mathematics

MDPI

Article

Topic Network Analysis Based on Co-Occurrence Time Series Clustering

Weibin Lin [1,2], Xianli Wu [1], Zhengwei Wang [1], Xiaoji Wan [1,*] and Hailin Li [1,3]

[1] College of Business Administration, Huaqiao University, Quanzhou 362021, China
[2] TSL Business School, Quanzhou Normal University, Quanzhou 362021, China
[3] Research Center of Applied Statistics and Big Data, Huaqiao University, Xiamen 361021, China
* Correspondence: wanxiaoji@hqu.edu.cn

Abstract: Traditional topic research divides similar topics into the same cluster according to clustering or classification from the perspective of users, which ignores the deep relationship within and between topics. In this paper, topic analysis is achieved from the perspective of the topic network. Based on the initial core topics obtained by the keyword importance and affinity propagation clustering, co-occurrence time series between topics are constructed according to time sequence and topic frequency. Subsequence segments of each topic co-occurrence time series are divided by sliding windows, and the similarity between subsequence segments is calculated. Based on the topic similarity matrix, the topic network is constructed. The topic network is divided according to the community detection algorithm, which realizes the topic re-clustering and reveals the deep relationship between topics in fine-grained. The results show there is no relationship between topic center representation and keyword popularity, and topics with a wide range of concepts are more likely to become topic network centers. The proposed approach takes into account the influence of time factors on topic analysis, which not only expands the analysis in the field of topic research but also improves the quality of topic research.

Keywords: topic network; co-occurrence time series; sliding window; community detection

MSC: 91Fxx

check for updates

Citation: Lin, W.; Wu, X.; Wang, Z.; Wan, X.; Li, H. Topic Network Analysis Based on Co-Occurrence Time Series Clustering. *Mathematics* **2022**, *10*, 2846. https://doi.org/10.3390/math10162846

Academic Editor: Pedro A. Castillo Valdivieso

Received: 21 July 2022
Accepted: 9 August 2022
Published: 10 August 2022

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Copyright: © 2022 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

In the era of big data, a large number of scientific literature contains valuable potential information. Accurately identifying the hot spots in the development of disciplines, precisely predicting the development trend of science and technology, and rationally allocating scientific research resources, which have become the focus of common attention of experts and decision-makers in various fields. In particular, as the basis of scientific and technological information analysis, the identification and evolution of the topic of scientific literature is attracting more and more attention.

Traditional literature topic analysis and research mostly use bibliometric methods, such as citation analysis [1] and co-word analysis [2]. Citation analysis is to analyze the relationship between scientific literature and relevant features (such as keywords) by counting the frequency of citations. Co-word analysis is to count the number of occurrences of a group of keywords in the literature, which reflects the correlation between keywords, and the structural changes of disciplines represented by the keywords are analyzed. However, the semantic expressiveness and importance of keywords in different pieces of literature are often ignored. With the development of big data technology, semantic analysis, and artificial intelligence methods have emerged. The knowledge graph is a visualization method combining citation analysis with co-word analysis [3], which displays the core structure, development, and evolution process of disciplines by constructing a semantic network. Research on literature topics based on text mining [4] is becoming more and

more common, which solves the problems to some extent such as the lack of integrity and objectivity caused by the analysis of external characteristics of documents and the lack of in-depth analysis level. Li [5] found out high-frequency keywords through co-occurrence analysis, then calculated the co-occurrence matrix by Ochiia coefficient, found literature topics through cluster analysis, defined the time series of topic popularity, and realized topic classification, and evolution trend analysis. Zhang et al. [6] proposed a new K-Means clustering method combined with a word embedding model, which can effectively extract topics from literature data. The method has the advantages of fast, simple, high efficiency, and scalability, but the defects of K-Means will have a great impact on topic analysis. By reviewing the literature on value justice in the energy sector, Wildt et al. [7] explained the difference between the proposed methods and the conventional method based on keywords. The new method provides a more comprehensive overview of the relevance of energy justice, but the physical meaning of singular value decomposition is difficult to understand, and the clustering effect of word meaning is difficult to control. Current topic models have been able to extract high-quality topic words, which can improve the current situation that co-word analysis can not effectively express the semantic relationship between words, and effectively reveal the true reflection information of the topic. However, the above mentioned approaches still have some defects. Many high-frequency but meaningless words will affect the effectiveness of the topic analysis, and the determination of the number of topics also has great controversy and subjectivity. Besides, linguistic features such as synonyms, polysemy, and homonymy need to be considered [8].

To accurately mine topics, many scholars have proposed specific topic models. Latent Dirichlet Allocation (LDA) is the most commonly used topic model, and improved models based on LDA have been proposed continuously [9]. Jung et al. argued that there are differences between topics and keywords, topics are fuzzy concepts conceived by the author. Therefore, an author-based topic model was proposed [10]. However, LDA models do not consider the correlation between words, the performance of LDA models is limited. Deep learning has significant advantages in text mining [11], and various topic models based on neural networks have been derived. Natural Language Processing (NLP) is an important application of deep learning in text analysis, which considers the language syntax and language semantics. Thus, NLP effectively improves the accuracy of tasks such as text classification or text mining [12]. Based on adversarial training, Wang et al. proposed a novel topic model, which is called the "Adversarial-neural Topic Model". With the rapid development of NLP, the text model is improved, and the topic mining is more accurate [13]. Compared with the traditional LDA model, the performance of the novel model has significant advantages. Through NLP techniques, the tasks of text mining can be effectively completed. Recently, Kim et al. proposed a model called "architext", which is designed for tight integration in interactive hierarchical topic modeling systems, and has good performance in large data sets [14]. Besides, a topic model based on a distributed system can greatly improve the training efficiency [15]. At present, topic models are constantly developing, and both accuracy and efficiency are optimized.

Time series can be used for both clustering [16] and transforming into complex networks [17]. Traditional analysis approaches such as clustering and classification can be used to study the relationship between topics, but lack of consideration of the impact of time factors on the study of topic evolution features. Meanwhile, although the research results can reflect the relationship between topics, the applicability of the conclusions to other topics needs further study. Topic analysis can be marked according to users' interests. Knowledge labels are subjective, and the relationship between topics is reflected in a parallel structure. There is a lack of explicit representation and analysis of subordination, and there is no strict knowledge architecture. The topic network structure has a certain level and subordinate relationship, which can be used to reflect the hierarchical relationship within the knowledge network according to certain knowledge processing steps and the constraints of the corresponding knowledge management norms. Time series data mining techniques and approaches can study the relationship of a certain object from the perspective of time

change. Through the time series data transformation of topic features, relevant techniques and approaches can be used to study the topic evolution relationship. Zhu et al. used LDA model to extract topics, and constructed "word-topic" coupling network [18]. Wu et al. proposed a new approach based on the LDA model, and predicted the trends of specific topics [19]. While time series clustering is a popular research approach [20]. Combined with complex networks, time series clustering can improve the clustering effect [21]. Using time series data, combining clustering with complex network algorithms to mine popular topics, is an innovative approach. Therefore, considering the different importance of different keywords in each literature, the keyword information is transformed into a topic co-occurrence time series that reflects the change of the topic, and the topic analysis is carried out from the perspective of complex network clustering. Therefore, by considering the different importance of different keywords in each literature, the keywords are transformed into a topic co-occurrence time series that reflects the topic change, and the topic analysis is achieved from the perspective of complex network clustering. To construct the similarity matrix and topic network of the topic co-occurrence time series data, the initial core topics in related fields are obtained by keyword importance and Affinity Propagation (AP) clustering algorithm [22], and the topic co-occurrence time series are constructed according to the time sequence and the frequency of topic co-occurrence. The topic co-occurrence time series are divided into subsequences through sliding windows, and the cosine similarity measure is used to search the most similar subsequence segments [23]. By community detection (Louvain) [24], the topic network is clustered to realize the topic extraction and analysis of literature. The literature data related to "network information security" from 2005 to 2019 collected by China National Knowledge Infrastructure (CNKI) [25] is taken as the data source. This paper constructs and analyzes the topic network in network information security, identify the core hot topics within a specific period of time and the deep-seated relationships within and between topics, then provide decision support for related scientific research institutions and personnel on subject research.

