Detecting emerging topics by exploiting probability burst and association rule mining: A case study of Library and Information Science

Min Xu, Guangjian Li* and Xiaodi Wang
Department of Information Management,
Peking University, Beijing 100871, CHINA
e-mail: kingfoxing@pku.edu.cn; *ligj@pku.edu.cn (corresponding author);
wangxiaodi@pku.edu.cn

ABSTRACT

The primary reason for detecting emerging topics is to reduce researchers' time in finding current related topic while maintaining awareness of current trends in a particular field. Nowadays, the amount of information is growing rapidly, but tracking the development of a research field by manually reading the literature is challenging. This study takes Library and Information Science (LIS) as a case study to present a new method for detecting emerging topics. This novel method could be applied to analyse various types of documents and detect emerging topics automatically. This method utilizes a Latent Dirichlet Allocation (LDA) model to generate topics and calculate probabilities. It discovers emerging topics by detecting probability burst in consecutive time spans. Association rule mining and lexical similarity computation are adopted to represent the topics. This work tests the method by comparing the results of emerging topics from the LIS data in the baseline paper. The validation demonstrates that the proposed approach is feasible.

Keywords: Latent Dirichlet Allocation; Emerging topic detection; Probability burst; Association rule mining; Library and Information Science research.

INTRODUCTION

With the development of science and technology, the conflict between the rapidly increasing amount of literature and the urgent demands for detecting emerging topics has intensified. It is important for scholars and researchers to track and detect the emergence of topics in a given research field. The reliance on manual reading and analysis is no longer applicable since several models could be used for topic detection. In 1996, the Defense Advanced Research Projects Agency (DARPA) proposed using a machine for the mining and analysis of text content. To this end, DARPA sponsored a scheme - Topic Detection and Tracking (TDT) - to investigate the state of the art in finding and following new events in a stream of broadcasted news stories (Allan et al. 1998). The objective of the TDT programme is to develop technologies for searching, organising and structuring multilingual, news-oriented textual materials from a variety of broadcast news media (Fiscus and Doddington 2002). Research on TDT was abundant and increased in 1998 and 1999 (Wayne 2000). In 2004, the TDT2004 project defined five main tasks for TDT: topic tracking, supervised adaptive tracking, hierarchical topic detection, new event detection,

and link detection. The TDT new event detection task aims to detect an event's first report in a chronologically ordered stream of information from multiple sources and in multiple languages (The NIST 2004). This study, which focuses on emerging topic detection, is an extension of TDT new event detection, and the approach can detect not only previously unseen topics but also rejuvenated topics. It defines a topic as a subject that is discussed or focused on by most authors. This work defines an emerging topic as a new finding that most people pay attention to or an existing subject that re-attracts public interest.

LITERATURE REVIEW

Emerging topic detection has become a popular research subject. There have been many studies to detect emerging topics in various research domain which aim to reduce researchers' time in finding current related topic while maintaining awareness of current trends in a particular field. The mainstream research relating to emerging topic detection can be grouped into three categories: emerging topic detection based on citation network analysis, emerging topic detection based on keyword analysis, and emerging topic detection based on topic models.

Based on Citation Network Analysis

Persson (1994) detected research fronts by using bibliographic coupling and co-citation network analysis. Morris et al. (2003) considered bibliographic coupling as a tool to detect and represent research fronts. Shibata et al. (2008) developed a citation network analysis method for emerging research front detection that performed topological clustering, tracked the positions of the papers in each cluster, and visualised the citation networks in terms of the characteristic words in each cluster. Their approach includes three phases: clustering, extracting the role of each paper, and topic detection by natural language processing. Fujimagari and Fujita (2014) proposed an approach for research front detection that combines weighted citation network analysis with a neural network model. Their method includes the following phases: creating weighted citation networks, topological clustering, and topological measurements to evaluate the citation networks. Shibata et al. (2009), Boyack and Klavans (2010) and Fujita et al. (2012) compared the performance of direct citation networks, co-citation networks and bibliographic coupling for detecting emerging research fronts based on the following indicators of clusters: visibility (measured by the normalised cluster size), speed (measured by the average publication year), and topological relevance (measured by the density). These researchers found that direct citation network analysis, which exhibited the best performance in detecting a research front, could detect large and young emerging clusters, whereas co-citation network analysis performed the worst. However, when applying these methods in science mapping, they found that bibliographic coupling slightly outperformed co-citation network analysis and that direct citation network analysis was the least-accurate science mapping approach. They also determined that the hybrid approach could effectively improve the bibliographic coupling results.

Based on Keyword Analysis

Xie (2014) established a co-word matrix based on high-frequency keywords from the China Academic Journal Network Publishing Database (CAJD) to detect hot topics in the research field of ecological/environmental risk. Kleinberg (2003) developed a burst detection algorithm to identify explosive term to analyse the text's underlying content whose ideas strongly influenced topic detection researches. Wang, Liu and Cheng (2014) and Hu and Zhang (2015) utilised co-word analysis to detect the emerging themes, research patterns

and research trends of a specific field in China. To this end, these researchers built a highly frequent words matrix, correlation matrix and dissimilarity matrix to analyse keywords from Chinese journals. Cobo et al. (2011) proposed a bibliometric method to observe the evolution of topics in a specific research field. Co-word analysis was the main tool utilised in their method to detect the various topics in a given research field and time span. Their approach includes four phases: theme detection, theme and thematic network visualisation, thematic evolution, and performance evaluation. Li and Chu (2017) proposed an enhanced co-word analysis method to detect research fronts that combined co-word analysis with burst term analysis and determined the relationship between keywords and burst words based on association rule mining. Ding, Chowdhury and Foo (2001) and Hu et al. (2013) employed co-word analysis to observe the pattern of a specific research field. Ding, Chowdhury and Foo (2001) manually extracted important keywords from titles and abstracts and added some keywords from the Science Citation Index (SCI) and Social Sciences Citation Index (SSCI) databases to perform co-word analysis and discover the trends and patterns of the Information Retrieval field. Hu et al. (2013) also employed co-word analysis to discover the intellectual structure of Library and Information Science. They extracted keywords from related documents in the Chinese Journal Full-Text Database and measured their correlation coefficients. Xu et al. (2016) put forward a measure index (TI value) to keywords for detecting interdisciplinary topics, which combines term frequency counting, TF-IDF and citation networks partitioning to identify the emerging topics.

Based on a Topic Model

Since it was originally proposed, the topic model has become a dominant method used to detect and track topics. Latent Dirichlet allocation (LDA) (Blei, Ng and Jordan 2003) is the best-performing Bayesian topic model used in modelling text and detecting topics. Many studies have investigated method of improving the LDA model. Additionally, some studies aimed to establish various types of LDA models for emerging topic detection. Blei and colleagues (Blei 2011; Blei and Lafferty 2005; 2006; Blei and Mcauliffe 2010; Boydgraber and Blei 2008; Chong et al. 2009) proposed a series of topic models, such as Correlated Topic Models, Dynamic Topic models, Markov Topic Models, Supervised Topic Models, and Syntactic Topic Models. Guo et al. (2013) and Alsumait, Barbará and Domeniconi (2008) developed an online topic detection method for detecting emerging topics in document streams according to an incremental pattern. Huang et al. (2017) used local weighted linear regression and the LDA model to evaluate the words and topics in documents. They detected an emerging topic based on the novelty index and fading probability of words and topics. Zhang et al. (2012) proposed a time-varying hot topic propagation model to find trends in hot online topics by gathering information on online users' behaviour. Rosen-Zvi et al. (2004) proposed a document generative model, the Author-Topic Model (ATM), combining the LDA model with authorship information. Bolelli, Erteki and Giles (2010) developed the Segmented ATM (S-ATM), which is based on LDA and is an extension of the ATM that detects topics and trends by adding a time-segmentation function to the ATM. Wu et al. (2014) proposed an approach that combined the LDA model, Hidden Markov Model and co-occurrence theory to segment topics and observe their evolution. Miao et al. (2016) developed a cost-effective system framework with algorithms for detecting the topic trends of Internet microblogs in real time.

