ISSN 2616-5775 Vol. 6, Issue 1: 70-76, DOI: 10.25236/AJCIS.2023.060111

Research on Topic Discovery and Evolution Trend Based on Temporal Keyword Characteristics Analysis

Wan Wang

College of Information Engineering, Nanjing University of Finance & Economics, Nanjing, China

Abstract: *In order to mine the research topics in massive articles, sort out the evolution context and correlation relationship of research topics, enhance the scientificity and vividness of evolution results, this paper proposes the concept of temporal influence factor as an important feature in keyword extraction. Using the time window method, the topic model is used to mine and identify topics and perform visual analysis. It is verified that keyword extraction with time series features can improve the effect of topic model. Through visualization, we can not only observe the overall trend of topic popularity, but also analyze the evolution of topic content in each time period and observe the trend of splitting and merging.*

Keywords: *Time Series Keyword, Topic Evolution, Keyword Extraction, Visual Analysis*

1. Introduction

It is of great significance to understand the development of academic research and explore the direction of academic research by analyzing the contents of published articles to detect the topic trend. With the development of science and technology and the popularity of the Internet, the growth rate of the number of documents is increasing. How to quickly and effectively discover research hotspots and frontiers in the face of massive literature data is an important content concerned by library and information science and other disciplines [1].

The evolution trend of research topics is to study how a topic changes over time, whether the attention received increases and which topics become more and more important or disappear gradually. The research on topic discovery and evolution trend can help researchers grasp hot information, promote knowledge transfer within and between fields, help funding institutions and policymakers track innovation and knowledge flow [2].

2. Related Work

2.1 Research Status of Keyword Automatic Extraction

Supervised, unsupervised and semi-supervised are three ways of keyword extraction. Supervised methods can be generally divided into two categories. The first category is to regard keyword extraction as a classification task, such as KEA system [3] and GenEx system [4]. The second category is to regard it as a sequence labeling task, such as using CRF to extract keywords based on the analysis of various features and expert knowledge [5].

Unsupervised keyword extraction methods include statistical-based methods and graph-based methods [6]. The statistical method is mainly based on the characteristics of the document itself to determine the characteristics of the extracted words. The TFIDF method proposed by Spark is the most basic statistical-based method [7]. In addition, the KeyCluster algorithm uses the idea of clustering to obtain representative phrases in key phrase clusters [8]. Some scholars have taken into account contextual information when performing keyword extraction. YAKE algorithm proposed in 2018 is not limited by external corpora, the length of text documents, language or domain [9].

Word graph based methods are more common in the field of keyword extraction. TextRank method based on PageRank algorithm was first proposed [10]. RAKE is also a method based on word graph [11]. The main difference between RAKE and TextRank is that RAKE considers the co-occurrence of

ISSN 2616-5775 Vol. 6, Issue 1: 70-76, DOI: 10.25236/AJCIS.2023.060111

candidate keywords rather than fixed windows and uses a simpler and more statistical scoring program. In order to solve the possible lack of information in a single document extraction, some algorithms with external resources have emerged to enhance the amount of information in the word graph, such as ExpandRank, CiteTextRank, WordAttractionRank. Other methods are based on multi-graph methods.

2.2 Research Status of Topic Evolution Trend

Documents in different periods can form a time flow sequence and the subject content of documents in different periods is also changing. In order to capture this changing feature, there are various researches on integrating timestamp and subject model. Some scholars [12] divided these into three types from the introduction order of document time: using time as an internal variable of the topic model, running the topic model first and then dispersing it into the time window, dividing the time period first and then running the topic model. TOT model is a method that takes time as an internal variable [13]. Each topic is related to a continuous timestamp distribution, which better fits the dynamic evolution process of topics. The second method is to run the topic model first and then discretize the time. Griffths proposed this method [14]. Before discretizing to each time window, first obtain all the topics on the whole text set using the LDA topic model. In contrast, the method of discretizing to the time window first and then running the topic model has greater flexibility. For example, Blei proposed a dynamic topic model DTM [15] based on the LDA model, also relevant is OLDA .

