.L.: online plsa: batch updating techniques including out-of-vocabulary words. IEEE Trans. Netw. Learn. Syst. **25**(11), 1953–1966 (2014)
4. Li, X., Ouyang, J., Zhou, X., Lu, Y., Liu, Y.: Supervised labeled latent Dirichlet allocation for document categorization. Appl. Intell. **42**(3), 581–593 (2014). https://doi.org/10.1007/s10489-014-0595-0

5. Hernández-Castañeda, Á., García-Hernández, R.A., Ledeneva, Y., Millán-Herná-Ndez, C.E.: Extractive automatic text summarization based on lexical-semantic keywords. IEEE Access **8**, 49896–49907 (2020)
6. Beliga, S., Meštrović, A., Martinčić-Ipšić, S.: Selectivity-based keyword extraction method. Int. J. on Semantic Web Inf. Syst. **12**(3), 1–26 (2016)
7. Tripathy, A., Agrawal, A., Rath, S.K.: Classification of sentiment reviews using n-gram machine learning approach. Expert Syst. Appl. **57**, 117–126 (2016)
8. Zhang, Y.T., Gong, L., Wang, Y.C.: An improved TF-IDF approach for text classlncatlon. Zhejiang Univ. Sci. **6**(1), 49–55 (2005)
9. Kang, D.-K., Sohn, K.: Learning decision trees with taxonomy of propositionalized attributes. Pattern Recogn. **42**(1), 84–92 (2009)
10. Biswas, S.K., Bordoloi, M., Shreya, J.: A graph based keyword extraction model using collective node weight. Expert Syst. Appl. **97**, 51–59 (2018)
11. Li, X., Ouyang, J., Lu, Y., Zhou, X., Tian, T.: Group topic model: organizing topics into groups. Inf. Retrieval J. **18**(1), 1–25 (2014). https://doi.org/10.1007/s10791-014-9244-9
12. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical dirichlet processes. J. Am. Stat. Assoc. **101**(476), 1566–1581 (2006)
13. Blei, D.M., Lafferty, J.D.: A correlated topic model of science. Annal. Appli. Stat. **1**(1), 17–35 (2007)
14. Jelodar, H., et al.: Latent Dirichlet allocation (LDA) and topic modeling: models,applications, a survey. Multimedia Tools Appli. **78**(11), 15169–15211 (2019)
15. Zhao, F., Zhu, Y.J., Jin, H., Yang, L.T.: A personalized hashtag recommendation approach using LDA-based topic model in microblog environment. Futur. Gener. Comput. Syst. **65**, 196–206 (2016)
16. Yin, Z.J., Cao, L.L., Han, J.W., Zhai, C.X., Huang, T.: Geographical topic discovery and comparison. In: Proceedings of the 20th International Conference on World Wide Web, pp. 247–256. ACM (2011)
17. Nallapati, R., Ahmed, A., Xing, E.P., Cohen, W.W.: Joint latent topic models for text and citations. In: Conference on knowledge Discovery Data Mining (KDD), vol. 14, pp. 542–550 (2008)
18. Shi, Q.W., Li, Y.N., Guo, P.L.: Dynamic finding of authors' research interests in scientific literature. J. Comput. Appli. **33**(11), 3080–3083 (2013)
19. Cui, W.W., et al.: Textfow: Towards better understanding of evolving topics in text. IEEE Trans. Visual Comput. Graph. **17**(12), 2412–2421 (2011)
20. Liu, H., Chen, Z., Tang, J., Zhou, Y., Liu, S.: Mapping the technology evolution path: a novel model for dynamic topic detection and tracking. Scientometrics **125**(3), 2043–2090 (2020). https://doi.org/10.1007/s11192-020-03700-5
21. Balili, C., Lee, U., Segev, A., Kim, J., Ko, M.: TermBall: tracking and predicting evolution types of research topics by using knowledge structures in scholarly big data. IEEE Access **8**, 108514–108529 (2020)
22. Borgatti, S.P., Everett, M.G.: A graph-theoretic perspective on centrality. Social networks **28**(4), 466–484 (2006)
23. Freeman, L.C.: A Set of measures of centrality based on betweenness. Sociometry **40**(1), 35–41 (1997)
24. Tsalouchidou, I., Baeza-Yates, R., Bonchi, F., Liao, K., Sellis, T.: Temporal betweenness centrality in dynamic graphs. Int. J. Data Sci. Analyt. **9**(3), 257–272 (2019). https://doi.org/10.1007/s41060-019-00189-x
25. Adebayo, I.G., Sun, Y.X.: A novel approach of closeness centrality measure for voltage stability analysis in an electric power grid. Int. J. Emerging Electric Power Syst. 3 (2020)
26. Hashemi, A., Dowlatshahi, M.B., Nezamabadi-pour, H.: MGFS: A multi-label graph-based feature selection algorithm via PageRank centrality. Expert Syst. Appl. **142**, 113024 (2019)

27. Cheung, K.F., Bell, M.G.H., Pan, J.J., Perera, S.: An eigenvector centrality analysis of world container shipping network connectivity. Transp. Res. Part E **140**, 101991 (2020)
28. Yin, H., Benson, A.R., Leskovec, J.: The local closure coefficient: a new perspective on network clustering. In: Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining, pp. 303–311 (2019)
29. Hu, Y.M., Yang, B., Wong, H.S.: A weighted local view method based on observation over ground truth for community detection. Inf. Sci. **355**, 37–57 (2016)
30. Fortunato, S.: Community detection in graphs. Phys. Report **486**(3–5), 75–174 (2010)
31. Souravlas, S., Sifaleras, A., Tsintogianni, M., Katsavounis, S.: A classification of community detection methods in social networks: a survey. Int. J. Gen Syst **50**(1), 63–91 (2021)
32. Blondel, V.D., Guillaume, J-L., Lambiotte, R., Lefebvre, E.:Fast unfolding of communities in large networks. J. Statistical Mech. Theory Experim., P10008 (2008)
33. Hu, Y.M., Yang, B., Duo, B., Zhu, X.: Exhaustive exploitation of local seeding algorithms for community detection in a unified manner. Mathematics **10**(15), 2807 (2022)