By investigating the existing studies, this work found the following:

(a) To use a citation network-based method for topic detection, a complete citation network data set is required. However, some documents, such as consulting reports, news, blogs, and web pages, have no or incomplete citation networks. In other words,

these methods can be adapted to only a limited number of types of documents. Furthermore, in a citation network, the value of a document's citation rate is related not only to the quality of its content but also to its time of publication; therefore, some relatively "old" documents are often highly cited in the citation network, whereas more recent documents could have low citation rates. As a result, the timeliness of the topic detection cannot be guaranteed, and thus, emerging topics cannot be found using this method.

(b) Not all documents have keywords, and thus, keyword-based emerging topic detection is not always plausible. Consequently, the problem associated with having a limitation on the document types that can be analysed remains unsolved. In addition, because authors primarily choose keywords, large numbers of homonyms, synonyms, super-ordinates and hyponyms exist. To address this problem, manual processing, such as removing duplicate words and merging similar words, is required. Furthermore, both the statistics of the word frequencies and the use of co-word analysis are based on high-frequency words; however, "high frequency" has no unified standard definition. Instead, it is defined subjectively by each analyst, and thus, the results vary from one document to another.

(c) To use topic model-based methods, it must consider several additional issues. On the one hand, if the words generated by the LDA model are used to represent the detected topics, then a wealth of prior professional background knowledge in related fields is required for the analysis. However, some topics are described by a variety of isolated words that often cannot be recognised, even by experts in related fields. On the other hand, many manual tasks (e.g., reading documents and choosing keywords) are used to interpret the topic words from the LDA model. In this regard, the reader's prior knowledge plays a major role in the process of analysing, summarising and tagging words. Indeed, the results of topic representation often vary from one person to another. Moreover, manual reading and manual labelling are time intensive. As a result, analysts may struggle to complete these tasks, especially when the collection of documents is massive and/or the number of topics is large.

METHOD

To address the problems as mentioned before, this paper proposes an emerging topic detection approach named Emerging Topic Detection Based on the LDA model and Association rule mining (ETDBLA), in which topic finding is based on probability burst, and topic representation is based on association rule mining. The former is used to overcome the limitation on literature types and find emerging topics automatically; the latter is used to recognise what the topic refers to.

This approach utilises only the four most basic attributions of a document: category, time, title and content. Most published documents have these attributions, and therefore, this approach can be used to detect emerging topics in cases where multi-source documents are mixed. Of these attributions, "Category" can limit the topic's scope. "Time" is used to divide the text set into several subsets according to a certain time span; "Title" is a refined representation of a paper, which directly reflects what the paper is about and can be used in the analysis to represent the emerging topics; and "Content" (i.e., full text or abstract) is an attribution that contains the maximum amount of information. Indeed, "Content" comprises the main idea and all of the details of the article, and thus, this attribution can be used to discover connections among documents and cluster related papers on the same topic.

The approach can be divided into four phases: document collection slicing, topic generation, emerging topic detection and topic representation. Figure 1 illustrates the overall process of the ETDBLA approach.

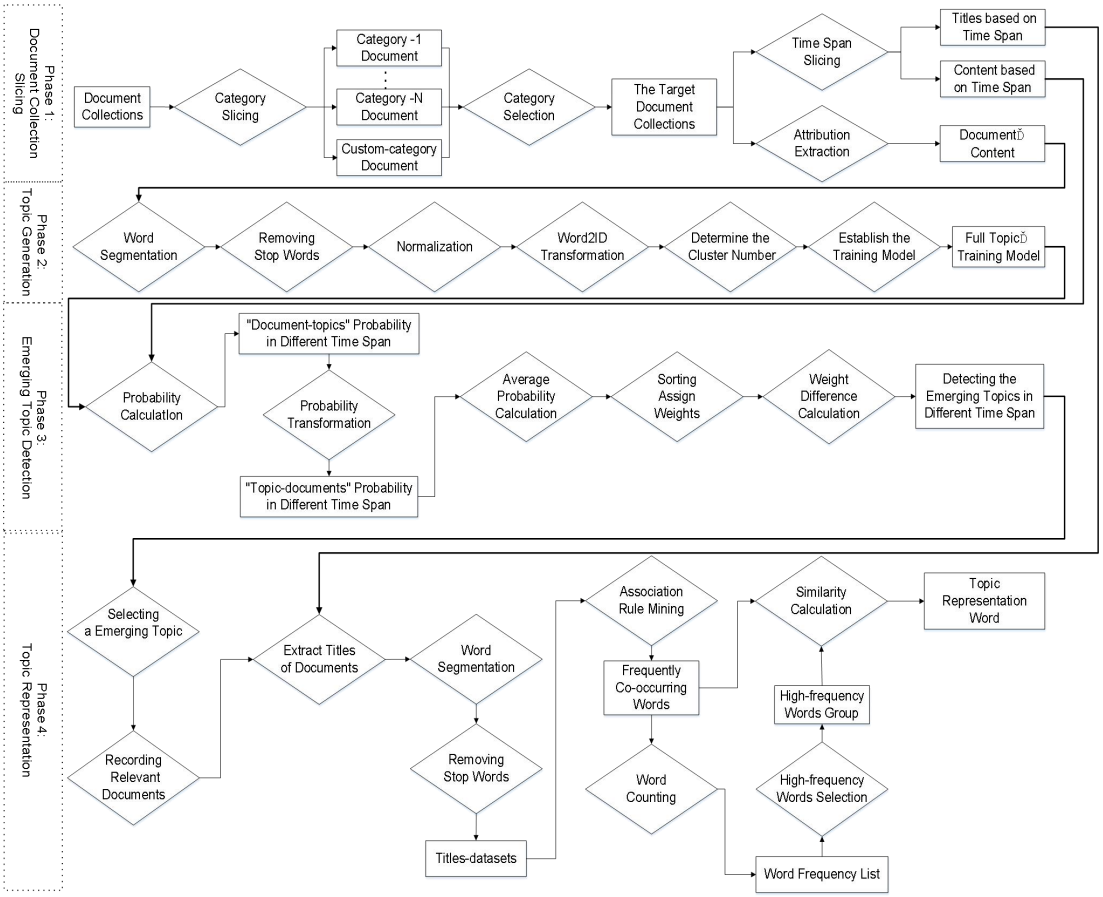

Figure 1: The Overall Process of the ETDBLA Approach

Document Collection Slicing

The main task in this phase is document collection slicing and attribution extraction. Category slicing can limit the scope of the topics. According to the demands of the analysis, a new data set that is relevant to the selected fields can be extracted from the entire document collections. This process ensures that the documents are highly related to certain fields. In other words, to analyse an interdisciplinary topic, it must merge the multi-category documents to ensure the accuracy of the results. For example, if researchers aim to identify the emerging topics relating to machine learning, they must gather documents that are related to linear mathematics, probability theory, statistics, and algorithms. Indeed, to analyse a specialised topic, they must focus on the specific field to avoid any ambiguity arising from interdisciplinary terminology. For example, the documents on “Apple” have different meanings in the fields of mobile device and agriculture.