This paper has made some contributions from research methods: (1) The topic co-occurrence sequence similarity matrix is constructed by sliding windows, which avoids the impact of traditional hard partitioning on topic results and reveals the deep relationships between topics in a fine-grained way. (2) Cosine similarity is used to measure topic co-occurrence subsequence similarity, and the correlation between topics is measured from the shape and trend of the sequence data to avoid the impact of zero co-occurrence frequency on the measurement results. (3) The relationship between topic time series is transformed into a topic network. The more similar subsequences in the two time series, the more similar the relationship between the two topics is. The relationship between time series data is preserved, and the influence of abnormal data in the series on the similarity measurement of the whole series is reduced. Meanwhile, combined with Matrix Profile (MP) [26], the time complexity of similarity measurements of topic co-occurrence time series is reduced.

The structure is as follows: Section 2 introduces the research framework, methods, and algorithms. Section 3 mainly visualizes the research results and provides suggestions. Finally, Section 4 summarizes the full text, and puts forward our shortcomings and prospects.

2. Methodology

This section introduces the acquisition of core topics, the similarity measurement of topic co-occurrence sequences, and the construction of community networks, including AP clustering, sequence division with matrix profile, and Louvain algorithm.

2.1. Research Framework

To address the lack of deep-seated relationships within and between topics in traditional research, based on the importance of keywords and the core topics obtained by AP clustering, the co-existence sequences between the two topics are calculated. Each co-existence sequence is divided into sequence segments according to sliding windows by matrix profile. The similarity between all subsequences is calculated by cosine similarity

measure to search for the most similar subsequences corresponding to each subsequence. The number of subsequences that are most similar between each topic sequence is counted, which is the similarity values between the two topics, and a similarity matrix is obtained. According to the similarity matrix, the network is constructed. The topic is the vertex of the network, and the two topics with the most similar subsequence segments are connected. Through community detection, the topic network is divided. The closely related topics are divided into a community, indicating that the topics in the same community are strongly correlated, and the topics in different communities are weakly correlated or uncorrelated, resulting in the "high cohesion and weak coupling" effect of topic division. The specific research framework is as shown in Figure 1.

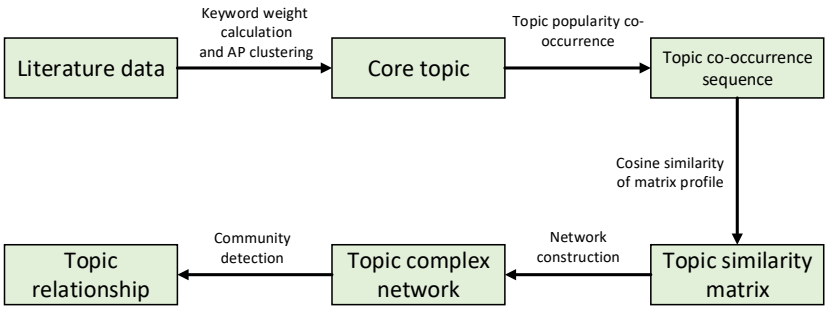

Figure 1. Research framework.

2.2. Core Topic

The existing literature topic recognition research mainly uses the co-word analysis to construct the keyword similarity matrix, and then uses the multi-dimensional scale analysis or hierarchical clustering to cluster, and extracts and analyzes the topic according to the specified multi-dimensional scale or clustering number. Traditional topic analysis lacks consideration for the topic performance of keywords in different literature, and the number of clusters needs to be set artificially. Therefore, the topic results may not necessarily reflect the connotation of the core topics. To compensate for such shortcomings, the order of keywords in different documents is considered, and the importance of keywords in literature data is calculated. The weighted similarity matrix of keywords is adaptively clustered by AP clustering, and then the core topics of the corresponding keyword clusters are represented by the center of each cluster. Meanwhile, the topic of evolution law changing with time is proposed. Finally, the initial core topics of related fields are obtained.

For the set with N literature, the keyword similarity measurement based on importance is used, the formula is as follows:

$$Sw(i,j) = \frac{\sum_{i,j \in Key_{p\bullet}}^{N} wKey_{pi'} \times wKey_{pj'}}{\sqrt{\sum_{i \in Key_{p\bullet}}^{N} wKey_{pi'}^2} \sqrt{\sum_{j \in Key_{p\bullet}}^{N} wKey_{pj'}^2}} \quad (1)$$

where $wKey_{pi'}$ and $wKey_{pj'}$ respectively represents the weight of the keyword i and j corresponding to the keyword i' and the keyword j' in literature P, $i,\ j \in Key_{p\bullet}$ represents the two keywords i and j jointly appear in the keyword set $Key_{p\bullet}$ of literature P.

If literature P has K_p keywords, the keyword set in order $Key_{p\bullet} = \{Key_{p1}, Key_{p2}, \cdots, Key_{pK_p}\}$ is listed. In literature P, the importance of the k-th keyword to the description of the document topic is as follows:

$$wKey_{pk} = \frac{K_p - k + 1}{\sum_{k=1}^{K_p} k} \quad (2)$$

The similarity of the target keywords *Keys* is calculated, and the similarity matrix Sw of the target keywords are obtained. To realize clustering analysis of keywords, AP clustering is used for the similarity matrix. The clustering results aggregate keywords with

high similarity into clusters, and each cluster forms a topic with keywords corresponding to the center of the cluster as the most descriptive object, and K topics T can be obtained. The formula is as follows:

$$T = AP(Sw, Keys) \tag{3}$$

where $T = \{T_1, T_2, \cdots, T_K\}$.

2.3. Co-Occurrence Time Series

Co-occurrence time series can detect the temporal relationship patterns of the "event" [27]. Considering that the topics are paid different attention in different time periods and have obvious temporal characteristics, according to the occurrence time of the topics, the calculation of co-occurrence frequency between each pair of topics over time can reflect the evolutionary correlation between topics. The information features of a topic h can be described by its co-occurrence features with other topics, and the co-occurrence features between topics are characterized by the co-occurrence between keywords contained in the topic. If the frequency of co-occurrence between topic T_i and another topic T_j in a certain period t is $F_t(i,j) = f_t(T_i \& T_j)$, then the information feature of the topic T_i is expressed as $P_t(i,:) = [F_t(i,1), F_t(i,2), \cdots, F_t(i,L)]$, which is the co-occurrence frequency vector of topic T_i and other topics, where L is the number of topics and $i \neq j$. Besides, a topic reflects the common information and common knowledge of multiple keywords, that is, keywords with common knowledge can be expressed by a core keyword through clustering. Therefore, the co-occurrence frequency of topic T_i and topic T_j in period t can be reflected by the co-occurrence frequency between the keywords covered by the two topics. The formula is as follows:

$$F_t(i,j) = f_t(T_i \& T_j) = \sum_{n=1}^{k} \sum_{m=1}^{s} f_t(T_i(n) \& T_j(m)) \tag{4}$$

where $T_i(n)$ and $T_j(m)$ are the n-th and m-th keywords in topic T_i and topic T_j respectively, k and s are the number of keywords covered in the two topics respectively.