In the study of the visualization of the topic evolution path, Wei Ling used the combination of the co-word network community and the alluvial map [16]. Gao Nan constructed the frontier technology evolution map through the technology category Condorcet cross-contrast matrix to judge the trend of the topic [17]. Hou Jianhua predicted the research frontier in the field of big data by drawing a knowledge map of literature co-citation and citation structure transformation [18].

2.3 Contribution

In the process of keyword extraction, the words appearing in the same document are composed of co-occurrence word pairs. The weight of the co-occurrence word pairs closer to the current time node is more meaningful and should be given a greater weight. The intensity of this temporal influence is called the temporal influence factor and the edge weight in the graph is corrected by the temporal influence factor. In the face of the problem that the keywords in different time periods are not completely consistent, the method used in this paper is to add the inconsistent keywords to the corresponding topics and set the probability of semantic similarity. The topic popularity of a topic over a period of time is calculated by summarizing the topic distribution of each document annually and topic co-occurrence network can describe the degree of correlation between topics. Sankey Chart and other charts are used to describe the state changes of the split and merge of local topics and the closeness of topic in various time periods. The main framework of this paper is shown in Figure 1.

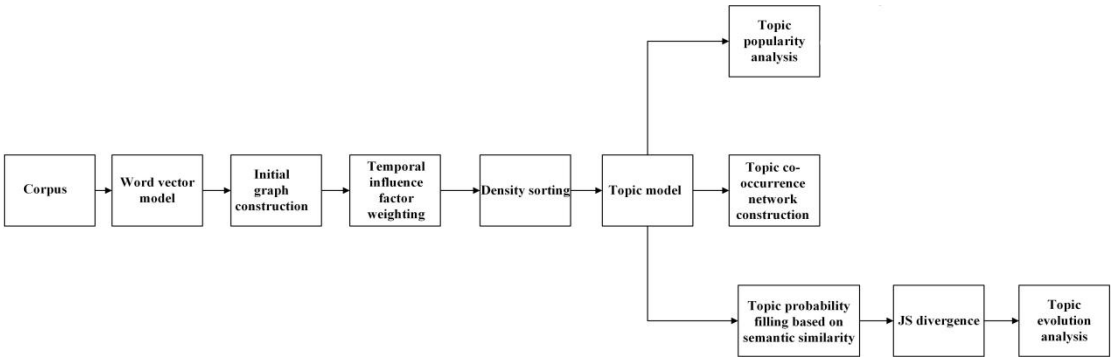

Figure 1: Research framework.

3. Keyword Weight Acquisition

3.1 Construction of Keyword Graph

Get the document data and then get the word set $W=\{w_1/s_1,w_2/s_2,w_3/s_3,...,w_n/s_n\}$ after word segmentation and part-of-speech labeling, where s_i is the part-of-speech tagging of the word w_i;

ISSN 2616-5775 Vol. 6, Issue 1: 70-76, DOI: 10.25236/AJCIS.2023.060111

Secondly, filter W, only retain the words whose part of speech is noun and record the filtered word segmentation set as W_1.

Take each word in the set W_1 as vertex V and add a connecting edge between all vertices, that is, E =V × V. The initial edge weight is set to the cosine similarity of adjacent vertex word vectors. When vectorizing words, the Skip-gram word vector model is used. The Skip-gram model establishes a multi-dimensional word vector for each word, and further calculates the cosine similarity between words according to the word vector and sets it as the weight of the edge. If the edge weight is less than 0, it is pruned to form the initial graph G= (V, E), then the initial adjacency matrix A_0 is obtained.