Time slicing can embody the changing of a topic’s features. According to a certain time span (e.g., year, month, or day), the data set is divided into a number of sub-data-sets. By

comparing the same topic in different time spans, it can find valuable information. Thus, time slicing is the basis for emerging topic detection. Title and content extractions are performed to generate a feature set of documents. The title and content (i.e., abstract or full text) can be extracted from each document. Then, the title can serve as the data for association rule mining in topic representation, and the content can serve as the data for clustering and topic generation. This work constitutes the basis of document comparison, document clustering and association rule mining.

Topic Generation

The topic generation method is as follows: First, a topic model is established based on the entire set of documents related to a specific field, and then, a trained model is applied to infer topic distributions for the sub-document collections.

Above all, word segmentation, stop word removal, and word normalisation are necessary. Word segmentation is the discretisation of the document and converts the text to a sequence of words. Stop word removal is performed to reduce the dimensionality of the documents and improve the computing efficiency. Word normalisation is utilised to improve the accuracy of the relationship calculations during the clustering. First, in the word segmentation of the document content (i.e., abstract or full text), white space can be seen as a word segmentation delimiter if the text is written in a Latin language; if the text is Chinese, certain open-source word segmentation tools (e.g., Institute of Computing Technology, Chinese Lexical Analysis System [ICTCLAS], jieba, and IKAnalyzer) can be employed. Second, stop word removal is performed to remove functional words and special symbols that have no actual meaning. Third, a special linguistic phenomenon in Latin languages must be acknowledged: Two words derived from the same root can have different meanings. Therefore, the function of word normalisation is to merge words with the same root and the same meaning and to distinguish between words with the same root but different meanings. For example, in English, "run" and "ran" are words that have the same root and the same meaning. These two words should be merged into one word. In contrast, "confident" and "confidential" have the same root but different meanings, and thus, these words should be distinguished. The merging of words should be based on the relationships between the words; namely, a word relationship dictionary, such as WordNet, can be used.

The word2ID transformation makes documents and words computable. All of the processed words in the documents must be mapped into the vector space to create a word distribution matrix. The feature vectors of the documents lie in a high-dimensional space, which means that a vector will have a large number of elements; however, for a single document, most of the values of the vector elements are zero. Thus, sparse storage is commonly used for the vectors of documents.

The primarily aim of clustering is to gather documents that are strongly or closely related. Subsequently, different categories of documents can be established by determining the numbers of clusters. Regarding document clustering and topic generation, selecting the LDA topic model is recommended. As one of the best techniques for TDT research, the LDA model can automatically cluster documents on the semantic level and generate a set of topics composed of related words and relevant probabilities. In an LDA analysis, the documents are considered to represent a mixture of topics. For each topic, there is a distribution over the terms of a fixed vocabulary (Blei, Ng and Jordan 2003). Therefore, in each category, all of the documents are related to a topic, but the value of the relevant probability varies from document to document. Within the document clustering results, a

document can be found in several subsets. For example, Document-A could appear in category M, category N, category L, and several others. However, the relevant probability values of Document-A in different categories are different. This is because a document usually addresses several topics rather than a single one. Clustering allows discovering what the authors of the document discussed.

Using the LDA model, the value of the probability between the words and topics within a document and that between topics and the documents themselves, is calculated as follows:

$$p(word|document) = \sum_{n=1}^{topic} p(word|topic) \times p(topic|document) \quad (1)$$

The value of p(word|topic) is the probability between the words and a topic in a document, and the value of p(topic|document) is the probability between a topic and a document. Using the value of p(word|document), the similarity between documents can be determined. The value of p(word|document) to find the semantic relationship among documents can be used and, thereby, cluster the documents.

Emerging Topic Detection

In this phase, the approach aims to identify the topics that had a probability burst during the selected time span. The purpose of the probability calculation is to obtain the value of p(topic|document) using the LDA model. Above all, the topic distribution of each text is computed by using the learnt/trained LDA model, which can discover how many topics each paper is associated with and the relevant probabilities. Thus, it can obtain a result such as "<document, relevant topics, relevant probabilities>"; namely, it can determine the sum value of p(topic|document). However, the core of this approach is the detection of the topic rather than the document, and thus, probability transformation is very important, which converts the previous structure to "<topic, relevant documents, relevant probabilities>", namely, the sum value of p(document|topic).

The probability sum and average probability are measures of the importance of the topic. Using the structure "<topic, relevant documents, relevant probabilities>" and the sum value of p(document|topic), the number of documents each topic is associated with and the relevant probabilities is known. Then, the probability sum, p(total), and the average probability, p(average), of the topic can be calculated as follows:

$$p(\text{total}) = \sum_{n=1}^{document} (p(document|topic)) \quad (2)$$

$$p(\text{average}) = \sum_{n=1}^{document} \frac{(p(document|topic))}{\text{document}} \quad (3)$$

The probability sum can reflect the ratio of a topic in the document collections. Interestingly, if a topic is related to an excessive number of documents and has a high probability value, it may have no real meaning or be mere gossip. For example, in technical research documents, a topic {theory, method, technology, framework, application.....} could exist that does not address the actual content but is related to a large number of documents and has a very high probability sum value. In this case, the topic is weakly related to the vast majority of documents. To avoid such situations, the average probability to indicate the topic's ratio is used.

The weight is the key factor for identifying emerging topics. Indeed, it is not believed that the absolute value of the probability can be used to detect emerging topics because, as time goes by, the probability of some previous hot topics will continue to increase, as

observed in citation networks. In the previous studies, some people use high-frequency term as emerging topic's characteristic, some people use high citation as emerging topic's characteristic, and some people use the self- built indicator as emerging topic's characteristic, while Kleinberg (2003) deemed that the signal of topic emerging in a document is a “burst of activity” that can be indicated by clear and definite characteristics such as rising sharply in frequency. Learning from Kleinberg’s ideas, the approach can discover emerging topics if their probabilities suddenly rise in the specific time span. Therefore, by sorting the topics in accordance with their average probabilities and corresponding weight values and comparing the changes in the weight value with those in the adjacent time span, there is a special phenomenon: Most of the topics’ weight values exhibit slight fluctuations, whereas the values of some topics suddenly rise or decline. The topics that exhibit such sudden increases are considered probability burst topics (i.e., emerging topics). The processes involved are as follow:

