Scientometrics (2020) 125:2043–2090
https://doi.org/10.1007/s11192-020-03700-5

Mapping the technology evolution path: a novel model for dynamic topic detection and tracking

Huailan Liu[1] · Zhiwang Chen[1] · Jie Tang[2] · Yuan Zhou[3] · Sheng Liu[1]

Received: 9 November 2019 / Published online: 14 September 2020

Abstract
Identifying the evolution path of a research field is essential to scientific and technological innovation. There have been many attempts to identify the technology evolution path based on the topic model or social networks analysis, but many of them had deficiencies in methodology. First, many studies have only considered a single type of information (text or citation information) in scientific literature, which may lead to incomplete technology path mapping. Second, the number of topics in each period cannot be determined automatically, making dynamic topic tracking difficult. Third, data mining methods fail to be effectively combined with visual analysis, which will affect the efficiency and flexibility of mapping. In this study, we developed a method for mapping the technology evolution path using a novel non-parametric topic model, the citation involved Hierarchical Dirichlet Process (CIHDP), to achieve better topic detection and tracking of scientific literature. To better present and analyze the path, D3.js is used to visualize the splitting and fusion of the evolutionary path. We used this novel model to mapping the artificial intelligence research domain, through a successful mapping of the evolution path, the proposed method's validity and merits are shown. After incorporating the citation information, we found that the CIHDP can be mapping a complete path evolution process and had better performance than the Hierarchical Dirichlet Process and LDA. This method can be helpful for understanding and analyzing the development of technical topics. Moreover, it can be well used to map the science or technology of the innovation ecosystem. It may also arouse the interest of technology evolution path researchers or policymakers.

Keywords Technology evolution path · Technology path mapping · Topic model · Dynamic topic detection · Hierarchical Dirichlet Process (HDP) · Path splitting and fusion

✉ Yuan Zhou
zhou_yuan@mail.tsinghua.edu.cn

[1] School of Mechanical Science and Engineering, Huazhong University of Science and Technology, Wuhan 430074, China

[2] Department of Computer Science and Technology, Tsinghua University, Beijing 100084, China

[3] School of Public Policy and Management, Tsinghua University, Beijing 100084, China

Introduction

The technology evolution path describes the emergence, transition, and extinction of a subject in this field, which can help researchers understand the history and current situation of the research field so that they can quickly identify research hotspots and gaps.

In the study of technological evolution path, discovery and presentation of topic information is a crucial problem. In recent years, an increasing number of researchers have begun to use machine-learning methods to identify the development of specific research domains based on literature data. Probabilistic topic models are useful in detecting different research topics and mining research hotspots. Especially Probabilistic Latent Semantic Analysis (PLSA) (Hofmann 1999) and Latent Dirichlet Allocation (LDA) (Blei et al. 2003), have drawn much attention in the field of topic discovery because of their effectiveness in analyzing sparse high-dimensional data, like literature data (Jeong and Min 2014; Yau et al. 2014).

There are usually two main questions when using topic models for technology mapping. First, is non-textual technical information useful for technical evolution analysis? If yes, how to add it to the topic model? Second, can we dynamically identify and track technical topics in different periods? It can help us discover the evolution path of technology more flexibly.

Most existing topic models only consider textual information. However, scientific literature contains textual information, citation information, co-author information, and so on. When topic models using only textual information are applied to analyze scientific literature, many useful features of literature are ignored. In particular, the citation relationship, which also contains robust technical evolution information, cannot be ignored when analyzing the development of specific research domains (Kajikawa et al. 2007; Zhou et al. 2016, 2019b, 2020).

The determination of the number of topics is essential for technical evolution analysis, but this is usually a troublesome problem. Generally, topic models, such as PLSA, LDA, and their extended models, need a preset number of topics. Two strategies can be used to handle this problem. One is comparing the experimental results for multiple times based on qualitative indicators such as perplexity or Normalized Mutual information (NMI) to determine the optimal number of topics. But this method requires a lot of experimentation, and the best result depends on the selected indicators. The second method is setting a relatively large number of topics, and then aggregating similar topics through Kullback–Leibler divergence, Cosine similarity, or another measure. By using the second method, the topics finally extracted are usually hard to understand (Griffiths and Steyvers 2004; Yao et al. 2011; Ding and Chen 2014).

When performing technology mapping, we hope that the algorithm can automatically determine the number of topics according to the structure of the data itself. In this way, we can not only have good adaptability to different data but also dynamically track changes of technical topics between different times. Teh et al. (2006) introduced HDP that can handle this problem. By utilizing the Dirichlet Processes feature of generating infinite clustering, the Hierarchical Dirichlet Processes (HDP) can automatically determine the appropriate number of mixture components.

In this paper, we combine textual information and citation information based on HDP to propose a new non-parametric topic model (no need to preset the parameters of the number of topics) to map the evolution path of the technology better. The novel non-parametric model based on HDP was named the citation-involved Hierarchical Dirichlet Process (CIHDP). Based on citation information, node2vec was used to convert papers in the

citation network into vector form. Then we calculated the similarity of each pair of papers in the given paper data set to construct a similarity matrix. Unlike the Hierarchical Dirichlet Process (HDP), topic distribution for each document in the CIHDP was influenced by all of the other documents with different degrees of impact. The similarity of the two articles determined the degree of influence. That is, the less similarity there was in the citation network, the smaller the impact was. As other researchers did, we used the Gibbs sampling inference to estimate parameters in our model. Quantitative experiments prove that CIHDP can achieve better subject modeling effects than LDA. Through case analysis, CIHDP can find complete path evolution information than HDP.

The technology evolution path dynamically tracked by CIHDP is visualized through D3.js finally. For those who are not very familiar with technology mapping methods, visualization helps to adjust the topic model (such as parameters adjustment), and also facilitates understanding and discussing the technology path effectively.

The rest of this paper is organized as follows. The "Related work" section briefly reviews the related works. The "Methodology" section presents the overall research process and method introduction. The "Result and discussion" section conducts a case study in the field of AI research and evaluates the validity of the model. The "Conclusions" section lays out our key findings and future works. In the end, the Appendix provides details about the improved algorithm and experiment results. The code of CIHDP and sample data are available on the GitHub repository.[1]

Related work

Technology evolution path

As a powerful presentation of the development of technology, the technology evolution path can track historical development, explore knowledge diffusion and predict future trends in technology (Adomavicius et al. 2007; Yu 2011; Huang et al. 2016; Huang et al. 2020). Given the explosive growth in the quantity of literature in the current research environment, analysis of the technology evolution path is usually based on data mining. There are two kinds of existing technology path research using literature data: bibliometrics and method based on the topic model.