3.2 Edge Weight Design Based on Weighted Temporal Influence Factor

This paper defines two words appearing in the same document to form a co-occurrence word pair (A, B) and counts the frequency of the co-occurrence word pairs in each time period as the weight, as shown in Formula (1).

$$\text{weight}(w_1, w_2, t_i) = | w_1 \cap w_2 |_{t_i} \tag{1}$$

Reduce the weight of the co-occurrence word pairs with a larger span from the current time node, emphasize the greater importance of the new co-occurrence word pairs and define the strength of this temporal relationship as the temporal influence factor, as shown in Formula (2). Here, t is the current time node, N is the normalized function and the value is limited to 0 to 1. The greater the value of the timing influence factor, the stronger the temporal relationship between words.

$$\text{TimeFac}(w_1,w_2) = N(\sum (e^{-(t-t_i)} \text{weight}(w_1,w_2,t_i))) \tag{2}$$

The value of each element in the initial adjacency matrix A is the cosine similarity between word vectors. The adjacency matrix A_0 can be obtained by modifying the value in adjacency matrix A with the time series influence factor, as shown in formula (3), where A_{ij} represents the value of the element in row i and column j of adjacency matrix A, that is, the modified weight of w_i and w_j.

$$A_{ij} = A_{0_{ij}} \text{TimeFac}(w_i,w_j) \tag{3}$$

3.3 Keyword Weight Measure Based on Density Ranking

The node density is calculated by the transfer probability of nodes. Nodes with larger total transfer probability have higher density. Based on this assumption, this paper extracts keywords. Let $A^l_{wj,wj}$ be the probability that word w_i is transferred to word w_j after step l, then the density at the specified step size l is set to Equations (4) and (5).

$$\text{density}(w_j) = \sum_{i=0,i\neq j} A^l_{w_i,w_j} \tag{4}$$

$$A^l_{w_i,w_j} = \sum_{s=0}^{n} A^{l-1}_{w_i,w_s} A_{w_s,w_j} \tag{5}$$

Here, $A_{ws,wj}$ is the initial adjacency matrix A and n is the number of nodes. The initialization step l of the algorithm is 1, then the density of each node is calculated circularly and arranged in descending order. If the density sorting of nodes is the same as the last time in a cycle, the cycle stops. After the density of all vertices is normalized, the words whose sorting is greater than the average value are the keywords of the document. Otherwise, the step size is increased in the cycle to continue the operation.

4. Topic Evolution Trend Analysis

4.1 Topic Relevance Analysis

The method of correlation analysis adopted in this paper is to build a topic co-occurrence network. With the topic as the node and the topic co-occurrence strength as the edge weight, an undirected weighted network G(V, E, W) is constructed, where V represents the collection of topic nodes, E represents the collection of associated topic edges and W represents the co-occurrence strength between topics. According to the Pareto rule, the topic with a cumulative probability of 0.8 can represent the topic feature of the document, so the number of selected topics is obtained. Defines the strength

ISSN 2616-5775 Vol. 6, Issue 1: 70-76, DOI: 10.25236/AJCIS.2023.060111

between two arbitrarily related topics globally as shown in Equation 6. Here, n is the number of documents that are co-existing with topic i and topic j, $P(T_{mi})$ represents the probability of topic i in the mth co-occurring document and $P(T_{mj})$ represents the probability of topic j in the mth co-existing document.

$$\text{Relation}_{ij} = \sum_{m=1}^{n} P(T_{mi})P(T_{mj}) \tag{6}$$

4.2 Topic Popularity Evolution Analysis

Research topic evolution includes how research topics change over time and whether they become increasingly important or disappear. To study this change, define a topic popularity index to indicate the popularity of a topic, which is represented by its annual proportion. The higher the proportion, the more popular the topic is. Based on time series analysis theory, variables with time series correlation can be predicted, so topic popularity indicators can be further predicted. This article defines the popularity of topic j in year i as Equation (7), where P_{mj} represents the probability distribution of the mth document on the jth topic.