- In every time span, it sorts the topics according to the ascending average probability, and the result is R={Topic$_{(p(average)min)}$, ……, Topic$_{(p(average)max)}$}.
- For each element in R, it set the weight value from 0 to the maximum number of topics in every time span. The structure is [R, Weight value]. Then, the result W={[R_1,0], [R_2,1], ……, [R_{max}, Maximum number of topics-1]} is obtained.
- For the same topic in W in an adjacent time span, it can find the difference (Topic$_{(difference)}$), namely, the current weight value (W$_{(Time\ node,\ Weight\ value)}$) minus the weight value in the previous time span (W$_{(Time\ node\text{-}1,\ Weight\ value)}$). Then, it sorts each topic in its respective time span and find the maximal value of the weight difference of a topic.

$$\mathrm{Topic}_{(difference)} = W_{(Time\ node, Weight\ value)} - \mathrm{W}_{(Time\ node-1, Weight\ value)} \quad (4)$$

$$\mathrm{Sort}(\mathrm{Topic}_{(difference)}) \rightarrow \mathrm{Max}(\mathrm{Topic}_{(difference)}) \quad (5)$$

This process can effectively reduce the interference by an order of magnitude in the comparison of topic probabilities and facilitate a fair comparison of topics with high and low probabilities. For example, Topic-A is relevant to tens of thousands of documents, and its probability sum change relative to the adjacent time span could be several hundred. In contrast, Topic-B is relevant to hundreds of documents, and its probability sum change relative to the adjacent time span could be only dozens. However, no evidence suggests that Topic-A is more important than Topic-B. Instead, the only criterion of detecting an emerging topic is the difference in its probability ratio, not its probability value.

Topic Representation

Because the previously generated topics were only a set of words and probabilities that are machine-readable rather than human-readable, understanding a topic often requires the reader to use his or her own knowledge to "guess". Furthermore, some topics cannot be guessed based on the generated words. The main task in this phase is to eliminate the fuzziness of the generated topic and to determine what the topic is talking about. For example, consider Topic-A: {prediction, artificial, scientific, paper, intelligence, synchronisation, indicator, triple, ideal}. An analyst with professional knowledge of information science and computer science can guess that the topic being discussed is "artificial intelligence prediction" and "scientific paper", but he/she must read a large number of documents related to the topic to verify this guess. In contrast, Topic-B: {metric, distance, sequence, clustering, parcel, mu, protein, modal, pair, deps} is about "Wireless Sensor Clustering Efficient". This topic-generated word set displays more technically

detailed words, but not words such as "efficient, wireless, sensor". Thus, the topic cannot be guessed.

The title is an essential attribution that uses the most concise words to display the most important content of the literature. Thus, the most frequently matched words-group in the title-set on a topic is the actual content that is of concern to most of the authors. For example, Convolutional Neural Network (CNN) technology is currently one of the most popular research areas in the field of image recognition. Therefore, in documents about image recognition technology research, the phrases "Convolutional Neural Network", "Deep Learning", "Neural Network", "Image Recognition", "OpenCV", and "Tensorflow" are frequently used. Thus, these phrases can be used to represent the topic, and the approach proposes a topic representation method to minimise the ambiguity.

Selectively processing the extracted titles can retain the maximum amount of information in the title. For emerging topics detected in a different time span, it extracts the titles of all documents related to the topics. However, only word segmentation and stop word removal are employed. Indeed, do not normalise the title text because normalisation is actually the process of text feature merging and text feature reduction. For topic representation, including more features will produce a better result.

Association rule mining is mainly used to find frequent patterns, correlation patterns and causality patterns in the data set; specifically, it discovers the links between the different items in the data set. It is generally agreed that within a data set, a strong correlation may exist among frequently co-occurring items. Therefore, association rule mining can find the most frequent item sets in a data set. This paper suggests that the algorithm "FP-growth" can be used for the association rule mining performed on the titles to find the maximum-frequency word group. Setting a certain threshold to control the size of the list of the maximum-frequency word group can facilitate finding the most frequent phrases.

High-frequency words are helpful for filtering frequent phrases. Because the maximum-frequency word group is a list that contains many co-occurring words, it is difficult to choose which items to use to represent the topic. Thus, counting the frequencies of all of the words in the previous title data set can generate a word frequency list and choose some high-frequency words to discover high-frequency phrases. In addition, taking CNN technology as an example, the statistical results demonstrate that the frequency values of the words "neural", "network", "convolutional", "deep", "learning" could be very high. Thus, it is necessary to combine these words into a word group.

The similarity calculation results are the criteria for choosing topic representation phrases. The Euclidean distance can be used to measure the similarity between the phrases in the vector space. X and Y are n-dimensional vectors that can be mapped as points in n-dimensional vector space, and thus, it can calculate the Euclidean distance, $\triangle(X,Y)$, between the two points as follows:

$$\Delta(X,Y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \qquad (6)$$

The combined high-frequency word group is used to calculate the distance between each item in the maximum-frequency phrases. It is important to choose several phrases with minimum distance values to represent the topic.

RESULTS

The following sub-sections present the experimental process and apply the approach to detect emerging topics, then validate and discuss the experimental results.

Experiment

In the absence of a definitive emerging topics list to be compared, it should fall back on homologous study whose researchers detect explicit and confirmed results from the specific data. A comparative study on interdisciplinary topics detection is provided by Xu et al. (2016). They believe that there are many technological frontiers and hotspots emerging in the domain of intersecting research and propose a new measurement index to detect interdisciplinary topics from the Web of Science data on Information Science & Library Science (LIS). Their study can be regarded as comparison based on following reasons:

(a) The online publication time is 5 January 2016, the latest research can ensure the novelty of the results.
(b) The experiment data originated from the Web of Science and Index Chemicus databases and researchers provide explicit search strategy, so it is easy to obtain the same data.
(c) The paper provides an unambiguous interdisciplinary topics list which can be evaluated the approach proposed in this paper.
(d) The domain of their case study is LIS. As the experts of LIS, it is able to analyse and compare the similarities and differences of two experimental results.

Experimental Data

In order to ensure the same data, the experiment obtain the data of article title, abstract and publication time from EXPANDED, SSCI, CPCI-S, CCR-EXPANDED database in the Web of Science. The search strategy involves: WC=Information Science & Library Science, the document type is article, the publication year is from 2001 to 2014. The result obtained is 39,446 records (Note: Because the date of collection data in Xu et al. (2016) is 07/25/2014, the data in the 2014 is incomplete, their result was 37,769 records). Given only the 2007, 2009, 2011 and 2013 data are analysed, the results cannot be disturbed by the incomplete data in 2014.

Topic Generation Example

In the pre-processing, the white spaces can be seen as delimiters for word segmentation. The experiment throws away some words from a customised English stop word list and remove the punctuation marks and symbols. It uses a WordNet backed Lemmatizer tool to normalise the words. In building the training model, it first needs to determine the number of topics. Blei , Ng and Jordan (2003) proposed that it can be used to evaluate the models by computing the perplexity of a held-out corpora. The formula is as follows:

$$\text{perplexity}(\mathrm{D}_{test}) = exp\left\{-\frac{\sum_{d=1}^{M}\log p(\boldsymbol{w}_d)}{\sum_{d=1}^{M} N_d}\right\} \qquad (7)$$

In the formula, D_{test} refers to the collection of corpora, $p(\boldsymbol{w}_d)$ denotes the generated probability of document-d in LDA model, N_{d} is the number of total words in document-d and M is the number of documents in the corpora. The experiment uses a corpus of LIS abstract from the Web of Science containing 39,446 abstracts with 168,996 unique terms. Compared the perplexity value of topics number from 20 to 200, the growth rate of topics number is set to 20. The other parameters of the LDA model are set to: alpha = 50/topics number, number of iterations = 1000. The results of perplexity value are shown in Figure 2.