Most methods of bibliometrics are based on citation analysis of scientific and technological literature (Zhou and Minshall 2014; Li et al. 2015, 2016b; Zhou et al. 2018; Xu et al. 2017, 2020; Nordensvard et al. 2018; Pan et al. 2019; Wang et al. 2018; Liu et al. 2019; Miao et al. 2020). Some methods can use to find simple information, such as keywords, influential authors, or core articles in the field literature. And then we can analyze the changes in this information over time to analyze the evolution of technology. These methods include co-word analysis (Callon et al. 1983), co-author analysis (Braun et al. 2001), bibliographic coupling (Kessler 1963), and co-citation analysis (Small 1973) and so on. Based on the citation network, some researchers also use the main path analysis method for path identification. For instance, Xiao et al. (2014) explore the knowledge diffusion path through an analysis of the main paths. Kim and Shin (2018) identify the main path of high voltage direct current transmission technology. Recently, researchers have begun to use citation network-based clustering methods, which can identify major research communities in a field.

[1] https://github.com/scientometrics-special-issue-2020-ml.

Chen et al. (2013) found that fuel cell technology consisted of several communities/clusters by clustering patent network. Moreover, the clusters used to detect and analyze technology evolution. However, the use of citation information alone is not convincing, and bibliometrics methods fail to consider both citation and text information.

The approach based on the topic model has gained more and more attention in recent years (Kong et al. 2017; Zhou et al. 2019a; Li et al. 2020). The content of the literature contains much information about technology development. By analyzing the distribution of words in the corpus, topic models perform well in extracting latent topics of documents. Of the topic models proposed in the early stage, TF-IDF, PLSA, and LDA are the most frequently used by researchers for mining topics in the corpus. Based on the topic model, some researchers explore the change of technology topics of each period, to analyze the development path of technology. For example, they are using TF-IDF cluster associated terms and phrases to constitute meaningful technological topics, Zhang et al. (2016) forecast future developments. Xu (2020) explores the identification method for innovation paths based on the linkage of scientific and technological topics. Wei et al. (2020) tracing the evolution of 3D printing technology in china using LDA-based patent abstract mining.

Nevertheless, topic models like LDA also have method flaws; that is, it needs to set the number of topics in advance. Because of the ability to automatically determine the number of topics in a given corpus, the HDP has attracted more scholars' attention. In contrast, the traditional topic model needs preset the topic number. See the next section for an introduction to the topic modeling domain.

To sum up, of the two methods used for path recognition, bibliometrics methods tend to use citation information, and topic model methods are good at using large amounts of textual information. However, the path drawn using one type of information alone is not convincing, and there are few attempts to combine citation information and text information. In this paper, we try to mapping the technology evolution path by a novel method that integrates citation information into the topic model.

Topic modeling

With the rapid increase in the amount of text data and the continuous improvement of machine learning, many latent topic discovery methods have been proposed (Hofmann 1999; Blei et al. 2003; Blei and Lafferty 2006; Teh et al. 2006; Chang and Blei 2010; Rosen-Zvi et al. 2012; Cheng et al. 2014; Fu et al. 2016; Chen et al. 2020). We first sort out the research context of the topic modeling. Among them, LDA and HDP are two typical representative algorithms. And then we introduce these two classic methods.

Most topic models, like LDA and HDP, only take the corpus as bags of words. Many data contain other information. For example, text data of web pages have hyperlinked information, comment text has user information, and scientific data have citation information and author information. Because of the excellent modularity of LDA, PLSA, and HDP, these models can be easily extended. BTM was used to integrate word co-occurrence information into LDA to solve the problem of inferring topics from large-scale short texts (Cheng et al. 2014).

Similarly, On-Line LDA (Alsumait et al. 2008) and Dynamic Online HDP (Fu et al. 2016) was used to integrate time information into LDA and the HDP to solve the problem of topic detection and tracking. Some researchers integrated author information into LDA, PLSA, or HDP to solve the problem of mining the author–topics distribution (Steyvers et al. 2004; Rosen-Zvi et al. 2012; Ming and Hsu 2016). Some other

researchers integrated information besides author information, like recipient information (Mccallum et al. 2007) and conference information (Jie et al. 2008). For example, Dai and Storkey (2009) integrated author information into the HDP to solve the author's disambiguation problem.

The above models have been shown to perform well under specific tasks and data. However, when these models are used to obtain scientific literature data, the citation information is ignored. And the citation information represents a strong topical relevance between the papers.

Several research advances have already incorporated citation information into topic modeling, and these works can be divided into two categories. One takes citation as an undirected link. For instance, the Relational Topic Model (RTM) (Chang and Blei 2010) uses LDA to model each document and uses the binary variables of a link whether or not there is a link between documents to optimizing model parameters. The other takes citation as a directed link. Based on PLSA and PHITS, Cohn and Hofmann (2000) proposed a joint probabilistic model link LDA, which generated terms and citations under a common set of underlying factors. Based on the link-based LDA, pairwise-ink LDA uses the Mixed Membership Stochastic Block (MMSB) model to generate a citation relationship alone to model the topicality of citations explicitly. Moreover, the link-PLSA-LDA method dividing data into citing part and cited part, and use the same global parameters to generate terms and citations (Nallapati et al. 2008). But, it could use PLSA to model the cited part while using LDA to model the citing part to reduce the calculation costs of a pairwise-link LDA. Kataria et al. (2010) proposed that cited-LDA and cited-PLSA-LDA extended link-LDA and link-PLSA-LDA. These two models explicitly model the influence propagation of words by citation. However, in these two models, words belonging to a citation are taken as clear information, yet the Inheritance Topic Model (ITM) views whether the word belongs to the citation as unclear information (He et al. 2009).

Specifically, LDA is a topic model based on the corpus (Blei et al. 2003), which treats the document as a set of words. LDA believes that the document contains only a limited number of hidden topics, and the number of topics corresponding to the corpus can be set to a fixed constant. Therefore, before being used, the number of topics needs to be preset (usually need). It can extract latent topics in the corpus, and each topic is composed of a set of words with different weights. At the same time, we can also obtain the probability value of each topic in the corpus (can be understood as the proportion of topics in the corpus).

The HDP is a topic model that automatically determines the number of expected topics, and can achieve dynamic topic mining. This model does not depend on the preset number of topics. As the data changes, the model can achieve adaptive changes, such as model parameter learning and automatic classification number update tasks. The model believes that the number of topics in the corpus can be infinite, and automatically learns the optimal set of topics based on the data. This model introduces the Dirichlet process and builds a hierarchical Dirichlet process, which provides a solution for sharing an infinite number of clusters among multiple documents. Similarly, its topic modeling process can mine latent topics in the corpus and output high-frequency words under each topic.

After the above overview, we focus on two problems with the topic model to make it better for technology path mapping. How to determine the number of topics automatically? How to use the citation information to mine more coherent topics? To solve these two problems, we aimed to propose a citation-involved topic model that automatically determines the number of topics (see "Methodology"). Since HDP has the specificity of automatically determining the number of topics, we selected it as the benchmark model of the improved algorithm. Unlike the link-and-content-involved topic model mentioned above, we used the citation information to calculate the similarity of each literature pair.

Besides, when sampling the topics of a specific document, the similarity information was used to adjust the impact of the topic distribution of other documents.