In topic analysis, some topics have a high degree of attention, while others have a low degree of attention, which means there are significant differences in popularity between topics. Therefore, it is necessary to eliminate the influence of the difference in topic information features caused by different disciplines or research directions. A description of the topic information features is defined as the ratio of the topic's co-occurrence relationship $f_t(T_i \& T_j)$ to the topic popularity $h_t(T_i)$. The topic popularity can be reflected by the frequency of keywords in a given period covered by the topic, and the topic popularity is defined as $h_t(T_i) = f_t(T_i) = \sum_{n=1}^{k} f_t(T_i(n))$. Therefore, the information features of topic T_i described by topic T_j in period t can be expressed as:

$$F_t^{T_i}(T_j) = \frac{f_t(T_i \& T_j)}{h_t(T_i)} = \frac{\sum_{n=1}^{k} \sum_{m=1}^{s} f_t(T_i(n) \& T_j(m))}{\sum_{n=1}^{k} f_t(T_i(n))} \tag{5}$$

The information features of topic T_i depicted by topic T_j in period t are reflected by the co-occurrence relationship between topic T_i and T_j in period t, and the information features reflect the correlation between the two topics. If period t is extended to several periods $t = 1, 2, \cdots, years$, then the information features of topic T_i represented by topic T_j can be expressed as the co-occurrence sequence between two topics, which is called "topic co-occurrence sequence". The formula is as follows:

$$L(i,j) = \left[F_1^i(j), F_2^i(j), \cdots, F_{years}^i(j)\right] \tag{6}$$

where *years* represents the length of time.

2.4. Similarity Measurement

By measuring the similarity between the co-occurrence sequences of each topic, the strength of the relationship between the two topics can be obtained. Traditional time series' similarity mostly adopts two measurements: Euclidean distance and Dynamic Time Warping (DTW) [28]. These two measurements measure the similarity between sequences from the numerical value, and have high requirements for numerical similarity. Combined with practical applications, many similar topic co-occurrence sequences cannot be completely consistent in all positions, and there will be some numerical differences. Such differences will be amplified when using Euclidean distance or DTW, resulting in the measurement error of the relationship between the following topics. Therefore, the similarity between topics is measured by the similarity of subsequence fragments. Each topic co-occurrence sequence is divided through the sliding window to get all co-occurrence subsequence segments. To search the topic corresponding to the most similar subsequence fragment, cosine similarity comparisons are made one by one with all other topic co-occurrence sequences, and the similarity between the two topics is increased by 1. In essence, the more similar subsequence fragments the two topic time series have, the more likely the two topic time series are to be similar. For all topic co-occurrence time series, the subsequence segments of all sequences are matched in the same time window, the most similar two subsequence segments are searched under the time window, and the correlation matrix between all topic co-occurrence sequences is calculated by the time window movements.

Assuming that there are two topic time series, they are $X = [x_1, x_2, \cdots, x_m]$ and $Y = [y_1, y_2, \cdots, y_m]$, the length of sliding time window is l. The two subsequence sets of time series X and Y obtained by moving the window are $X^{'} = \left\{X_1^{'}, X_2^{'}, \cdots, X_{m-l+1}^{'}\right\}$ and $Y^{'} = \left\{Y_1^{'}, Y_2^{'}, \cdots, Y_{m-l+1}^{'}\right\}$. The distance between $X_i^{'}$ and each element in $Y^{'}$ is calculated, and a vector with length $m - l + 1$ is obtained. According to matrix profile, the minimum distance between all subsequences in $X_i^{'}$ and $Y^{'}$ is $MP_{X_i^{'}Y^{'}}$. The formula is as follows:

$$MP_{X_i^{'}Y^{'}} = \min_j D_{X_i^{'}Y_j^{'}} = \min\left\{D_{X_i^{'}Y_1^{'}}, D_{X_i^{'}Y_2^{'}}, \cdots, D_{X_i^{'}Y_{m-l+1}^{'}}\right\} \tag{7}$$

where $D_{X_i^{'}Y_j^{'}}$ is a variant measure of Euclidean distance or cosine similarity of sequence segments $X_i^{'}$ and $Y_j^{'}$. Topic co-occurrence time series data usually have a value of 0, cosine similarity variants are more likely to be chosen as the vector distance measurement of sequence segments, and the formula is as follows:

$$D_{X_i^{'}Y_j^{'}} = 1 - \frac{\sum_{k=1}^{l} X_i^{'}(k) Y_j^{'}(k)}{\sqrt{\sum_{k=1}^{l} X_i^{'2}(k)} \sqrt{\sum_{k=1}^{l} Y_j^{'2}(k)}} \tag{8}$$

For a data set H composed of n topic co-occurrence time series, the length of the time moving window is l, and the subsequence fragment set of all topic co-occurrence sequences in the k-th time window is recorded as $H_k^{'} = \left\{H_k^{'}(1, 1:l), H_k^{'}(2, 1:l), \cdots, H_k^{'}(n, 1:l)\right\}$, $H_k^{'}(i, 1:l)$ represents the subsequence segment of the i-th topic time series under the k-th time window, and the length of the sequence segment is l. Assuming that $X_i^{'} = H_k^{'}(i, 1:l)$, and Y' represents a sequence fragment set that excludes sequence fragments $H_k^{'}(i, 1:l)$ from the subsequence fragment set $H_k^{'}$, that is, $Y' = H_k^{'} - H_k^{'}(i, 1:k)$. To get that the subsequence most similar to topic i appears in topic J_k^i under the k-th time window, $X_i^{'}$ and Y' are substituted into Equation (4) to calculate $MP_{X_i^{'}Y^{'}}$, the formula of J_k^i is:

$$J_k^i = \arg\min_j D_{X_i' Y_j'} \tag{9}$$

where $Y_j^{'} \in Y'$. By moving the time window $m - l + 1$ times with a length of l, the similarity matrix between topics is expressed as $\overrightarrow{S}$, and the similarity $\overrightarrow{S}(i,j)$ between topic i and topic j is as follows:

$$\overrightarrow{S}(i,j) = \sum_{k=1}^{m-l+1} \left(1 - \left|Sign\left(J_k^i - j\right)\right|\right) \tag{10}$$

where $Sign(x)$ is a symbolic function. The Equation (7) counts the number of times the subsequence fragments of topic i are most similar to those of topic j in different time windows. The optimal matching of similarity between subsequence segments in the same time window is not symmetrical, and the similarity or correlation between topic co-occurrence time series is described by the similarity or correlation between the parts of subsequence segments. Therefore, the correlation between the final topics is not symmetrical, that is, $\overrightarrow{S}(i,j) \neq \overrightarrow{S}(j,i)$. In topic analysis, the relationship between topics usually changes in different periods, and the asymmetric similarity measurements reflect the relationship between topics in different periods.

2.5. Topic Complex Network and Community Detection

Based on the similarity matrix $\overrightarrow{S}$, the topic complex network is constructed as $G = (V, E\ W)$, $V = (v_1, v_2, \cdots, v_n)$ is the network node set, v_i is the topic i in the dataset, and n is the size of the dataset. $E = \left(e_{1,2}, e_{1,3}, \cdots, e_{i,j}, \cdots, e_{n,n-1}\right)$ is the directed edge set of a network, where $e_{ij} = \left(v_i, v_j\right)$. $W = \left(w_{1,2}, w_{1,3}, \cdots, w_{i,j}, \cdots, w_n\right)$ is the edge weight set, and $w_{i,j} = \overrightarrow{S}(i,j)$ represents the correlation between topic i and topic j, topics T_i and T_j are sequences represented by points v_i and v_j respectively.