$$\text{Popularity}_{ij} = \frac{\sum_{m\square \text{ yeari}} p_{mj}}{\sum_{j}\sum_{m\square \text{ yeari}} p_{mj}} \tag{7}$$

4.3 Topic Content Evolution

The analysis of topic evolution includes the splitting and merging of topics. If a topic is a combination of multiple topics, make the multiple topics its precursor; If a topic can be split into multiple topics, the split topic is called a subsequent topic. We get the keyword set $T_i=\{v_1,v_2,...v_n\}$ for a certain period of time, $P_i=\{p_1,p_2,...,p_n\}$ for the probability distribution matrix of topic words and incomplete overlap of topic words in adjacent periods. The method used in this paper is to add the non-overlapping topic word v to the non-occurring topic T and set the probability of the topic word as shown in Formula (8), so that the vocabulary contained in the two topics is identical. Here, n is the number of original topic words in T and $\text{cosSimilar}(v,v_i)$ is the cosine similarity of the two topic words.

$$p_T(v) = (\sum_{i=1}^{n} p_T(v_i)\text{cosSimilar}(v, v_i)) / n \tag{8}$$

5. Experiments

5.1 Datasets

The experimental dataset comes from the “Science Network”, with a total of 23318 initial data. The time span is from 2007 to 2021, including 37 types of data. The basic statistics of the data, depicting the annual trend of the number of papers is shown in Figure 2, which is divided into four time intervals: germination period, rapid growth period, peak period and fluctuation period. The data information of the four time periods is shown in Table 1.

Table 1: Dataset information.

Dataset number	Dataset 1	Dataset 2	Dataset 3	Dataset 4
Time interval (year)	2007-2009	2010-2012	2013-2016	2017-2021
Number of data	814	5511	9329	7664
Number of categories	22	32	36	33

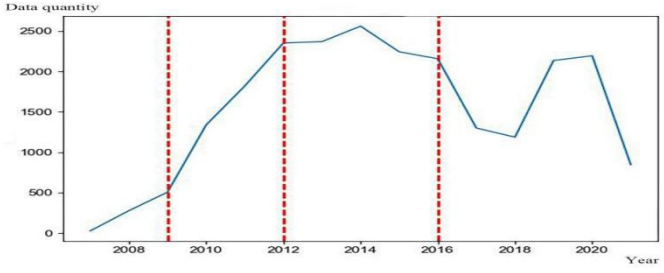

Figure 2: Annual change trend of data quantity.

ISSN 2616-5775 Vol. 6, Issue 1: 70-76, DOI: 10.25236/AJCIS.2023.060111

5.2 Results Analysis

5.2.1 Model Comparison

Topic consistency evaluation: The consistency score of the LDA model measures the semantic similarity between words in each topic. When all other conditions are the same, the higher the consistency score, the better, because this indicates that the meaning of the words in each topic has a high similarity. The results of the comparative experiment under different topic numbers are shown in Figure 3. It can be observed that the graph screening LDA model with temporal influence factor has higher topic consistency.