Path visualization

The process of technology path mapping can be divided into two parts: one is to mine the path information based on the topic model and other methods; the other is to visualize the evolution path information of technology effectively.

The visualization of the path can help us understand and analyze the development process of technology more intuitively, and there are many existing visualization methods. Citespace (Chaomei 2006), based on Java development, can perform citation analysis and timing network visualization. However, this approach cannot show the whole path of technology evolution on a single graph. Similarly, TopicRiver (Havre et al. 2002) uses rivers with varying widths to symbolize technical topics, and changes in the width of rivers to indicate changes in the strength of the topics. This method is suitable for displaying the topic of continuous development, but it is challenging to represent isolated technical topics and the developing relationship between different topics. TextFlow is a more intuitive method, which can show the split and fusion of topics by the confluence and diversion of rivers (Cui et al. 2011). Based on the semantic similarity calculation, the topic association can be used to get the split and fusion information of the topic. However, in the case of a large number of topics, this approach makes the final evolution path look messy, so that the information presentation may be inappropriate. From the perspective of the data set, a more targeted visual model is designed for different data sets based on the above two river graph representation methods. For example, the TopicFlow (Malik et al. 2013) and the OpinionFlow (Wu et al. 2014) can visualize Twitter data further be used to analyze public opinion communication. Besides, Guo et al. (2012) used a technology roadmap-style chart to represent the evolution trend, which is difficult to express complex information when the evolution path is complicated (such as cross-development path, topic intensity changes).

D3.js (data-driven documents) is a JavaScript library (Bostock et al. 2011), which is also called an interactive and dynamic data visualization tool library and can be visualized with great flexibility through programming. Based on D3.js, CellWhere showing the local interaction network organized into subcellular locations (Heberle et al. 2017), SPV simplifies biological signaling pathways visualization (Calderone and Cesareni 2018).

Combining the advantages of TopicRiver and TextFlow, we use D3.js to present the evolution path of technology. Visualize the technical topic information proposed by the novel topic model, and use the changes of rivers to represent the development of technology. It is worth mentioning that the topic model and visualization method are not isolated. They together serve to map the path of technology evolution.

Methodology

How do we map the technology evolution path? This section summarizes the overall research process, introduces the integration of citation information and topic model, and a dynamic topic detection model. Then, the process of mapping the evolution path is explained in detail.

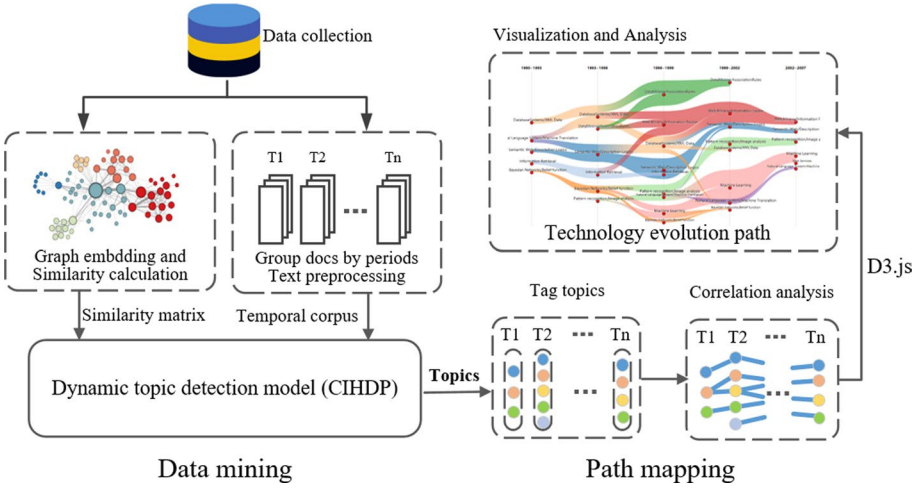


Fig. 1 A methodological framework for mapping the technology evolution path

Framework

To make the mapping process of the technological evolution path proposed in this article clearer, we have drawn the overall methodological framework, as shown in Fig. 1. Firstly, the processing of document data collection is divided into two aspects. On the one hand, we construct the citation network formed by the documents and embed the citation information of each document into a vector. Then calculate the similarity of citation information of two documents, and finally construct a document similarity matrix. On the other hand, based on the year of publication, the documents were grouped by period, and the interval of the period was uniform. Secondly, we integrated document content information with document similarity information, CIHDP was used to detect the topics of documents in each period dynamically. Thirdly, based on the topic information of each period obtained by CIHDP, we conduct topic path tracking. This part of the work is divided into two steps, the first step is to tag the topic of each period, and the second step is the correlation analysis of topics in the adjacent period to obtain the evolution path of the topics. Finally, D3.js was used to visualize the evolution path. By mapping out the evolution path, we can see the major technology branches of the field, as well as the splitting and fusion of technology evolution paths.

Measuring the similarity between documents

A citation network is a graph that contains information about the paper in each vertex, and an edge is the citation relationship between them. Vertex attributes are details about the paper, such as id, publication year, abstract, keywords, and content. Moreover, when the paper P_i referenced paper P_j, there was an arrow extending from the vertex representing P_i to the vertex representing P_j. Therefore, the citation network had the following characteristics. (1) The citation network was a directed graph in which each edge was an arrow going from one paper to the other. (2) All of the citation arrows almost always pointed backwards in time to older papers. Therefore, the graph of the citation network was acyclic

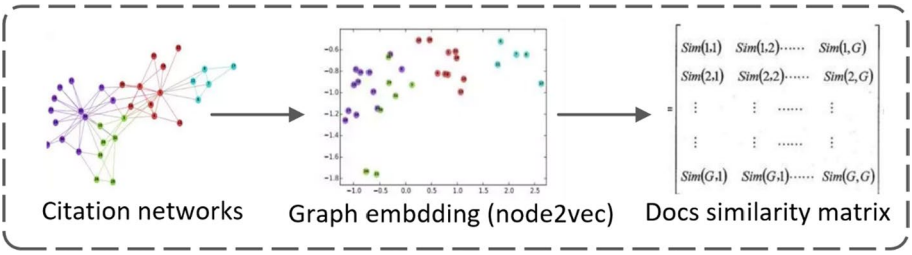


Fig. 2 Flow chart of document similarity calculation. Adapted from Perozzi et al. (2014)

and showed the development of the research field over time. (3) The most important characteristic of the citation network was the immediate relevance between topics of the paper and mentioned topics in other papers that it cited.