Louvain is a community detection algorithm based on modularity, which can detect the hierarchical community structure. The optimization goal is to maximize the modularity of the community network, and the optimization efficiency and optimization effect are relatively good. Modularity [29] can evaluate the effect of community network division, whose physical significance is that the number of sides of a community is only different from that in random cases, and the value range is $[-1, 1]$. Modularity is defined as follows:

$$Q = \frac{1}{2s} \sum_{i,j} \left[W_{ij} - \frac{k_i k_j}{2s}\right] \delta(C_i, C_j) \tag{11}$$

where W_{ij} is the weight of the edges between nodes i and j. If the complex network is not a weighted graph, the weight of all edges is the same and is 1, $k_i = \sum_j W_{ij}$ is the sum (degree) of the weights of the edges connected to node i. C_i is the community of node i, $s = \frac{1}{2}\sum_{i,j} W_{ij}$ is the sum of the weights of all edges.

The community detection algorithm based on modularity takes the maximization of modularity Q as the optimization goal. The algorithm moves node i to community C_j, and then calculates the change of modularity, measures if i should be planned for community C_j by ΔQ, where ΔQ is as follows:

$$\Delta Q = \left[\frac{\sum_{in} + 2k_{i,in}}{2s} - \left(\frac{\sum_{tot} + k_i}{2s}\right)\right] - \left[\frac{\sum_{in}}{2s} - \left(\frac{\sum_{tot}}{2s}\right)^2 - \left(\frac{k_i}{2s}\right)^2\right] \tag{12}$$

where $\sum_{in}$ is the sum of the weights of the internal connecting edges of the community C_i, $\sum_{tot}$ is the sum of the edge weights connected to all nodes in the community C_i. k_i is the sum of weights of all edges connected by node i, $k_{i,in}$ is the sum of edge weights connected between node i and nodes in community C_i, and s is the sum of the weights of all edges

in the network. Algorithm 1 calculates the change in modularity when i is removed from the community.

Algorithm 1 Louvain algorithm.

Input: Network G
Output: Divided community C

1: Initialize and set each node as a community
2: Calculate the modularity change ΔQ of each node i according to Equation (12)
3: Select the node k with the largest ΔQ
4: Divide node i into the community C_k where k is located if the maximum ΔQ is greater than 0
5: Repeat the second step until the communities of all nodes no longer change.
6: Compress all nodes in the same community into a new node to get a new G
7: Repeat the first step until the modularity of the whole network remains unchanged, and then return C

In the first step of Algorithm 1, each node of the network is assigned a different community, the number of communities is the same as the number of nodes. In the second step, for each node i, to evaluate the modular gain change, the algorithm needs to remove i from the corresponding community, and then put i in the community of neighbor node j for calculation. In the third step and the fourth step, node i is placed in the community C_k with the largest gain and positive value (in the case of a draw, a random rule is used). If the modularity changes to a negative value, i will remain in the original community. The more closely related topics are divided into a community, the topics of the same community are strongly related, and the topics of different communities are weakly related or irrelevant.

3. Literature Topic Analysis

This section shows the results of our proposed approach to the dataset, including clustering of core topics, topic correlation analysis, construction, and clustering of topic complex networks.

3.1. Dataset

The literature on "network information security" in the full-text database of China National Knowledge Infrastructure (CNKI) journals was taken as data analysis materials, the search keyword was "network information security", and the publication time range was from January 2005 to October 2019. Multiple kinds of literature such as notices, solicitation documents, and interviews were excluded, 18,085 literature and 17,544 literature keywords were collected. The threshold of high-frequency keywords was set to 30, and 278 high-frequency keywords were obtained. To observe the year distribution of the publication volume, the distribution map of the publication volume is drawn, as shown in Figure 2.

In Figure 2, the number of literature published in 2005 was the least (41), and the number of literature published in 2018 is the highest (2453). Meanwhile, the number of literature during the six years from 2005 to 2011 is relatively low, the main reason is that the database maintains the integrity of the literature of the past 10 years, and regularly cleans up the literature more than 10 years. Over time, the literature of more than 10 years is gradually deleted from the database. To avoid the unbalanced impact of the annual literature collection, the annual literature data was standardized.

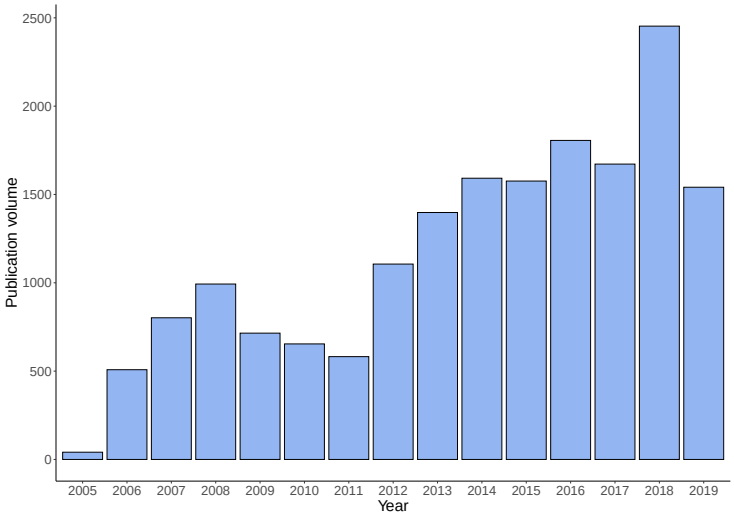

Figure 2. Publication volume in each year from 2005 to 2019.

3.2. Core Topic Acquisition

Through literature research approach [5], 278 high-frequency keywords were clustered [16], and 42 topics such as "security", "intrusion detection", "enterprise management" and "hierarchical protection" were obtained. The topic clustering results are shown in Figure 3.