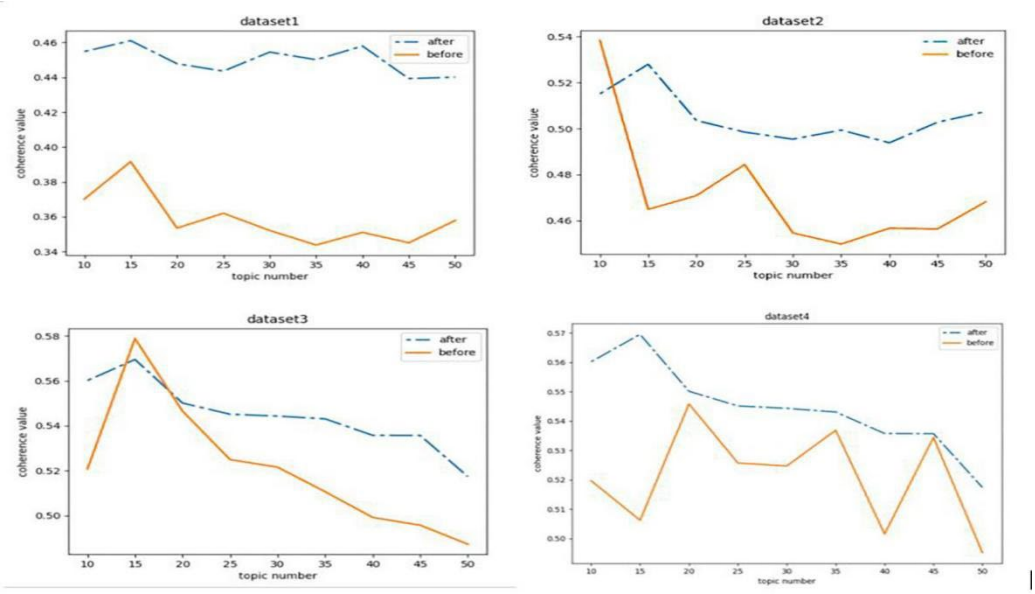

Figure 3: Topic consistency comparison.

Cluster effect evaluation: Set the number of topics to the number of categories K of the dataset, and use the topic with the highest probability in the document topic distribution obtained as the document aggregation class. Use ACC to evaluate the clustering effect, as shown in Equation (9).

$$ACC = \frac{\sum_{d=1}^{D} \delta(y_d, map(c_d))}{D} \tag{9}$$

The real label set and the estimated cluster label set are represented by Y and C respectively. For each document d, its true label and estimated cluster label are represented by y_d and c_d respectively. δ is an indicator function and map(c_d) is a mapping function obtained using the Hungarian algorithm. The experimental results are shown in Table 2.

Table 2: ACC evaluation.

	Dataset 1	Dataset 1	Dataset 1	Dataset 1
Standard LDA	0.213	0.214	0.202	0.238
Graph screening LDA fusing temporal influence factor	0.240	0.254	0.271	0.257

5.2.2 Visual Analysis

Topic co-occurrence network: This paper randomly selects 100 documents and uses Origin to draw the cumulative probability map of the literature topic. According to the Pareto rule, the cumulative number of topics corresponding to 0.8 on the fitting curve is 4.66 and the integer is 5. Therefore, the five topics with the highest probability of each document are selected to construct the topic co-occurrence network. The visualization results are shown in Figure 4, where the node sizes are arranged according to degree centrality and the weight of the edge indicates the weight of the topic co-occurrence.

ISSN 2616-5775 Vol. 6, Issue 1: 70-76, DOI: 10.25236/AJCIS.2023.060111

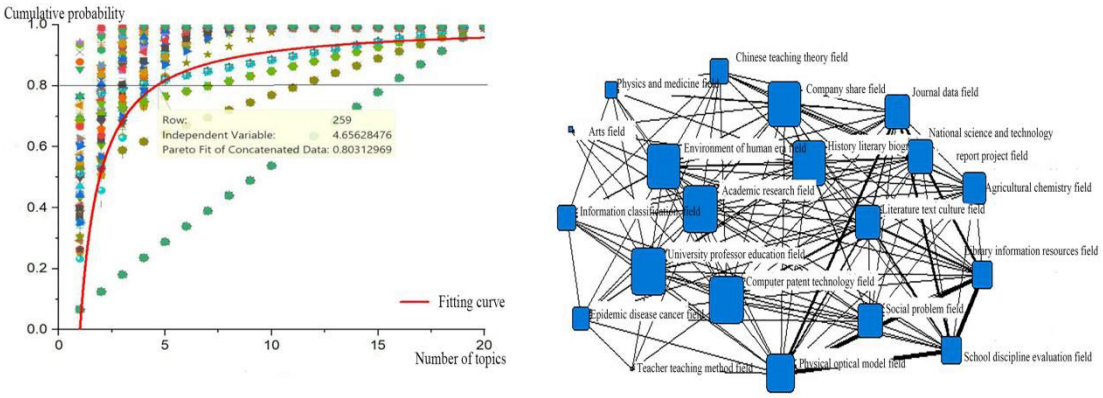

Figure 4: Cumulative distribution of document topic probability and topic co-occurrence network.