In bibliometrics, more researchers are using citation networks to identify and forecast development in the field of science and technology (Kajikawa et al. 2007). Therefore, many algorithms have been proposed to identify the similarity of each pair of documents in the citation network, such as bibliographic coupling, co-citation, Amsler, SimRank, P-Rank. Bibliographic coupling takes only out-links into account, so the similarity between two papers is computed based on the number of papers directly cited by both of them (Kessler 1963). Unlike bibliographic coupling, co-citation considers only in-links, and the similarity between the two papers depends on the number of papers that directly cite both of them (Small 1973). Amsler considers both direct in-links and out-links by combining the results of co-citation and bibliographic coupling (Amsler 1972). SimRank, a recursive version of co-citation, considers only in-links recursively, so the similarity between two papers is computed based on the papers that cite them (Jeh and Widom 2002). P-Rank, a recursive version of Amsler, considers both in-links and out-links recursively, so the similarity between two papers is computed based on the papers that cite them and are cited by them (Zhao et al. 2009).

However, because these models use co-citing and co-cited papers to calculate the similarity of each pair of documents, these models do not perform well in the task of calculating document pairs that have a direct citation relationship.

As shown in Fig. 2, Node2vec (Grover and Leskovec 2016) was chosen to calculate the similarity of each pair of documents. This model uses the random-walk procedure to catch the features of similarity between nodes and then embeds the node into a low-dimensional space. First, node2vec was used to acquire the vector representation of each node in the citation network in this study. Then, cosine similarity (cosineSim) was used to calculate the similarity of each pair of documents. To avoid the negative value of similarity, we finally use the following formula to calculate document similarity (docSim), whose range is [0,1]:

$$\mathrm{docSim}\left(\mathrm{doc}_1,\ \mathrm{doc}_2\right) = 0.5 + 0.5 \times \mathrm{cosineSim}\left(\mathrm{node}_1, \mathrm{node}_2\right) \tag{1}$$

where doc denotes the document, i and j denote the order of document (or node), $i \in [0, N]$, docVector denotes the embedding vector of the node in citation networks, and cosineSim denotes cosine similarity calculation function.

At last, an $N*N$ matrix was used to adjust the degree of influence between topics of different documents, where N is the number of documents in a given data set. It can be seen that the citation information is transformed into the similarity matrix of the document by

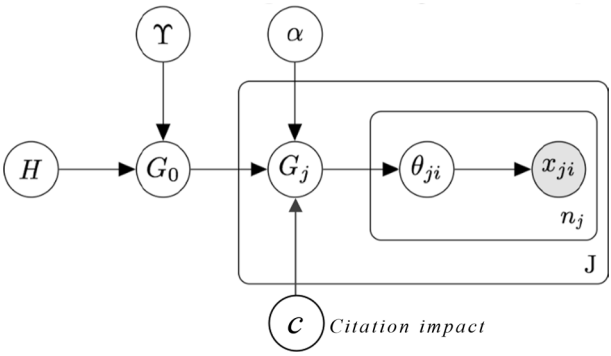


Fig. 3 Directed graphical representation of CIHDP

way of graph embedding. The similarity matrix will be used to affect the topic allocation of the document to improve the quality of topic detection and tracking.

Citation involved Hierarchical Dirichlet Process

In this section, the citation information is indirectly introduced into the novel dynamic topic model (CIHDP) by using the document similarity matrix. Group documents by time, and by using the CIHDP algorithm, we can dynamically get the topics corresponding to each period. This part will be the primary work of path mapping.

In most cases, when researches try to extract topics from scientific literature, only the textual information (such as title and abstract) are used. In most topic models (such as LDA and HDP), topics are considered as the distribution of words, and the corpus is considered as batches of words. However, how can citations be appropriately used in topic extraction? The main idea of our improved model was to use citation information as a means to enhance the textual representation of documents, to discover more coherent technology evolution paths. Node2vec was used to construct a similarity matrix based on the citation network in the scientific literature. In our model, the topics of a paper were influenced by all other papers by different degrees in the corpus. And the similarity between papers determined the degree of influence. The similarity between documents will affect the topic modeling process.

The directed graphical representation of CIHDP is shown in Fig. 3. In the directed graph, open circles represent variables, shaded circles represent observable measurements, rounded rectangles represent parameters or basic distributions, and rectangular boxes represent iteration cycles. The numbers in the lower right corner of the rectangular boxes represent the number of cycles. Among them, G_0 represents the global topic distribution of all documents, G_j represents the local topic distribution of the j-th document, and θ_{ji} represents the distribution of words under the topic. x_{ji} represents observed words in the document.

CIHDP is very similar to HDP, but CIHDP has an additional influencing factor "c", namely the influence of document similarity. In simple terms, the topic distribution of each document is not only affected by the overall topic distribution, but also by the topic distribution of similar documents. If the similarity of the two documents is higher (calculated based on the citation networks), the topic distribution between the two documents tends to be more similar, so that we can identify more coherent topics and better topic modeling effects. The citations we added are based on the global citation network, so the evolution

of the topic will be more coherent, which will also improve the effect of topic tracking. To better explain our model, we provide a brief introduction of HDP (see "Appendix A"). We also proposed a metaphor for the CIHDP to explain our model. The metaphor was named the "USA Local Specialties Restaurant Franchise." (See "Appendix B").

To verify the effectiveness of the proposed algorithm in topic detection and topic tracking, we compared the LDA, HDP, and CIHDP. On the one hand, compare the advantages and disadvantages of the algorithm level, on the other hand, compare the effectiveness of the algorithm in the technology path mapping. In this article, we use perplexity indicators to compare algorithms between models.

Perplexity is an important measurement in information theory. It is a common way of evaluating language models. The lower the perplexity is, the better the model trains the dataset. The perplexity formula is as follows:

$$\text{perplexity (D)} = \exp\left\{-\frac{\sum_{d=1}^{M}\log\left(p\left(\omega_d\right)\right)}{\sum_{d=1}^{M}N_d}\right\} \tag{2}$$

where D denotes the data set, $\sum_{d=1}^{M}N_d$ denotes the number of words in the data set, and $p\left(\omega_d\right)$ denotes the probability of that document generating a word in the data set.

Dynamic topic tracking and path identification

In order to complete the technology path mapping work, this section post-processes the topic modeling results to obtain the technology evolution path. Additionally, we designed the path expression and used visual methods to present the path.

The process of dynamic topic detection and tracking using CIHDP is shown in Fig. 1. First, documents were grouped by periods, and each group of documents was modeled to detect the topic in each period. In the topic modeling process, the document similarity matrix calculated in the last part would be used to affect the distribution of document topics. The topic name had to be determined by manually reading the corresponding high-frequency topic words.

Through the topic model, we get the distribution of each topic in each period (the probability of each word appearing under each topic). First, the tag of each topic is determined by manual calibration, and then the correlation analysis of the topics in different periods is carried out.

To determine the tag of each topic, we output the 25 words with the highest occurrence probability in the distribution of topics and words. Based on the tags given in the original data set, through manual reading, determine the topic for each tag. There are two more critical issues in this process. (1) For tiny topics (the total number of word frequencies corresponding to the topic is less than 100), and there are no highly directed words, we consider these topics to be background topics and filter them. (2) If the topic with the same tag appears at the same time, we believe that the domain represented by this tag has evolved into sub-domains within this period. In our research, we do not discuss the issue of the technical level, so we will do the fusion processing on the same tag in the same period (see "Appendices D and E"). The same topic in the same period retains the one with the highest word frequency.