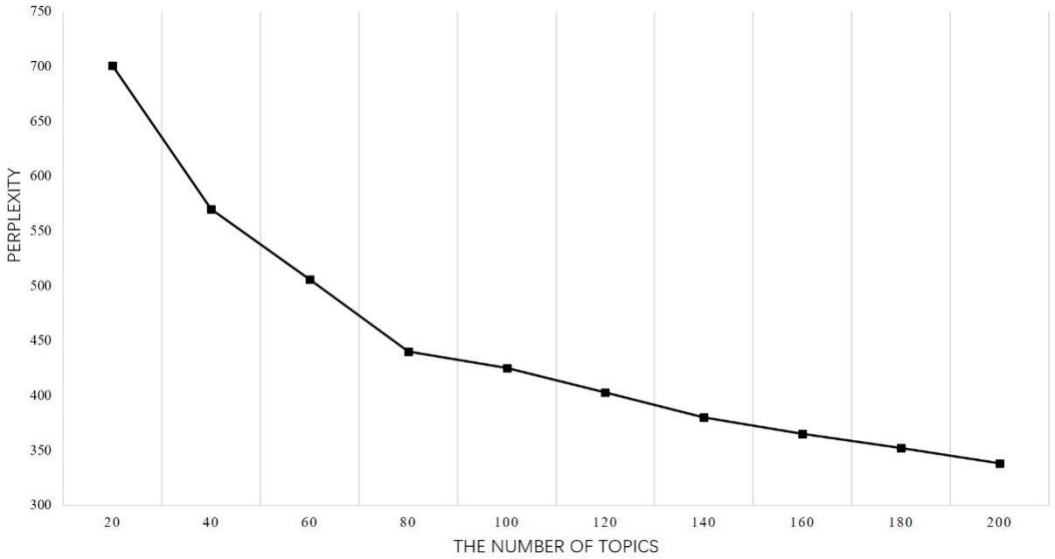

Figure 2: The Results of Perplexity Value

As is shown in Figure 2, topic number 80 is the key point. The perplexity value of model declines rapidly when the topic number is from 20 to 80. After the number exceeds 80, the perplexity value tends to decrease more slowly. To avoid over-fitting of the corpus data in the modelling process, the key point (80 topics) is selected as the definitive topic number.

Topic generation by the LDA model produces 80 topics, which consist of words and probabilities. Some of the topics are readable (e.g., Maybe topic 11 is related to the study of classification), whereas some are unreadable (e.g., the meaning of topic 54 is hard to distinguish). These findings are shown in Table 1.

Table 1: Some of the Results from the Topic Generation by the LDA Model

Topic 11—readable		**Topic 54—unreadable**	
LDA Words	**Probability**	**LDA Words**	**Probability**
classification	0.060	theory	0.027
feature	0.028	framework	0.025
approach	0.027	paper	0.024
selection	0.021	concept	0.023
intelligence	0.020	approach	0.023
path	0.018	theoretical	0.016
modeling	0.017	understanding	0.014
agent	0.017	conceptual	0.013
methodology	0.016	literature	0.012
graph	0.015	perspective	0.012

By calculating the probability, it can measure the relationship between the document and topic: <one document--related topics--probabilities>. In other words, it is able to find out what topics are associated with each document. Take document 1 in 2013 as an example: This document is related to 34 topics (Table 2). The most relevant topics of this document are: Topic 56 (0.162), Topic 45 (0.058), Topic 19 (0.056), Topic 10 (0.051), Topic 33 (0.050) and so on. Based on this way, it can calculate the probability values for all

document—relevant topics, namely p(topic|document) using equation (1). Subsequently, the probability values of all topic-relevant documents can be readily obtained via probability transformations.

Table 2: Document 1-related topics in 2013

Document 1							
Related topics	**Probability**	**Related topics**	**Probability**	**Related topics**	**Probability**	**Related topics**	**Probability**
Topic 2	0.033	Topic 17	0.026	Topic 38	0.028	Topic 60	0.049
Topic 3	0.018	Topic 19	0.056	Topic 39	0.012	Topic 62	0.024
Topic 5	0.013	Topic 20	0.014	Topic 42	0.012	Topic 63	0.010
Topic 8	0.014	Topic 22	0.018	Topic 45	0.058	Topic 64	0.037
Topic 10	0.051	Topic 23	0.039	Topic 46	0.027	Topic 70	0.014
Topic 11	0.012	Topic 29	0.012	Topic 48	0.013	Topic 76	0.013
Topic 12	0.021	Topic 33	0.050	Topic 52	0.025	Topic 78	0.047
Topic 14	0.031	Topic 34	0.013	Topic 56	0.162		
Topic 16	0.013	Topic 37	0.016	Topic 57	0.012		

Emerging Topic Detection Example

After probability transformation, it should convert the structure to calculate each topic's probability sum and average probability by using equations (2) and (3). Taking four topics (i.e., topics 10, 11, 31 and 72) as examples, the probability sum value of topic 72 is a minimum, whereas the average probability values exhibit the opposite relationship (Table 3). The average probability value is the standard used for topic ranking.

Table 3: Examples of Topic's Probability Sum and Average Probability

Topic	Probability sum	Related document number	Average probability
10	42.265	1306	0.032
11	48.285	1553	0.031
31	50.775	1596	0.032
72	41.570	1197	0.035

After ranking and weighting the topics, equations (4) and (5) are used to identify topics that increased suddenly in an adjacent year. Consider topics 10,11,31 and 72 as examples (Table 4) (i.e, from 2008 to 2009 and from 2012 to 2013, the weight value of topic 11 and topic 72 have risen dramatically). These are candidate emerging topics.

Table 4: Ranking and Weighting the Topics

Topic	Year 2013	Year 2012	Year 2011	Year 2010	Year 2009	Year 2008	Year 2007	Year 2006
10	51	48	47	49	49	52	48	50
11	47	52	53	50	53	46	49	48
31	50	50	48	46	48	50	46	49
72	58	49	44	52	46	44	47	46

Topic Representation Example

The FP-Growth algorithm is used to calculate the frequent words set. Taking topic 72 in 2013 as an example, there is a list of frequent co-occurring phrases/words (Table 5). Then, by calculating the frequency of every word in this set can identify the high-frequency words, some of which are selected to create the word group. Equation (6) is used to calculate the similarity between the word group and each term in the frequent co-occurrence phrases/words list ; the result is the topic representation. Table 6 shows that as the value of the distance decreases, the similarity of the phrases increases. Therefore, the most likely representation of topic 72 in 2013 is " Social media analysis ".