Analysis of topic popularity evolution trend: topic popularity is expressed by its annual proportion as shown in Figure 5. Topic popularity is expressed by height. The higher the height, the more popular the topic is. It can be observed that: In each year, the topics numbered 1-3 are popular, while the topics numbered 11-20 have received less attention. Topic numbers 1-4 have accounted for a large proportion since 2007, but these topics showed a significant downward trend in 2018. Some topics that were less popular in the past increased sharply, such as topic numbers 6-11. After that, they gradually returned to a flat state, and the attention of the first few topics recovered.

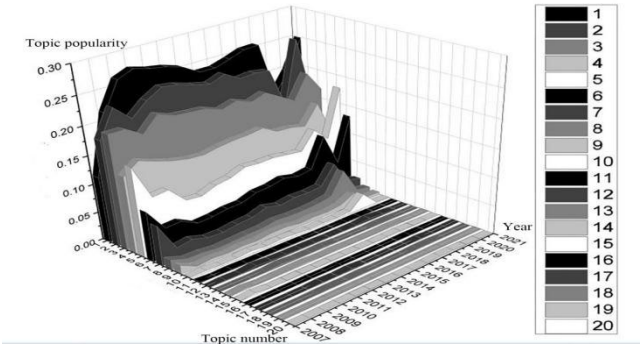

Figure 5: Annual change of topic popularity.

Topic content evolution analysis: A Sankey Chart of topic evolution is drawn to describe the trend of splitting and merging between topics, Set the threshold to filter the threshold with low topic similarity and use the five words with the highest probability of each topic as the content label of the topic. Drawing the Sankey Chart to show the evolution trend of topic content can not only observe the change of topic with time flow, but also trace its source. For example, select topic 8 (content label: intelligence - author - informatics - theory - science) in the rapid growth period to draw the evolution trend as shown in Figure 6.

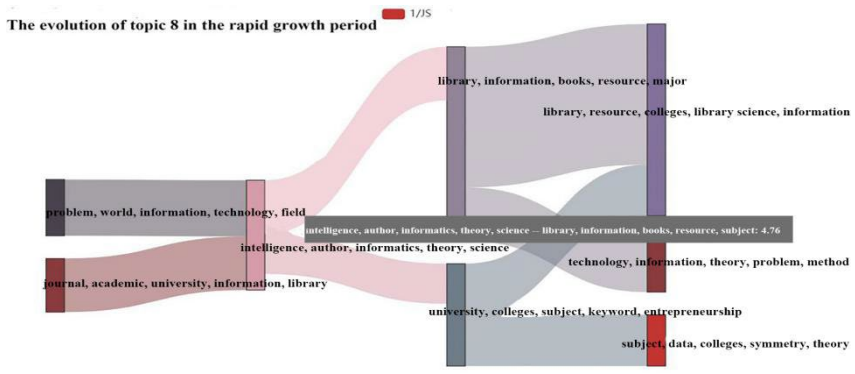

Figure 6: Content evolution display of topics in different time periods.

ISSN 2616-5775 Vol. 6, Issue 1: 70-76, DOI: 10.25236/AJCIS.2023.060111

6. Discussion

The existing keyword extraction lacks the measurement of the temporal relationship of words. This paper proposes the concept of temporal influence factor as an important feature in keyword extraction. To sum up, keyword extraction combined with temporal keyword features can improve the effect of the topic model and better conduct subsequent topic evolution analysis. This research mine and identify research topics under multiple time series windows. When measuring topic similarity, it pays attention to combining semantic characteristics and polynomial distribution characteristics between words and fully considers multiple visualization methods for display.