We think the topics with the same tag can be connected directly in the two adjacent periods. Topics in the latter period are used to continue developing the topics in the previous period. As for topics with different tags in the two periods, the association

between topics needs to be judged by the similarity. Here we use Jensen-Shannon divergence to characterize the similarity of two topics:

$$\mathrm{JS}\left(T_1||T_2\right) = \frac{1}{2}\mathrm{KL}\left(T_1||\frac{T_1+T_2}{2}\right) + \frac{1}{2}\mathrm{KL}\left(T_2||\frac{T_1+T_2}{2}\right) \tag{3}$$

where $\mathrm{KL}(T_1||T_2)$ is the Kullback–Leibler divergence:

$$\mathrm{KL}(T_1||T_2) = -\sum T_1(x)\log\frac{1}{T_1(x)} + \sum T_1(x)\log\frac{1}{T_2(x)} = \sum T_1(x)\log\frac{T_1(x)}{T_2(x)} \tag{4}$$

The value range of JS divergence is 0–1. The smaller the JS divergence, the higher the similarity of the topics. After sorting the JS divergence, we set a similarity threshold (S) between topics, and the association within this threshold will be presented.

Based on d3.js, this paper presents the evolution path visually, which is convenient for understanding development trends in the technology field. On the graph, there are tags for each topic, the intensity of the topic, and the topic relations in the adjacent periods. The topic of the topic modeling output was a set of words. The topic intensity was used to indicate the research heat of the topic. In this study, the number of words under the topic was used to measure topic intensity. The calculation formula of topic intensity strength (T_i) was as follows:

$$\mathrm{strength}\left(T_i\right) = \frac{n\left(T_{si}\right)}{\sum_{j\in s} T_{sj}} \times n\left(doc_s\right) \tag{5}$$

In the s_th period, $n(T_{\mathrm{si}})$ represents the total word frequency of the i-th topic, $\sum_{j\in s} T_{sj}$ represents the total word frequency of all topics, and $n\left(doc_s\right)$ represents the total number of documents.

There were two elements in our visual design: points and lines. Points represented the topic in the period, whereas lines (rivers) represented the relationship between topics. Each river in the figure represents the technology evolution path, which reflects the intensity change of the technical topic and the starting and ending time of the topic.

Result and discussion

To verify the effectiveness and usefulness of the proposed methodology in technology path mapping, we select the field of artificial intelligence for a case study. Also, a comparison of similar methods was conducted.

In the first section of this part, three data sets were selected for model comparison and case study, and the data were described and preprocessed. In the second section, the model parameters are set. The third section compares the topic modeling performance of CIHDP, HDP, and LDA based on the perplexity index, and performs dynamic topic detection on the Aminer data set. The fourth section carries out path identification based on manual calibration and topic similarity calculation. The fifth section is based on D3.js to visualize the path and compare the technology path mapping capabilities of CIHDP and HDP.

Table 1 Data sets information

Data set	Category	Publication	Citation	Vocabulary	Words/Doc
Cora	7	2708	5429	1433	18.2
Citeseer	6	3312	4608	3703	31.8
Aminer	10	1000	1109	670	29.4

Data collection

The data sets used in this paper are shown in Table 1. Our study's data sets included two virtual data sets (Citeseer and Cora) and one real data set (Aminer). These data sets were used to verify the effectiveness of the CIHDP. We conducted a case study of an evolution path based on the Aminer data set. Details about the data are described below.

Citeseer[2]: This dataset contained 3312 scientific publications. All these papers had a unique category label. There were 6 categories in the data set, namely Agents, Artificial Intelligence, Database, Human–Computer Interaction, Machine Learning, and Information Retrieval. The citation network consisted of 4732 links, and the data set had 3703 unique words after stemming and removing the stop words. Moreover, there was no spelling information in this data set. When a word was repeated multiple times in a paper, we only counted it once.

Cora[3]: This dataset contained 2708 scientific publications. All these papers also had a unique category label. There were 7 categories in this dataset, namely Neural Networks, Rule Learning, Reinforcement Learning, Probabilistic Methods, Theory, Genetic Algorithms, and Case-Based. The citation network consisted of 5429 links, and the data set had 1433 unique words after the stemming process and removal of stop words. Like Citeseer, there was no spelling information in the Cora data set of each word in the vocabulary, and the words that appeared multiple times in the same paper were only recorded once.

Aminer[4]: The papers in this data set were mainly from the Aminer team and included about 10 research fields of artificial intelligence: "Data Mining/Association Rules" (DM/AR), "Web Services", "Bayesian Networks/Belief function" (bayesian networks), "Web Mining/Information Fusion"(web mining), "Semantic Web/Description Logics" (SW/DL), "Machine Learning", "Database Systems/XML Data" (DS/XD), "Information Retrieval", "Pattern recognition/Image analysis", and "Natural Language System/Statistical Machine Translation" (NLS/SMT). Since the raw data did not have the abstract information, we used the Web of Science database to complete the abstract information and delete some data that could not be found in the raw database. In addition, the paper of "Database Systems/XML Data" accounted for almost half of the total data. So, some papers from this category were also deleted to avoiding data skewness. At last, the data set contained 1000 scientific publications and information for 1109 citations. After the stemming process and removal of stop words, we had 670 unique words. The year range of the final data was 1990–2007. As with Cora and Citeseer, the words that appeared multiple times in the same paper were only recorded once.

[2] https://s3.us-east-2.amazonaws.com/dgl.ai/dataset/citeseer.zip.
[3] https://s3.us-east-2.amazonaws.com/dgl.ai/dataset/cora_raw.zip.
[4] https://lfs.aminer.cn/lab-datasets/soinf.

Table 2 Number of documents each period

Period	T1	T2	T3	T4	T5
Year span	1990–1993	1993–1996	1996–1999	1999–2002	2002–2007
Number of documents	189	278	386	285	167

Table 3 Orthogonal test level setting table

Parameters	Distribution	Ranges	Level				
			level1	level2	level3	level4	level5
β	–	[0, 0.5]	0.1	0.2	0.3	0.4	0.5
α	$\Gamma(0.1, 0.1)$	[0, 2]	0.3	0.5	1	1.5	2
γ	$\Gamma(5, 0.1))$	[0, 1]	0.1	0.3	0.5	0.7	0.9

Based on Aminer data, we conducted a case study on mapping the technology evolution path in the field of artificial intelligence. The overall time span of our analysis is from 1990 to 2007, and the overall time is divided into five periods. Among them, each period includes four consecutive years. Due to the small number of documents contained in 2006 and 2007, these two years are included in the last period (T5). The final period setting and the corresponding number of papers are shown in Table 2.