Table 5: List of Frequent Co-occurring Phrases/Words for Topic 72 in 2013

Multiword	Twice-word	Single-word
Social media analysis	Social media	Analysis
Social knowledge analysis	Web media	Network
Social network exploring	Analysis theory	Study
Social media system	Social study	Exploring
Knowledge management	Online exploring	Theory
Social media knowledge	Web analysis	Public
……	……	……
Social system study	Network data	Online

Table 6: The Representation of Topic 72 in 2013 (Euclidean distance—similarity)

Euclidean distance	Representation
1.414	Social media analysis
1.732	Social analysis
1.732	Social knowledge
……	……

Validation and Discussion

To verify the approach proposed in this paper, the results in the Xu et al.'s (2016) paper can be seen as the validation baseline. That means this study compares the two results from the same data. Xu et al. (2016) analysed the LIS data in 2007, 2009, 2011, 2013 and identified the important sub-networks, the interdisciplinary disciplines, the important topic terms in the corresponding year. The important sub-network means the significant topic network, namely it is the key research field. The interdisciplinary discipline means that the discipline with the closest relationship with LIS. The important topic term means that it is an emerging topic term or hot topic term. All of these can be seen as the reference lists for the identification of emerging topics in this study. The results vary from one experimental method to another; it cannot obtain exact same topic terms. As the experts in LIS field can manually analyse the relationship between the two results and judge whether the topic terms in two results are consistent. In the comparison tables (Tables 7, 8, 9 and 10), the top half of the table lists the sub-networks, the interdisciplinary disciplines and the important topic terms of baseline, the bottom half of the table lists the 10 detected emerging topics. There are three conditions (direct correlation, indirect correlation, not in baseline list) in the analysing and matching the topics. If the "**√**" is in the status column behind a topic, the topic is consistent with the baseline, i.e. direct correlation. If the "**△**" is

in the status column behind a topic, the topic is indirectly related to the baseline, i.e. indirect correlation. If the "○" is in the status column behind a topic, the topic is not in the baseline list.

Table 7 presents the emerging topics of LIS field in 2007. First, the emerging topic 2 and 6 are directly related to bibliometrics and bibliometric analysis. The emerging topic 4 is directly related to information management. The emerging topic 5 is directly related to information retrieval. The emerging topic 7 is directly related to information technology, computer science (interdisciplinary applications), information system and so on. The emerging topic 10 is directly related to information technology and information system. Second, the emerging topic 8 is indirectly related to management, information technology and information system. The emerging topic 9 is indirectly related to management. Third, the emerging topic 1 is not in baseline list. Using literature investigation, the results show that it is also a significant topic. To solve the problem of clinical decision support (CDS) systems being not widely used in U.S, the American Medical Informatics Association published The Roadmap for National Action on Clinical Decision Support in June 2006 to advance the widespread adoption of effective CDS systems (Kawamoto and Lobach 2007). This topic is related to the LIS, but cannot be judged whether it is an emerging topic. So it is not a definitive emerging topic. The result has not found the emerging topics which are related to the history of social science.

Table 7: The Comparison Table of Emerging Topics in LIS field (2007)

The analysis results of baseline		
The important sub-networks:	**The interdisciplinary disciplines:**	**The important topic terms:**
Bibliometrics Information Management Information Retrieval	The history of social science Computer science (interdisciplinary applications) Computer science (information system) Management	Information technology Information retrieval Information system Information management Bibliometric analysis

This study		
	Emerging Topic	**Status**
1	National clinical action decision support	○
2	Journal citation analysis, Science citation study	√
3	Electronic health record	√
4	Knowledge management	√
5	Retrieval language study, Retrieval approach, Retrieval study	√
6	Journals citation impact factor, Citation analysis	√
7	Frequency identification(RFID) adoption exploratory	√
8	Student behaviour and electronic services	△
9	Public libraries, Library services	△
10	Digital libraries	√

Table 8 presents the emerging topics of LIS field in 2009. First, the emerging topic 1 and 2 are directly related to computer science (interdisciplinary applications), computer science (information system) and Information technology. The emerging topic 4 and 5 are directly related to citation data and bibliometrics. The emerging topic 6 and 8 are directly related to management and information management. The emerging topic 7 is directly related to Complex network and bibliometrics. The emerging topic 9 is directly related to computer science (interdisciplinary applications), citation data and bibliometrics. The emerging topic 10 is directly related to communications technology. Second, the emerging topic 3 is indirectly related to semantic relationship technology. The results of this approach have not found the emerging topics related to user research, user satisfaction, open access, classification technology.

Table 8: The Comparison Table of Emerging Topics in LIS field (2009)

The analysis results of baseline		
The important sub-networks:	**The interdisciplinary disciplines:**	**The important topic terms:**
User research Open access Information management	Computer science (interdisciplinary applications) Computer science (information system) Management	User satisfaction Open access Information management Citation data Information technology Information system Complex network Communications technology Classification technology Bibliometrics Semantic relationship technology

This study		
	Emerging Topic	**Status**
1	Clinical analysis system, Unified medical language system	√
2	Computerized clinical decision support	√
3	Web 2.0, Web knowledge	△
4	Science citation, Citation analysis	√
5	Cancer Prevention, Cancer Screening, Bibliometric, Comparative study	√
6	Knowledge management	√
7	Collaboration analysis, Collaboration networks, Bibliometric analysis	√
8	Knowledge sharing, Scholarly library, Academic library, Community library, Transfer knowledge	√
9	Google science citation analysis	√
10	Online communication, Social communication	√

Table 9 presents the emerging topics of LIS field in 2011. First, the emerging topic 1 and 9 are directly related to bibliometrics and citation analysis. The emerging topic 2 is directly related to computer science and information technology. The emerging topic 3 is directly related to user acceptance. The emerging topic 4 is directly related to information technology and information system. The emerging topic 5, 7, 10 are directly related to computer science, information technology, information system, technology practice. The emerging topic 6 is directly related to business administration and economics. Second, the emerging topic 8 is not in the baseline list. It is difficult to find the relationship between this topic and LIS; it is rather the topic in medical field than in LIS field. So the detection of topic 8 has failed.

Table 10 presents the emerging topics of LIS field in 2013.. First, the emerging topic 1 is directly related to digital libraries. The emerging topic 2 is directly related to information technology, medical informatics, health care science and service. The emerging topic 3 is also directly related to medical informatics, health care science and service. The emerging topic 4 is directly related to social media. The emerging topic 5 is directly related to information technology, information retrieval, information science, medical informatics and computer science. The emerging topic 6 is directly related to knowledge sharing and information science. The emerging topic 7 is directly related to user studies (user acceptance), information technology, information system, information science and computer science. The emerging topic 8 is directly related to social network and information science. The emerging topic 9 is directly related to user research, medical informatics, health care science and service. Second, the emerging topic 10 is indirectly related to knowledge sharing. Unfortunately, the emerging topics which are related to business management and economics are not found in the results.