This paper constructs a topic co-occurrence network in the visualization link, which can measure the degree of connection between the overall topics; By calculating the topic popularity, hot topics can be found, which is helpful for scholars to grasp the changing rules of research hot spots; The Sankey diagram is used to represent the split and merge of the topic content of each time period. Compared with the single visualization method in the current research, this paper is more comprehensive, scientific and understandable.

References

[1] Wang K, Gao JP, Pan YT, Chen Y. (2021) Research on Multi-position Research Topic Recognition and Evolutionary Path Method. Library and Information Work. 65(11), 113-122.
[2] Chen, B., Tsutsui, S., Ding, Y., Ma, F. (2017) Understanding the topic evolution in a scientific domain: an exploratory study for the field of information retrieval. Journal of Informetrics. 11(4), 1175-1189.
[3] FRANK E, PAYNTER G W, WITTEN I H. (1999) Domain-specific keyphrase extraction. International Joint Conference on Artificial Intelligence. 2, 668-673.
[4] Turney, P. D. (2000) Learning algorithms for keyphrase extraction. Information Retrieval. 2(4), 303-336.
[5] Gollapalli, S. D., Li, X., Peng, Y. (2017) Incorporating expert knowledge into keyphrase extraction. National Conference on Artificial Intelligence. 3180-3187.
[6] Papagiannopoulou, E. T. G. (2020) A Review of Keyphrase Extraction. Wiley interdisciplinary reviews. Data mining and knowledge discovery, 10(2).
[7] Jones, K. S. (2004) A statistical interpretation of term specificity and its application in retrieval. Journal of Documentation, 60(5), 493-502.
[8] Liu, Z., Peng, L., Zheng, Y., Sun, M.. (2009) Clustering to find exemplar terms for keyphrase extraction. Conference on Empirical Methods in Natural Language Processing. 257-266.
[9] Pasi, G., Piwowarski, B., Azzopardi, L., Hanbury, A.. (2018) A text feature based automatic keyword extraction method for single documents. European conference on information retrieval. 63, 684-691.
[10] Mihalcea, R., Tarau, P.. (2004) Textrank: bringing order into texts. Proceedings of 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona, Spain. 404-411.
[11] Danesh, S., Sumner, T., Martin, J. H.. (2015) SGRank: Combining Statistical and Graphical Methods to Improve the State of the Art in Unsupervised Keyphrase Extraction. Proceedings of the Fourth Joint Conference on Lexical and Computational Semantics. 117-126.
[12] Shan B, Li F. (2010) Overview of research methods of topic evolution based on LDA. Journal of Chinese Information Processing. 24(6), 43-50.
[13] Wang, X., Mccallum, A.. (2006) Topics over time: a non-Markov continuous-time model of topical trends. Acm Sigkdd International Conference on Knowledge Discovery & Data Mining. ACM.
[14] Griffiths, T. L., Steyvers, M.. (2004) Finding Scientific Topics. Proceedings of the National Academy of Sciences of the United States of America. 101(S1), 5228-5235.
[15] Blei, D. M., Lafferty, J. D.. (2006) Dynamic topic models. Machine Learning, Proceedings of the Twenty-Third International Conference (ICML 2006), Pittsburgh, Pennsylvania, USA, June 25-29, 2006.
[16] Kui L, Xu HY, Hu ZY. (2016) Multi-pattern recognition and prediction of subject evolution path. Library And Information Service. 60(13), 71-81.
[17] Gao N, Peng DY, Fu JY. (2020) Analysis of Advanced Technological Evolution Based on Patent IPC Classification and Text Information--A Case Study in the Field of Artificial Intelligence. Information studies: Theory & Application. 43(4), 123-129.
[18] Hou JH, Li LJ, Yang XC. (2018) Frontier Prediction of Large Data Research Based on the Transformation of Citation Network Structure. Information Science. 36(6), 142-148.