Parameter setting

When using HDP and CIHDP models for topic modeling, how to determine the model parameters (β, α, γ) is a tricky question. We have found different parameter combinations in the previous literature and applied these parameter combinations to our data set. With the number of topics (since we used a labeled data set, the number of labels is known, so we hope the number of topics is closer to the number of labels) and the perplexity as the basis for judgment, the result is not ideal. Therefore, we conducted orthogonal experiments to obtain the optimal parameter combination.

Combining the distribution of parameters $(\alpha \sim \Gamma(5, 0.1), \gamma \sim \Gamma(0.1, 0.1))$ and the parameter values set in the previous HDP model, we determine the value ranges of the three parameters. We divided the value range of the parameter into 5 levels and conducted orthogonal experiments. The parameter values range and level division are shown in the following Table 3, and the results of the orthogonal experiments are shown in the "Appendix".

When conducting orthogonal experiments, we need to consider the effect of different parameter combinations on the number of topics and perplexity. Ding and Chen (2014) designed an S value to consider the number of topics and perplexity in the parameter selection process:

$$S(\beta, \alpha, \gamma) = \sqrt{\lg(prep)^2 + \lg(K)^2} \tag{6}$$

The goal of parameter selection is that the perplexity is small enough and too many topics are unnecessary, so we chose the parameter combination that generates the lowest

Table 4 Parameter settings for LDA

Algorithm	α	β	Iterations
LDA	50/K	0.01	2000

Table 5 Parameter settings for CIHDP and HDP

Dataset	Algorithm	β	γ	α	Iterations
Cora	CIHDP	0.3	0.3	1.5	150
	HDP	0.3	0.7	0.3	150
Citeseer	CIHDP/HDP	0.4	0.9	1	150
Aminer	CIHDP/HDP	0.2	0.7	0.5	150

S value as the optimal parameter combination. Since the number of tags in the dataset we use is known, the number of topics determined using the S value may be very different from the actual number of tags. Therefore, for the labeled data set, we assign different weights to the perplexity part and the number of topics:

$$S(\beta, \alpha, \gamma) = \sqrt{a \cdot \lg(prep)^2 + b \cdot \lg(K)^2} \tag{7}$$

In the case of different weight distributions, the optimal parameters obtained by the S value are used for multiple repeated experiments to ensure that the number of topics finally obtained by the topic model is about 10. Finally, we determine the S value's weight distribution value under the three data sets and the optimal parameter combination of the three data sets under the two subject models.

In this paper, the algorithm comparison experiment requires that the number of topics identified by different algorithms is basically the same, and then the subsequent perplexity index comparison and path mapping comparison. For simplicity, if there is a parameter combination in the orthogonal test table that meets our requirements, we will directly set this group of parameters as the model parameters.

According to the orthogonal experiment table (see "Appendix C"), we find that there are parameter combinations with 7 and 6 topics in the table, and set the corresponding parameter combinations as the topic model parameters of Cora and Citeseer data, respectively. Since the aminer data needs to be compared for path drawing, we want to find a better combination of parameters, using weights $a = 0.7$, $b = 0.3$ to set the parameter combination for the aminer data. The parameter settings of the three data are shown in Table 5.

In our experiment, several parameters had to be set for CIHDP and the benchmark models LDA and HDP. See "Appendix C" for details of parameter settings. Since it is selected as the empirical test data, the following uses Aminer as an example to set the parameters.

For LDA, parameters α, β, K, and iteration need to be set (see Table 4). To compare the three models, we try to make the number of topics, that generated by the three models, consistent with the number of data set categories. Thus, the number of topics, K = 10 (for Aminer), was set the same as the number of categories of data set. Dirichlet prior parameters α and β will influence the performance of the model. We set α = 50/K, β = 0.01, which has been proved to be effective for the LDA model (Heinrich

2005; Cheng et al. 2014; Li et al. 2016a; Liu et al. 2016). In all of the experiments, the number of iterations of Gibbs samples was set to according to the perplexity index. The perplexity index of the LDA topic model has a slower convergence rate. Iterations is set to 2000.

For the CIHDP and the HDP, we used a symmetric Dirichlet distribution with parameters β for the prior H over the topic distribution, and concentration parameters γ and α that influence hierarchical DP. For CIHDP and HDP, we set $\beta=0.2$, $\gamma=0.7$, $\alpha=0.5$, and iteration = 150 (these two models can fast achieve good performance than LDA).

In our method, we had to mine the topic of relevance between documents for CIHDP. Node2vec was used to capture the information from the network via a biased random walk (Grover and Leskovec 2016). Then, we can calculate the similarity of each document pair. Two parameters were used to control the process: the return parameter p and the in–out parameter q. The same with Kim et al. (2018), the goal of our study was to identify nodes that are closely interconnected and belong to the same communities (homophily equivalence), we set $p=2$ and $q=0.125$. The other parameters involved in node2vec were set as $d=128$, $r=10$, $l=10$, and $k=10$, where d, r, l, and k denote embedding dimensions, walk per node, walk length, and context size, respectively. Parameter values also were selected based on the parameter-sensitive part of the original paper (Grover and Leskovec 2016) for the best performance.

Topic modeling

The third section compares the topic modeling performance of CIHDP, HDP, and LDA based on the perplexity index, and performs dynamic topic detection on the Aminer data set. According to the model parameters set above, we set the model parameters for CIHDP, HDP, and LDA separately. Taking the Citeseer data set as an example, we use the three topic model algorithms to perform topic modeling on the data set and obtain the perplexity data in the topic modeling process. We conduct five repeated experiments for each algorithm and use the average perplexity to draw the perplexity change curve during the topic model sampling process.

We run LDA topic modeling on Citeseer, and draw the perplexity curve with the number of iterations in the topic detection process, as shown in Table 4. It can be seen from the figure that the perplexity of LDA changes slowly, and the model needs many iterations to achieve better results.

To facilitate the comparison of the three algorithms' performance, we plot their perplexity curves in the same coordinate system. The number of iterations of LDA is much greater than the number of iterations of CIHDP and HDP, so we set a secondary abscissa system for LDA so that the three algorithms can be compared in the same graph (see Table 5a).

It can be seen from the figure that the perplexity of CIHDP and HDP is similar, and the final convergence value of the perplexity is also almost equal. It can be seen that their algorithm performance on the perplexity is similar. However, the perplexity of LDA decreases very slowly (the number of iterations needs to be 2000), and the final convergence value of the perplexity is higher than others. It can be seen that the algorithm performance of CIHDP and HDP on the perplexity is better than LDA (Fig. 4).

In the process of topic modeling for Cora and Aminer, we also found the same conclusion, the corresponding perplexity is shown in the subplot (b) and (c) of Fig. 5.