Table 9: The Comparison Table of Emerging Topics in LIS field (2011)

The analysis results of baseline		
The important sub-networks:	**The interdisciplinary disciplines:**	**The important topic terms:**
Information technology Technology practice Citation analysis	Computer science Business administration and economics Arts and humanities	Information technology User acceptance Information system
This study		
Emerging Topic		**Status**
1	h-core, h-index, Bibliometric citation	√
2	Online social communication	√
3	Technology acceptance model	√
4	Electronic health record	√
5	Medical diagnosis and image retrieval	√
6	Social capital and knowledge management	√
7	Semantic question answering, question answering system	√
8	Chronic pulmonary obstructive disease	○
9	Bibliometric analysis, Bibliometric study	√
10	Geographic object-based image analysis(GEOBIA)	√

Table 10: The Comparison Table of Emerging Topics in LIS field (2013)

The analysis results of baseline		
The important sub-networks:	**The interdisciplinary disciplines:**	**The important topic terms:**
User research Information technology Information system Digital library Information retrieval Information science Social network	Computer science Business management and economics Medical informatics Health care science and service	User acceptance Information technology Information retrieval Information system Social network Knowledge sharing Social media
This study		
Emerging Topic		**Status**
1	Digital libraries	√
2	Electronic health care data	√
3	Scientometric analysis, Medical analysis, Health and Cancer	√
4	Social media	√
5	Medical natural language processing	√
6	Knowledge communication, Social knowledge sharing, Communication study	√
7	Transactive memory system	√
8	Social network analysis	√
9	Health trends survey, National health survey(HINTS)	√
10	Knowledge management	△

From the value of precision, recall and F1, it can analyse and evaluate the results of four years. The value of true positive (TP) is 38, the value of false positive (FP) is 2 (the detected topic 1 in 2007 and the detected topic 8 in 2011 are not in baseline list), the value of false negative (FN) is 6 (it does not find 1 emerging topic, 4 emerging topics, 1 emerging topic in 2007, 2009, 2013 respectively). As shown in Figure 3, the value of precision (95%), recall (86%) and F1 (90%) proved that the emerging topic detection approach proposed in this paper is feasible. Furthermore, this approach can find the more detailed emerging topics because representation terms originated from the title.

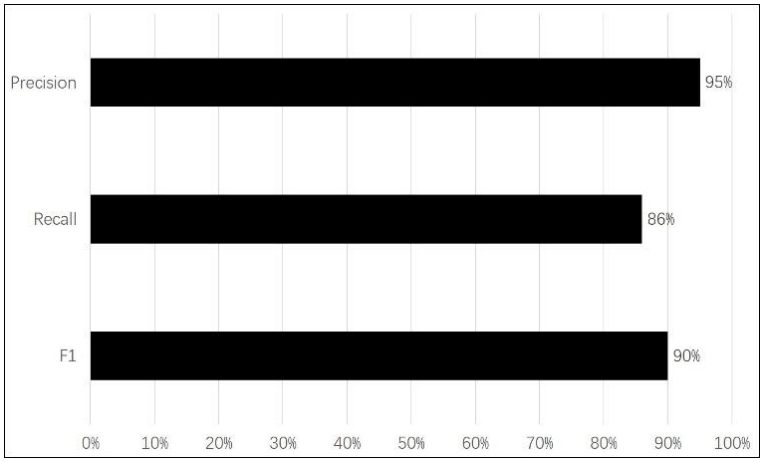

Figure 3: The Evaluation of the ETDBLA Approach

CONCLUSION

This paper present a new approach, ETDBL, for detecting emerging topics using the LDA model and association rule mining, which have several significant advantages. First, it requires only the four basic attributes of a document (i.e., category, title, time, and content). Thus, it can be applied to analyse most types of documents, including papers, patents, reports, and web pages, and can also be applied to analyse mixed types of documents. Indeed, it has wider adaptability. Second, the probability burst of topics in a specific time span can reflect a growing interest in such topics. In other words, only valuable and promising topics are interesting to most people. Additionally, by weighting each topic in its respective time span, it can decrease the magnitude of the difference between the higher-probability-value topics and the lower-probability-value topics. Third, the title reflects the essence of a document; indeed, people can determine the main idea of an article by reading its title. Thus, to find the most frequently used and co-occurring phrases in the title data set it should extract the titles of the documents related to a topic and utilise the association rule mining as a tool. These terms tell what a topic is about. To automatically find better choices within the list of the most frequently used and co-occurring phrases, it counts the frequency of each word in the list, combine high-frequency words into a word group and calculate the Euclidean distance (i.e., similarity) between the word group and each phrase. The phrases with the minimum distance values are better topic representations. Finally, regarding a newest topic detection study's results as baseline, it evaluate the feasibility of the method.

In the future, this approach can be improved in two ways. First, because small document collections and short text (e.g., Twitter) collections are less informative, have sparse word distributions, and depend strongly on context, the LDA model is not an effective tool for clustering these texts or for topic generation. Thus, it should extend the amount of information associated with such texts before creating an LDA training model. Second, some complex information documents may focus on multiple topics rather than a single, specific topic. For example, a book or dissertation is composed of many chapters, and each chapter addresses a variety of topics. Therefore, it should segment the complex information of a document into chapters and then calculate the similarities between chapters.

ACKNOWLEDGEMENT

This work was supported by the Major Project of the National Social Science Foundation of China [grant number: 15ZDB129]; the Key Project of the National Social Science Foundation of China [grant number: 14ATQ005].

REFERENCES

Allan, J., Carbonell, J., Doddington, G., Yamron, J., and Yang, Y. 1998. Topic detection and tracking pilot study: Final report. *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop,* pp. 194-218. Available at: http://ciir.cs.umass.edu/pubfiles/ir-137.pdf

Alsumait, L., Barbará, D. and Domeniconi, C. 2008. On-line lda: adaptive topic models for mining text streams with applications to topic detection and tracking, *Eighth IEEE International Conference on Data Mining*, 5-19 Dec. 2008, Pisa, Italy. pp.3-12. Available at: https://doi.org/10.1109/ICDM.2008.140.

Blei D M. 2011. Probabilistic topic models. *Proceedings of the 17th ACM SIGKDD International Conference Tutorial*s. ACM, 2011. Available at: https://doi.org/10.1145/2107736.2107741.

Blei, D. M. and Lafferty, J. D. 2005. Correlated topic models. International Conference on Neural *Information Processing Systems,* Vol. 18, no. 1: 147-154. Available at: https://doi.org/10.1145/1143844.1143859.

Blei, D. M. and Lafferty, J. D. 2006. Proceedings of the 23rd International Conference on Machine Learning, pp.113-120. Available at: https://doi.org/10.1145/1143844.1143859.

Blei, D. M. and Mcauliffe, J. D. 2010. Supervised topic models. *Advances in Neural Information Processing Systems,* Vol. 3, no. 1: 327-332. Available at: https://doi.org/10.1109/ICPR.2014.65.

Blei, D. M., Ng, A. Y. and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research*, Vol. 3, no. 1: 993-1022. Available at: https://doi.org/10.1162/jmlr.2003.3.4-5.993.

Bolelli, L., Erteki, S. and Giles, C. L. 2010. Topic and trend detection in text collections using latent dirichlet allocation. *Lecture Notes in Computer Science,* Vol. 5478, no.6: 776-780. Available at: https://doi.org/10.1007/978-3-642-00958-7_84.

Boyack, K. W. and Klavans, R. 2010. Co-citation analysis, bibliographic coupling, and direct citation: which citation approach represents the research front most accurately?. *Journal of the Association for Information Science & Technology*, Vol. 61, no.12: 2389-2404. Available at: https://doi.org/10.1002/asi.21419.

Boydgraber, J. and Blei, D. M. 2008. Syntactic topic models. *Advances in Neural Information Processing Systems,* Vol. 1, no. 1: 185-192.