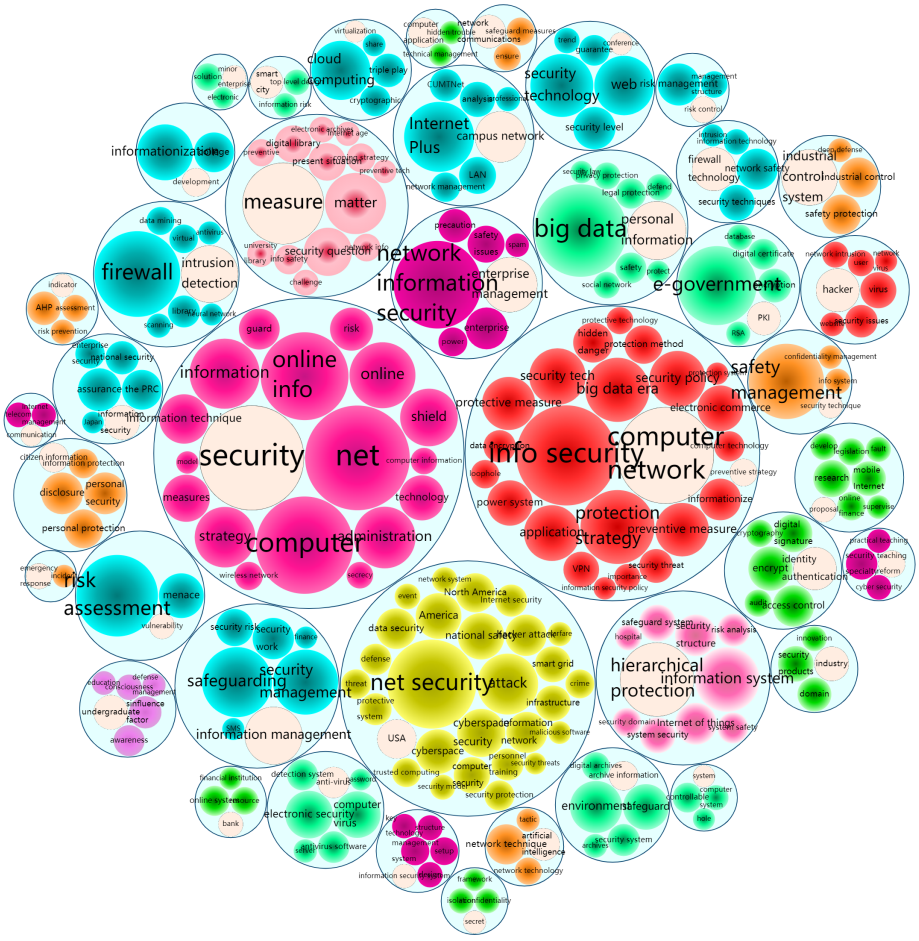

Figure 3. Topic clustering results of high frequency keywords.

In Figure 3, each circle with a border represents a cluster (class or topic), while the largest borderless circle in the cluster represents the keyword with the greatest attention in the cluster, and the circle with a different color from that in the cluster represents the center representative object of the cluster. For instance, in the cluster where "info security" is the most concerned, "computer network" is the cluster center, which better summarizes the topics expressed by other cluster members, that is, the related research with "computer network" as the topic usually includes the research of "info security", "protection strategy" and "big data era".

According to the clustering analysis of keywords, Topics related to security measures (protection strategy, intrusion detection, hierarchical protection, etc.) are usually popular, and the popular topics have received high attention in the field of network information security in the past 16 years. According to Figure 3, the high attention keywords in each cluster are not necessarily the representative points of the topic, because the representative points in the center of the cluster are keywords that are more closely related to other topic members, not keywords with high attention. Therefore, in the same topic study, the more focused keywords do not mean that they are more closely related to other keywords, which is an innovative conclusion. For example, the cluster whose central point is "USA", that is, the cluster of "net security" with the most attention, includes "net security", "attack", "cyberspace", "cyberspace security", "America", "North America", "national safety", "smart grid", "hacker attack", etc. Most of the keywords in such a cluster reflect topics that seem to be related to network security, which is only the surface semantics of the keywords. The literature topics contained in a such cluster are mainly related to network security studies related primarily to the United States or the United States of America. Therefore, keywords with high attention may not reflect the central idea of the topic of the corresponding cluster, while keywords with low attention may be the central representative of the topic.

3.3. Topic Correlation Analysis

In Figure 3, 42 core topics can be obtained, and the number of occurrences of each topic in the corresponding year can be obtained in chronological order, which can be reflected by the frequency of occurrences of other keywords covered by the topic. Therefore, the co-occurrence time series between topics can be expressed as the co-occurrence frequency change between keywords covered by two topics over a period. For instance, the topic co-occurrence sequence of topic A relative to topic B in 2005–2019 is a sequence with a length of 15. The value of each element in the sequence is obtained by dividing the co-occurrence frequency of two topics (all keywords covered respectively) in the corresponding year by the occurrence frequency of topic A. The co-occurrence time series of topics not only consider the influence of time factor, but also reflect the co-occurrence relationship between the two topics by the co-occurrence relationship of the keywords covered, which better observe the change or evolution relationship between topics. For 42 topics, the corresponding topic co-occurrence sequence is established according to the Equations (4)–(6). A particular topic needs to establish a co-occurrence sequence with 42 other topics, including a co-occurrence time series with the topic itself. Therefore, each topic builds 42 co-occurrence time series with other topics, and then 42×42 topic co-occurrence sequences can be obtained. Each topic co-occurrence time series reflects the changing co-occurrence relationship between a specific topic and another topic. As shown in Figure 4, the blue time series $L(i, j)$ records the co-occurrence relationship between topic i and topic j over time, and the red lines denote their subsequences. Considering the influence of the great difference in the research popularity of each topic, the topic popularity was standardized according to Equation (6), and the topic co-occurrence sequence $L(i, j)$ was not equal to $L(j, i)$, that is, $L(i, j) \neq L(j, i)$.

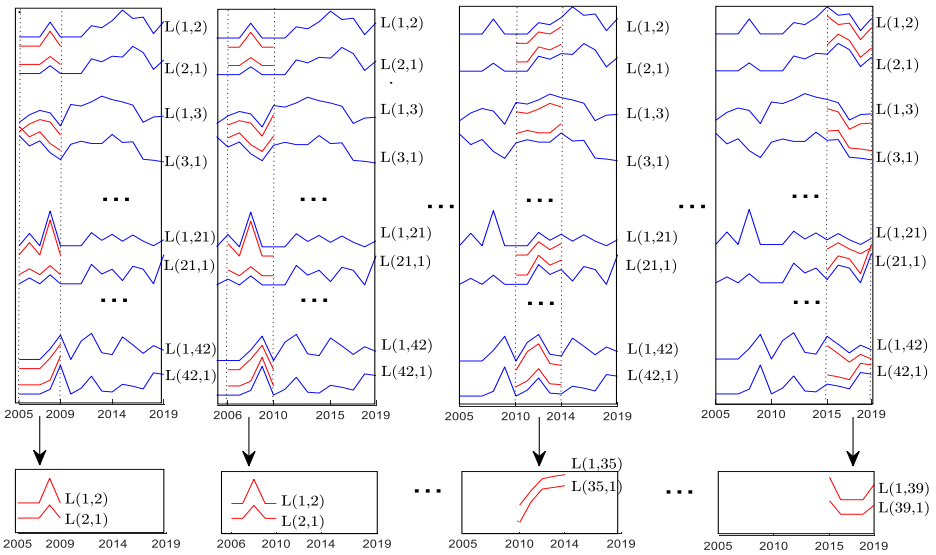

Figure 4. The measurement process for the first topic relative to all other topics.

The sliding window was set to 5, and 42 × 42 topic co-occurrence sequences were obtained. To measure the correlation between the 42 topics, each topic co-occurrence sequence with a length of 15 (2005 to 2019) was divided into 11 $(15 - 5 + 1)$ subsequences with a length of 5. In Figure 4, the correlation calculation process between the first topic and all other topics is displayed. The blue sequence $L(i, j)$ represents the co-occurrence sequence of the i-th topic relative to the j-th topic, which can be calculated from Equations (4)–(6), and the red sequence represents the sub-sequence of the corresponding year. The similarity of subsequence fragments of different topic co-occurrence sequences under each same time window is calculated, and two most similar subsequence fragments are obtained, which can reflect that the corresponding topics of the two subsequence segments are most relevant in the period. For example, according to the calculation as shown in the first subfigure in Figure 4, topic 1 and topic 2 are the most relevant from 2005 to 2009, which means that among the correlations between research topics, topic 1 and topic 2 are the most relevant during 2005–2009, thus the total correlation between the two topics increased by 1, which also shows that the total correlation between topics is reflected by the correlation between local subsequence fragments. Similarly, the second subfigure is the window that slides to 2006–2010, and the calculation results show that topic 1 is the most relevant to topic 2. The third subfigure is the window that slides to 2010–2014, and the most relevant topic related to topic 1 is topic 35. In the fourth subfigure, in 2015–2019, topic 39 is most relevant to topic 1. In such a way, the total correlation between topic 1 and the rest of the topics can be obtained.