As can be seen from the above, the algorithm performance comparison between CIHDP and HDP is not apparent. Since these two algorithms can automatically

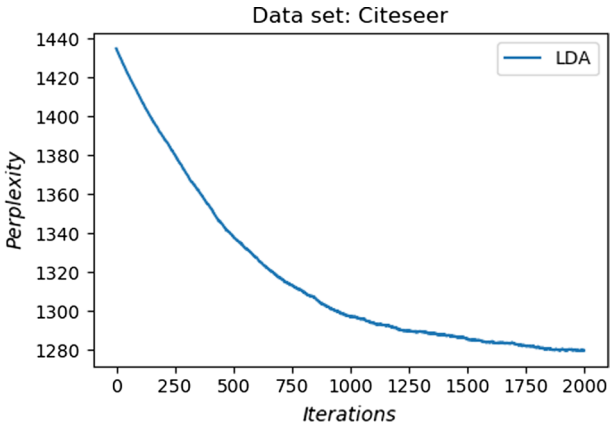


Fig. 4 Perplexity curve of LDA trained by Citeseer

determine the number of the topics, to perform dynamic topic detection and subsequent dynamic topic tracking, the following will compare the advantages and disadvantages of the two algorithms in mapping. The following section uses HDP as a benchmark model.

To compare the technology path mapping, we use CIHDP and HDP to model the topic of the Aminer data, identify the topics of each period, and get the word distribution under each topic (see "Appendix E").

Path identification

Before topic calibration in each period, we first subject the comprehensive data set to topic modeling and pre-calibration (see "Asppendix D"). Pre-calibration can make the calibration work more directional and guiding, and improve the efficiency of each calibration work. According to the topic modeling results, we first perform manual calibration (see "Appendix E"), and then perform path tracking on the calibration results.

We divide the evolution path into two categories, one is the evolution path of the same topic, and the other is the evolution path between different topics. In two adjacent time slices, the topics with the same tag can establish an association relationship directly. As for topics with different tags in the two periods, the association relationship between topics needs to be judged by the semantic similarity. As mention above, if the semantic similarity between two topics on adjacent time slices is high, we think there is an evolutionary relationship between the two topics. In this paper, the JS divergence is used to measure the semantic similarity, set the similarity threshold, and connect the topics with higher similarity to the path. First, calculate the JS divergence of all the associations of the topics with the different tags in adjacent periods and rank the JS divergence.

We set the similarity thresholds to different values, and we can get different evolution paths. In this paper, we set the similarity thresholds (S) to 10%, 20%, and 30%, respectively, and we can get the six different evolution paths in Fig. 6. Taking machine learning as an example, the red circle in Fig. 6 indicates different paths made using different similarity thresholds. We conduct information validity analysis on the paths under different similarity thresholds, and finally, determine an appropriate threshold so that the obtained evolutionary paths present the most useful information.

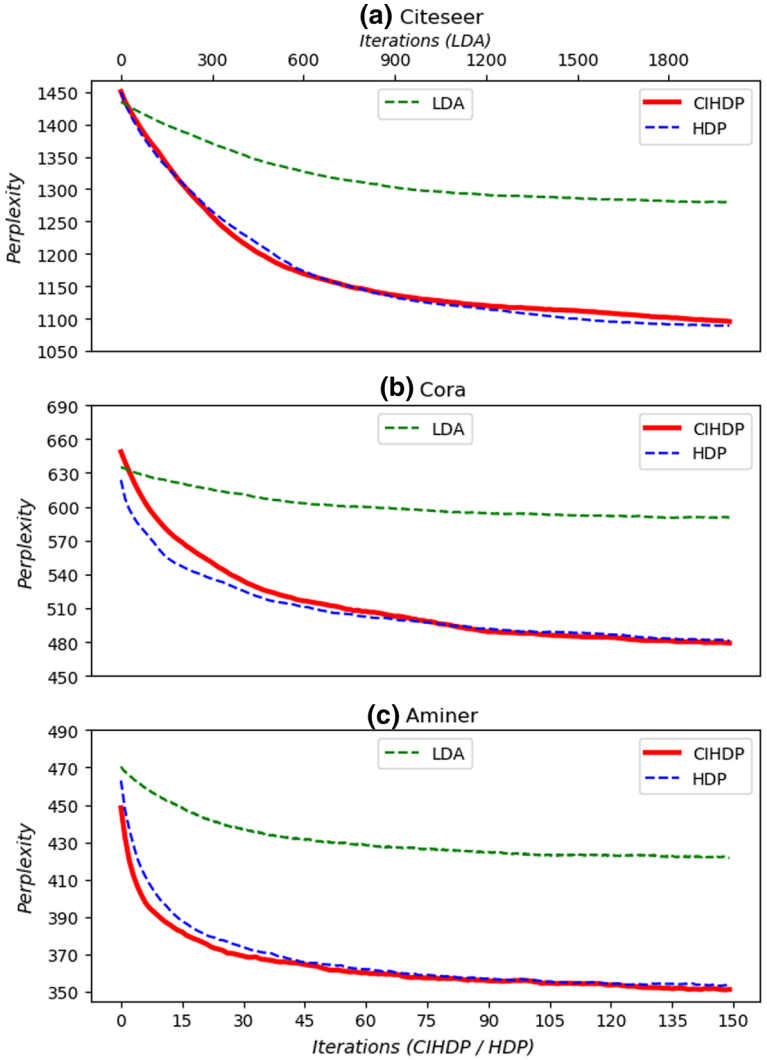


Fig. 5 Comparison of the perplexity of different topic models (LDA, HDP, CIHDP)

Mapping the evolution path

Based on Aminer data, we conducted a case study on the path identification in the field of artificial intelligence. We use CIHDP and HDP to mapping the evolution path in this field separately, and conduct a comparative analysis to prove the usefulness and advantage of the method proposed (CIHDP). After performing topic detection and topic tracking analysis for Aminer data, we can get the correlation of artificial intelligence technology topics from 1990 to 2007. That is to say, we can obtain multiple paths of technological evolution.

To intuitively analyze the path evolution in this field, we use D3.js to visualize the path evolution process. The previous section calculated the semantic similarity between topics