Chong, W., Bo, T., Meek, C. and Blei, D. M. 2009. Markov topic models. Proceedings of the *Twelfth International Conference on Artificial Intelligence and Statistics*, PMLR 5: 583-590. Available at: http://proceedings.mlr.press/v5/wang09b/wang09b.pdf.

Cobo, M. J., López-Herrera, A. G., Herrera-Viedma, E. and Herrera, F. 2011. An approach for detecting, quantifying, and visualizing the evolution of a research field: a practical application to the fuzzy sets theory field. *Journal of Informetrics*, Vol. 5, no.1: 146-166. Available at: https://doi.org/10.1016/j.joi.2010.10.002.

Ding, Y., Chowdhury, G. G. and Foo, S. 2001. Bibliometric cartography of information retrieval research by using co-word analysis. *Information Processing & Management,* Vol. 37, no.6: 817-842. Available at: https://doi.org/10.1016/S0306-4573(00)00051-0.

Fiscus, J. G.and Doddington, G. R. 2002. Topic detection and tracking evaluation overview. In: Allan J. (eds) Topic Detection and Tracking. *The Information Retrieval Series*, Vol. 12. Springer, Boston, MA.

Fujimagari, H. and Fujita, K. 2014. Detecting research fronts using neural network model for weighted citation network analysis. *2014 IIAI 3rd International Conference on Advanced Applied Informatics*, Kitakyushu, 2014, pp. 131-136.

Fujita, K., Kajikawa, Y., Mori, J. and Sakata, I. 2012. Detecting research fronts using different types of weighted citation networks. *Technology Management for Emerging Technologies,* Vol. 32, no. 1: 267-275. Available at: https://doi.org/10.1016/j.jengtecman.2013.07.002.

Guo, X., Xiang, Y., Chen, Q., Huang, Z. and Hao, Y. 2013. LDA-based online topic detection using tensor factorization. *Journal of Information Science,* Vol. 39, no.4: 459-469. Available at: https://doi.org/10.1177/0165551512473066.

Hu, C. P., Hu, J. M., Deng, S. L. and Liu, Y. 2013. A co-word analysis of library and information science in china. *Scientometrics*, Vol.97, no.2: 369-382. Available at: https://doi.org/10.1007/s11192-013-1076-7.

Hu, J. and Zhang, Y. 2015. Research patterns and trends of recommendation system in china using co-word analysis. *Information Processing & Management,* Vol. 51, no.4: 329-339. Available at: https://doi.org/10.1016/j.ipm.2015.02.002.

Huang, J., Peng, M., Wang, H., Cao, J., Gao, W. and Zhang, X. 2017. A probabilistic method for emerging topic tracking in microblog stream. *World Wide Web-internet & Web Information Systems,* Vol. 20, no. 2: 325-350. Available at: https://doi.org/10.1007/s11280-016-0390-4.

Kawamoto, K. and Lobach, D. F. 2007. Proposal for fulfilling strategic objectives of the u.s. roadmap for national action on decision support through a service-oriented architecture leveraging hl7 services. *Journal of the American Medical Informatics Association JAMIA,* Vol. 14, no. 2: 146-155. Available at: https://doi.org/10.1197/jamia.M2298.

Kleinberg, J. 2003. Bursty and hierarchical structure in streams. *Data Mining & Knowledge Discovery*, Vol. 7, no. 4: 373-397. Available at: https://doi.org/ 10.1145/775047.775061.

Li, M. and Chu, Y. 2017. Explore the research front of a specific research theme based on a novel technique of enhanced co-word analysis. *Journal of Information Science,* Vol. 43, no. 6: 725-741.

Miao, Z., Chen, K., Fang, Y., He, J., Zhou, Y., Zhang, W., and Zha, H. 2016. Cost-effective online trending topic detection and popularity prediction in microblogging. A*CM Transactions on Information Systems,* Vol. 35, no. 3: Article 18. Available at: https://doi.org/10.1145/3001833.

Morris, S. A., Yen, G., Wu, Z., and Asnake, B. 2003. Time line visualization of research fronts. *Journal of the Association for Information Science & Technology,* Vol. 54, no.5: 413-422. Available at: https://doi.org/10.1002/asi.10227.

National Institute of Standards and Technology (NIST). 2004. The 2004 Topic Detection and Tracking (TDT2004) task definition and evaluation plan. Available at: http://itl.nist.gov/iad/mig/tests/tdt/2004/TDT04.Eval.Plan.v1.2.pdf.

Persson, O. 1994. The intellectual base and research fronts ofjasis 1986-1990. *Journal of the Association for Information Science & Technology*, Vol. 45, no.1: 31-38. Available at: https://doi.org/10.1002/(SICI)1097-4571(199401)45:13.0.CO;2-G.

Rosen-Zvi, M., Griffiths, T., Steyvers, M. and Smyth, P. 2004. The author-topic model for authors and documents. *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence* (UAI '04). AUAI Press, Arlington, Virginia, USA, pp.487-494. Available at: http://arxiv.org/abs/1207.4169v1.

Shibata, N., Kajikawa, Y., Takeda, Y., and Matsushima, K. 2008. Detecting emerging research fronts based on topological measures in citation networks of scientific publications. *Technovation*, Vol. 28, no. 11: 758-775. Available at: https://doi.org/10.1016/j.technovation.2008.03.009.

Shibata, N., Kajikawa, Y., Takeda, Y. and Matsushima, K. 2009. Comparative study on methods of detecting research fronts using different types of citation. J*ournal of the Association for Information Science & Technology,* Vol. 60, no. 3: 571–580. Available at: https://doi.org/10.1002/asi.20994.

Wang, X., Liu, J. and Sheng, F. 2014. Analysis of hotspots in the field of domestic knowledge discovery based on co-word analysis method. *Cybernetics & Information Technologies,* Vol. 14, no. 5: 145-158. Available at: https://doi.org/10.2478/cait-2014-0051.

Wayne, C. L. 2000. Multilingual topic detection and tracking: successful research enabled by corpora and evaluation. *Proceedings of the 2nd International Conference on Language Resources and Evaluation (LREC)*, Greece, May 31-June 2, 2000, European Language Resources Association, pp. 1487-1494. Available at: http://www.lrec‐conf.org/proceedings/lrec2000/pdf/168.pdf.

Wu, Q., Zhang, C., Hong, Q. and Chen, L. 2014. Topic evolution based on lda and hmm and its application in stem cell research. *Journal of Information Science*, Vol. 40, no. 5: 611-620. Available at: https://doi.org/10.1177/0165551514540565.

Xie, Y. 2014. Hotspots of ecological and environmental risk research in china based on co-word analysis. *Journal of Information & Computational Science*, Vol. 11, no. 4: 1185-1192. Available at: https://doi.org/10.12733/jics20103065.

Xu, H., Guo, T., Yue, Z., Ru, L. and Fang, S. 2016. Interdisciplinary topics of information science: a study based on the terms interdisciplinarity index series. *Scientometrics,* Vol. 106, no. 2: 583-601. Available at: https://doi.org/10.1007/s11192-015-1792-2.

Zhang, B. Beibei Zhang,Xiaohong Guan,Muhammad Junaid Khan,Yadong Zhou. 2012. A time-varying propagation model of hot topic on bbs sites and blog networks. *Information Sciences,* Vol. 187, no.1: 15-32. Available at: https://doi.org/10.1016/j.ins.2011.09.025.