As shown in Figure 5, the first topic is "teaching reform". Through the statistical analysis of the optimal correlation of local subsequence fragments, the topic correlation between topics can be obtained. The topic is regarded as a vertex in the network graph, and the relationship between topics is regarded as an edge. Figure 5 shows the connection diagram of other topics related to the topic of "teaching reform", "USA" is the most relevant, followed by "vulnerability" and "information management", which shows that in 11 sliding time windows, the topic "teaching reform" is the most relevant to the topic "USA" in 3 times, and the topic "teaching reform" is the most relevant to "information management" in 2 times.

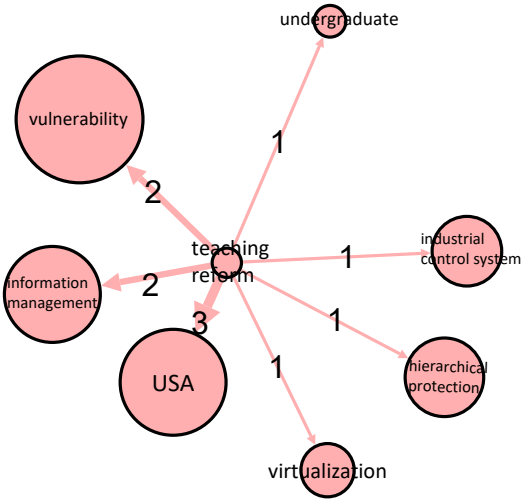

Figure 5. Connection diagram of topics related to "teaching reform".

3.4. Topic Network and Community Division

The relationship between a specific topic and other topics can be drawn as a topic connection diagram as shown in Figure 5, which means that the topic of "teaching reform" can be locally most relevant to 7 of the other 42 topics. Therefore, the network diagram formed by the nodes with the topic of "teaching reform" as the output and the nodes with the other 7 topics as the input is constructed. The local correlation between a specific topic and other topics can be obtained, and then the edges of all topics are displayed. By merging all topic connection diagrams into the same network, a complex network diagram of 42 topics is formed, as shown in Figure 6.

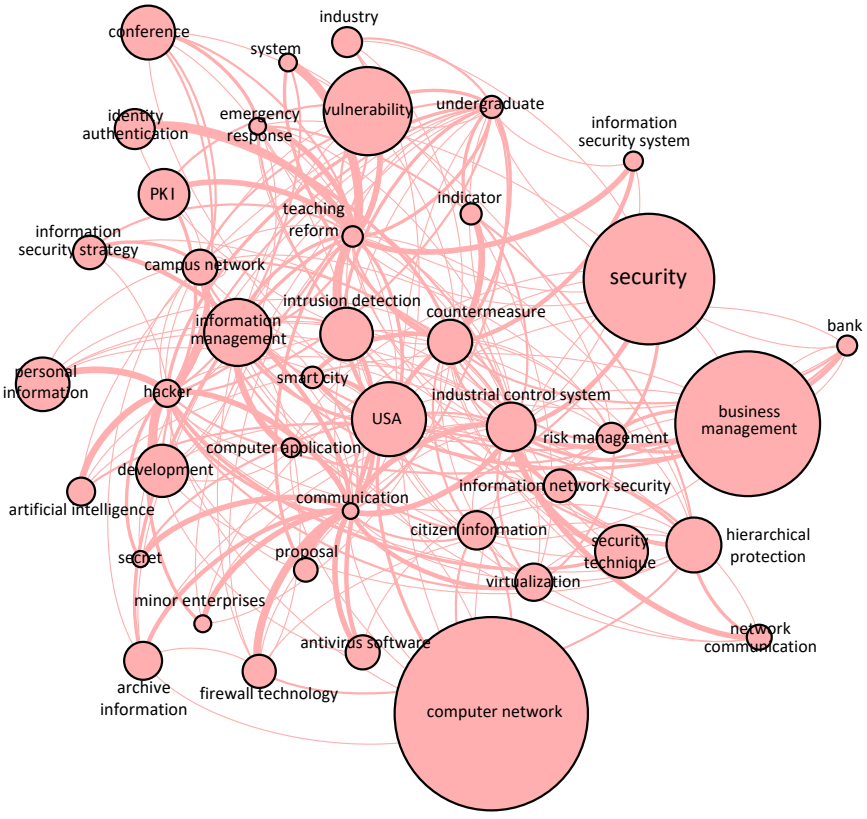

Figure 6. Complex networks of 42 topics.

The network of 42 topics in Figure 6 are divided by the Louvain algorithm. The topics with strong correlation are divided together, and then the division results of constructing topic complex network communities in the form of sliding windows are obtained, as shown in Figure 7.

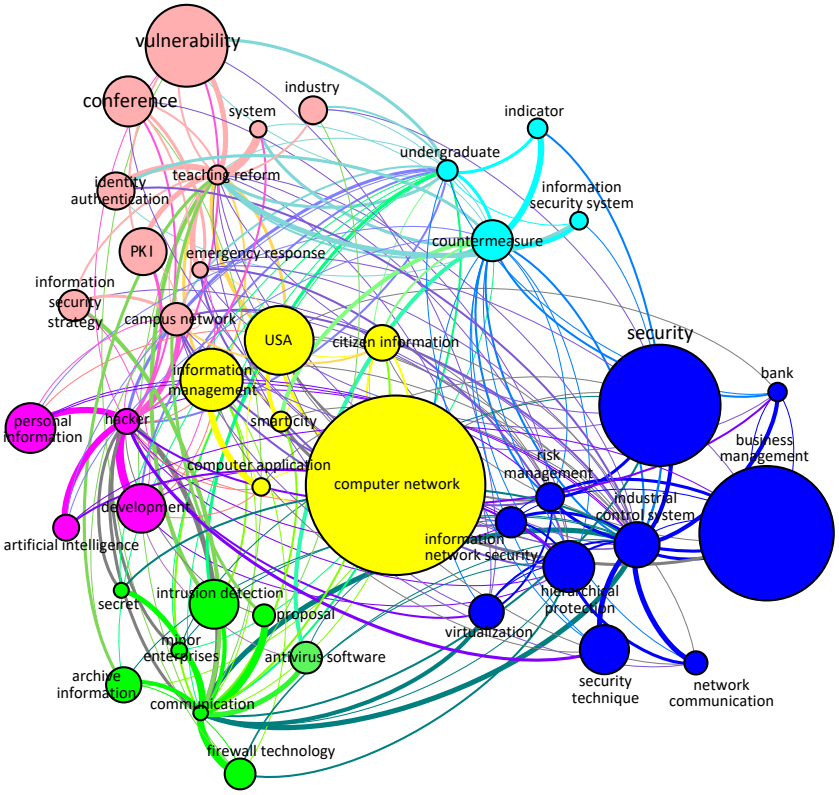

Figure 7. Topic network relationship constructed by sliding window.

In Figure 7, 42 topics are contained. Each topic is represented by a circle. The identifier on each circle is the topic name. The size of the circle indicates the frequency of the topic in 2005–2019. The more the number of occurrences, the larger the circle. The color of circles represents the clustering results of topics. Circles with the same color belong to the same cluster. The edge between circles represents the correlation between the connected two topics. The thicker the edge, the more relevant the two topics. According to the community division, 42 topics are divided into 6 clusters. The topics within and between clusters are connected, and the topics within clusters are more closely connected. Each of the 6 clusters has a relatively central topic, namely "teaching reform", "hacker", "communication", "industrial control system", "USA" and "measure", which have a broader meaning. Due to the greater likelihood of co-occurrence with many topics in the literature, topics with a wide range of meanings are more likely to become central links between topics. Therefore, when obtaining literature in related fields, users can search for topics in the form of "specific keywords + fuzzy keywords", so that the obtained literature topics are more accurate and the results are more in line with the research needs. Besides, in Figure 8, the frequency of topics with central connections are usually not high. The reason is that such topics have a wide range of concepts. In the compilation process of actual literature, words with wider meaning are less used as keywords, and the content of literature is generally specific. However, since fuzzy retrieval is easier to query results [30], topics with broader concepts are likely to appear alongside other topics.

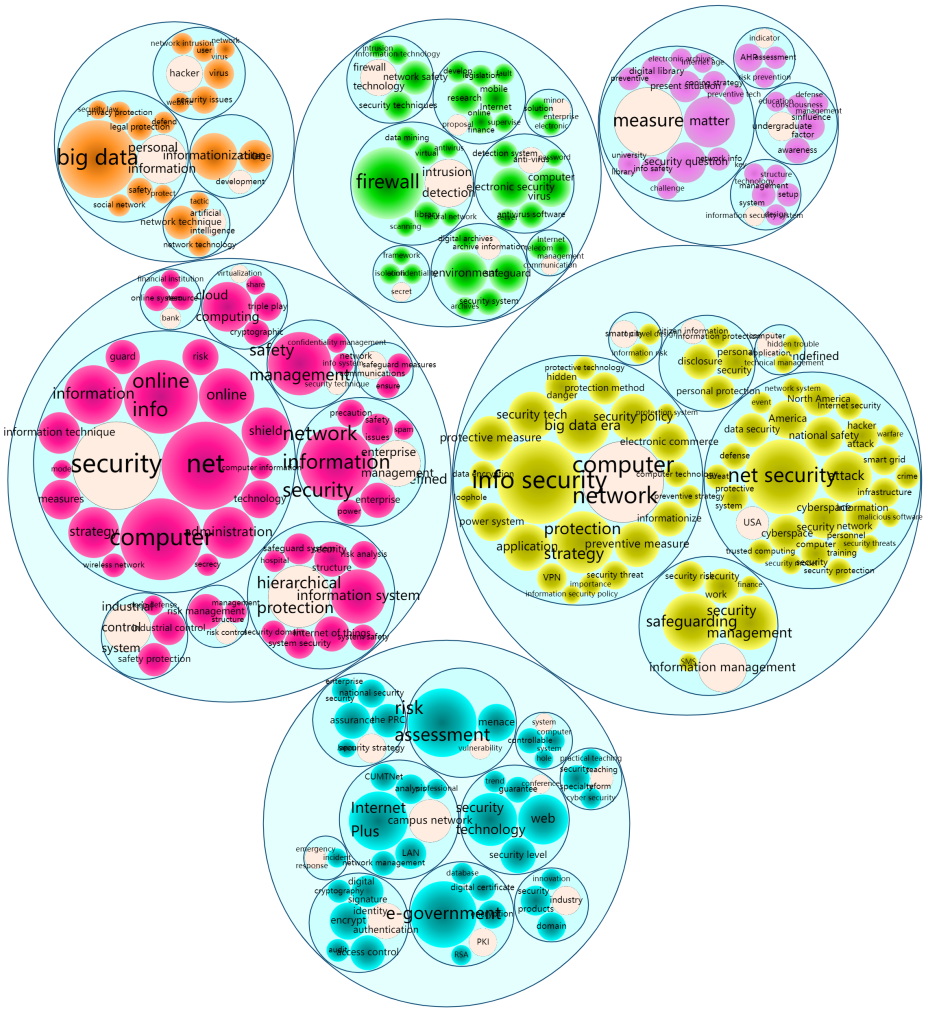

Figure 8. Clustering results of network constructed by sliding window.

From the content of literature, for instance, the cluster of topic "communication", which also includes the topics "anti-virus", "firewall technology", "minor enterprise", "archive information", "secret", "proposal" and "intrusion detection". Only according to the word meaning, the general meanings of the three topics, "anti-virus", "firewall technology" and "intrusion detection" are relatively related, the word meanings of other topic members are quite different, which is because the clustering measurement not only considers the word meaning, but also includes the influence of time factors and the relationship of co-occurrence.

In most clusters, the network structure is star distributed, and a central keyword is closely connected with other cluster members. The network structure measures the correlation degree of the members in the cluster. For example, in Figure 7, "computer network", "USA", "information management", "citizen information", "smart city" and "computer application" are divided into a cluster. Although "computer network" appears frequently, its relationship with other members in the cluster is not very close. "USA" and "smart city" are closely related, "information management" and "computer application" are closely related, "citizen information" and "computer network" are relatively close with "USA". According to the actual analysis of the literature and the actual situation, the relationships are reasonable [31,32], which shows that the topic network analysis based on co-occurrence time series clustering can effectively display hidden and unknown topic relationships in the literature.

4. Conclusions

This paper proposes a topic network analysis approach for co-occurrence time series clustering, which takes the importance of keywords and the initial core topics obtained by AP clustering as the initial point. We construct the co-occurrence time series between topics in time order, and use the time window and the matrix profile to obtain the local correlation of the co-occurrence time series data, Then the similarity matrix reflecting the total correlation is obtained. Based on the similarity matrix, a topic complex network is constructed, and the community of the topic network is divided by the Louvain algorithm, which realizes the re-clustering of topics in related fields, and analyzes the deep relationship between and within topics. Through the analysis of the literature in the field of network information security, the results show that the new approach can reveal the popular topics in a specific field and the deep relationships in a fine-grained way from the perspective of the time change, which not only complements the application of time series data mining in topic analysis, but also provides decision support for institutions and personnel to grasp the subject research direction. The main innovations of this study are as follows: (1) The topic co-occurrence time series are constructed through the co-occurrence relationship between the keywords covered by the topic, and the relationship between topics is displayed from the fine-grained perspective of time and keywords. (2) The pattern matching of subsegments of co-occurrence time series of different topics is achieved through the time moving window, and the total correlation between topics is reflected by local correlation. Combined with complex network analysis, the analysis and research of topic relations are more comprehensive and in-depth. (3) Topic clustering centers are independent of keyword popularity, and topics with broader concepts are more likely to become topic network centers.

In addition, there are some limitations. (1) The data is not pre-processed in detail. Some meaningless topics cannot provide substantive conclusions, and deleting meaningless topics in advance may obtain higher quality results. (2) Topic analysis approach provides a 5-year time window for the subsequence division of topic co-occurrence, which is referential but not objective enough. (3) When constructing the topic network, the similarity times of the co-occurrence topic time series segments are taken as the network edge weight, which may ignore the influence of the real distance of the optimal similar segments. The determination of the network edge weight by the true value of the similarity of the topic time series segments is a work that needs further research in the future. Moreover, to provide constructive suggestions for scholars in more fields and improve the applicability of conclusions, in the future work, the data scope can be expanded to multiple fields, and the data can be processed in more detail.

Author Contributions: Conceptualization, W.L.; data curation, X.W. (Xianli Wu); formal analysis, Z.W.; funding acquisition, H.L.; investigation, W.L.; project administration, H.L.; resources, W.L.; software, X.W. (Xianli Wu); supervision, X.W. (Xiaoji Wan) and H.L.; validation, W.L. and H.L.; writing—original draft, X.W. (Xianli Wu) and Z.W.; writing—review and editing, W.L. and Z.W. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China (71771094).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The funders had no role in the study design, data collection, analysis and interpretation, manuscript writing, or decision to publish the results.

References

1. Tahamtan, I.; Bornmann, L. What do citation counts measure? An updated review of studies on citations in scientific documents published between 2006 and 2018. *Scientometrics* **2019**, *121*, 1635–1684. [CrossRef]
2. Mokhtarpour, R.; Khasseh, A.A. Twenty-six years of LIS research focus and hot spots, 1990–2016: A co-word analysis. *J. Inf. Sci.* **2021**, *47*, 794–808. [CrossRef]
3. Ji, S.; Pan, S.; Cambria, E.; Marttinen, P.; Philip, S.Y. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Trans. Neural Netw. Learn. Syst.* **2021**, *33*, 494–514. [CrossRef] [PubMed]
4. Jung, H.; Lee, B.G. Research trends in text mining: Semantic network and main path analysis of selected journals. *Expert Syst. Appl.* **2020**, *162*, 113851. [CrossRef]
5. Li, H.; Wu, X. Research on topic discovery and evolution based on time series clustering. *J. China Soc. Sci. Tech. Inf.* **2019**, *38*, 1041–1050. [CrossRef]
6. Zhang, Y.; Lu, J.; Liu, F.; Liu, Q.; Porter, A.; Chen, H.; Zhang, G. Does deep learning help topic extraction? A kernel k-means clustering method with word embedding. *J. Inf.* **2018**, *12*, 1099–1117. [CrossRef]
7. de Wildt, T.E.; Chappin, E.J.; van de Kaa, G.; Herder, P.M. A comprehensive approach to reviewing latent topics addressed by literature across multiple disciplines. *Appl. Energy* **2018**, *228*, 2111–2128. [CrossRef]
8. Chauhan, U.; Shah, A. Topic modeling using latent Dirichlet allocation: A survey. *ACM Comput. Surv. (CSUR)* **2021**, *54*, 1–35. [CrossRef]
9. Jelodar, H.; Wang, Y.; Yuan, C.; Feng, X.; Jiang, X.; Li, Y.; Zhao, L. Latent Dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimed. Tools Appl.* **2019**, *78*, 15169–15211. [CrossRef]
10. Jung, S.; Yoon, W.C. An alternative topic model based on Common Interest Authors for topic evolution analysis. *J. Inf.* **2020**, *14*, 101040. [CrossRef]
11. Young, T.; Hazarika, D.; Poria, S.; Cambria, E. Recent trends in deep learning based natural language processing. *IEEE Comput. Intell. Mag.* **2018**, *13*, 55–75. [CrossRef]
12. Shankar, V.; Parsana, S. An overview and empirical comparison of natural language processing (NLP) models and an introduction to and empirical application of autoencoder models in marketing. *J. Acad. Mark. Sci.* **2022**, 1–27. [CrossRef]
13. Otter, D.W.; Medina, J.R.; Kalita, J.K. A survey of the usages of deep learning for natural language processing. *IEEE Trans. Neural Netw. Learn. Syst.* **2020**, *32*, 604–624. [CrossRef] [PubMed]
14. Kim, H.; Drake, B.; Endert, A.; Park, H. Architext: Interactive hierarchical topic modeling. *IEEE Trans. Vis. Comput. Graph.* **2020**, *27*, 3644–3655. [CrossRef]
15. Zhao, K.; Cong, G.; Li, X. PGeoTopic: A Distributed Solution for Mining Geographical Topic Models. *IEEE Trans. Knowl. Data Eng.* **2022**, *34*, 881–893. [CrossRef]
16. Li, H.; Liu, Z. Multivariate time series clustering based on complex network. *Pattern Recognit.* **2021**, *115*, 107919. [CrossRef]
17. Li, H.; Jia, R.; Wan, X. Time series classification based on complex network. *Expert Syst. Appl.* **2022**, *194*, 116502. [CrossRef]
18. Zhu, H.; Qian, L.; Qin, W.; Wei, J.; Shen, C. Evolution analysis of online topics based on 'word-topic'coupling network. *Scientometrics* **2022**, *127*, 3767–3792. [CrossRef]
19. Wu, F.; Xu, W.; Lin, C.; Zhang, Y. Knowledge Trajectories on Public Crisis Management Research from Massive Literature Text Using Topic-Clustered Evolution Extraction. *Mathematics* **2022**, *10*, 1966. [CrossRef]
20. Zhang, Y.; Shi, Q.; Zhu, J.; Peng, J.; Li, H. Time Series Clustering with Topological and Geometric Mixed Distance. *Mathematics* **2021**, *9*, 1046. [CrossRef]
21. Li, H.; Du, T. Multivariate time-series clustering based on component relationship networks. *Expert Syst. Appl.* **2021**, *173*, 114649. [CrossRef]
22. Frey, B.J.; Dueck, D. Clustering by passing messages between data points. *Science* **2007**, *315*, 972–976. [CrossRef] [PubMed]
23. Chen, J.; Du, S.; Yang, S. Mining and Evolution Analysis of Network Public Opinion Concerns of Stakeholders in Hot Social Events. *Mathematics* **2022**, *10*, 2145. [CrossRef]
24. Blondel, V.D.; Guillaume, J.L.; Lambiotte, R.; Lefebvre, E. Fast unfolding of communities in large networks. *J. Stat. Mech. Theory Exp.* **2008**, *2008*, P10008. [CrossRef]
25. Wan, J.k.; Hua, P.h.; Rousseau, R.; Sun, X.k. The journal download immediacy index (DII): Experiences using a Chinese full-text database. *Scientometrics* **2010**, *82*, 555–566. [CrossRef]
26. Pariwatthanasak, K.; Ratanamahatana, C.A. Time series motif discovery using approximated matrix profile. In *Third International Congress on Information and Communication Technology*; Springer: Berlin/Heidelberg, Germany, 2019; pp. 707–716. [CrossRef]
27. Li, J.; Chen, S.; Zhang, K.; Andrienko, G.; Andrienko, N. COPE: Interactive exploration of co-occurrence patterns in spatial time series. *IEEE Trans. Vis. Comput. Graph.* **2018**, *25*, 2554–2567. [CrossRef]
28. Li, H. Time works well: Dynamic time warping based on time weighting for time series data mining. *Inf. Sci.* **2021**, *547*, 592–608. [CrossRef]
29. Chen, S.; Wang, Z.Z.; Bao, M.H.; Tang, L.; Zhou, J.; Xiang, J.; Li, J.M.; Yi, C.H. Adaptive multi-resolution modularity for detecting communities in networks. *Phys. A Stat. Mech. Its Appl.* **2018**, *491*, 591–603. [CrossRef]
30. Fang, Z.; Wang, J.; Wang, B.; Zhang, J.; Shi, Y. Fuzzy search for multiple Chinese keywords in cloud environment. *Comput. Mater. Contin.* **2019**, *60*, 351–363. [CrossRef]

31. Sancino, A.; Hudson, L. Leadership in, of, and for smart cities–case studies from Europe, America, and Australia. *Public Manag. Rev.* **2020**, *22*, 701–725. [CrossRef]
32. Du, M. Application of information communication network security management and control based on big data technology. *Int. J. Commun. Syst.* **2022**, *35*, e4643. [CrossRef]