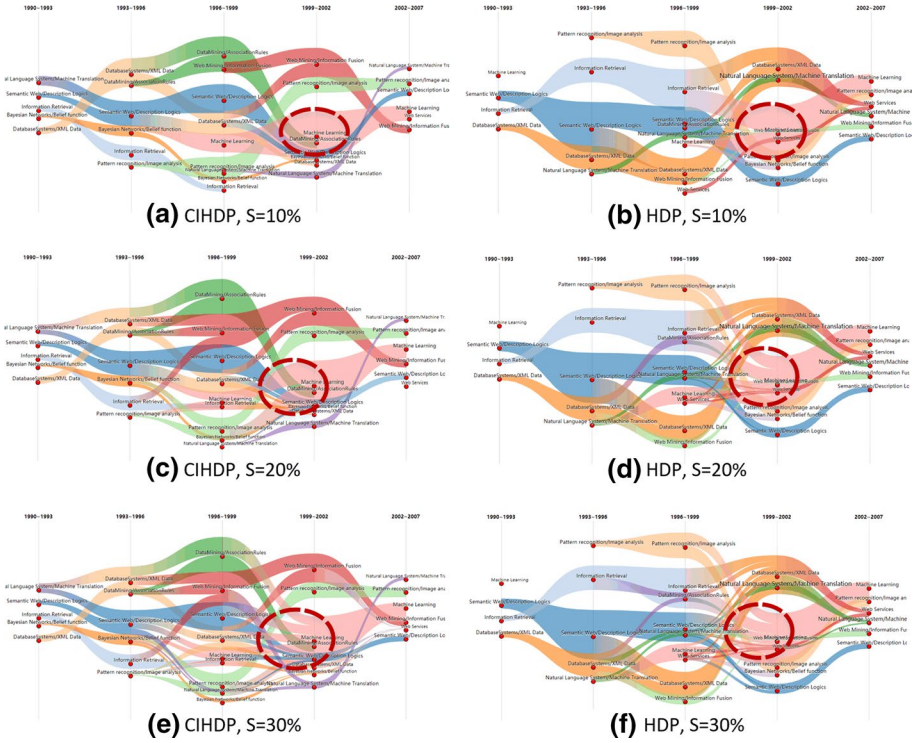


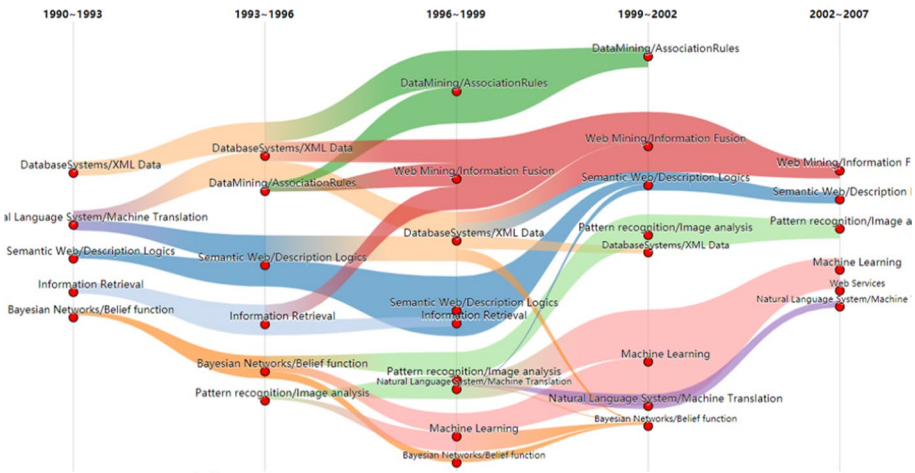
Fig. 6 Comparison of technology path mapping using different methods and topic similarity threshold

with different tags and connected every two topics with a semantic similarity larger than the similarity threshold. In this paper, the similarity threshold for CIHDP and HDP is set to $S = 20\%$, because the most effective path evolution information can be obtained in this way.

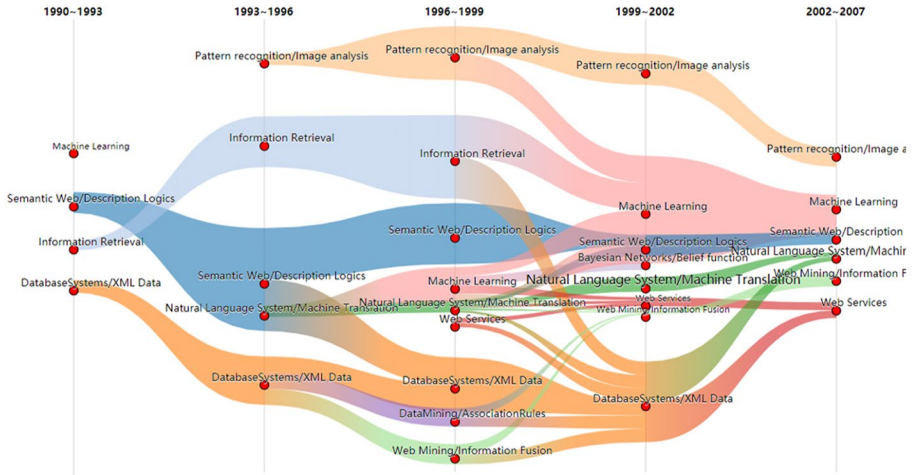
The visualization results are shown in Fig. 7. In the visual design, we use different rivers to describe the evolution path of technology. (1) Each vertical line represents a period and is marked with a year interval. From left to right, the year is getting closer to the present. (2) The topics included in each period are presented on the corresponding time vertical line, and the red dots with tags represent the technical topics. (3) Connect related topics with lines to form a series of rivers, and use different rivers to represent different technical evolution paths. (4) Different colors represent different paths, and the gradation of colors indicates the fusion and splitting of paths. (5) The width of the river expresses the topic intensity. The stronger the topic, the wider the river.

After visualizing the path information, we can conveniently conduct technical evolution analysis, and at the same time, we can also compare the effectiveness of CIHDP and HDP in path mapping.

In the sub-figure (a) in Fig. 7, we can analyze the technology evolution path mapped using CIHDP. Judging from the overall time, CIHDP has identified a total of 10 types of topics, and the overall topic recognition effect is satisfactory. In general, we first analyze the development trend of each topic. The topics that first appeared represent some basic and supporting research areas. These topics include “database systems and XML data” (DS/XD), “semantic web and description logic” (SW/DL), “natural language systems and



(a) Mapping the technology evolution path using CIHDP



(b) Mapping the technology evolution path using HDP

Fig. 7 Mapping the technology evolution path of the artificial intelligence field. (S = 20%). **a** Mapping the technology evolution path using CIHDP. **b** Mapping the technology evolution path using HDP

statistical machine translation" (NLS/SMT), "bayesian network", "information retrieval", in which DS/XD and SW/DL exist almost throughout the entire period.

The topics that emerged later were research areas that were more application-oriented and high-end R&D. These topics include "data mining", "web mining", "machine learning", "pattern recognition", and "web services". Among them, in the T4 and T5 periods, the intensity of the topic of machine learning has increased dramatically, and this kind of research is hot and popular, occupying the mainstream research status. In T5, the new topic of "web service" appeared, and the topic intensity value is not high, which means that a new path has emerged, and it is in the initial stage of path evolution.

Next, with the help of the color gradient effect, we analyze the path evolution details between different topics. Following the direction of time development, we analyze path splitting, fusion, emergence, and even disappear. (1) In the process from T2 to T3, "database

system", "information retrieval" and "data mining" merge to form "web mining". (2) During the process from T3 to T4, "web mining" continues to integrate DS/XD. (3) Based on the development of related technologies such as database systems, information retrieval technology, and data mining technology, the emerging of "web mining" path is reasonable. Moreover, similar path fusion and splitting conditions still exist. (4) For example, DS/XD, as a basic technology, always undergoes path splitting during its development and merges with "data mining", "web mining" and SW/DL. (5) Another example is that the "bayesian network" path splits, and merges with "pattern recognition", "machine learning". (6) From the figure, we can also find some interesting phenomenon. "bayesian network" and "machine learning" have interactive fusion and splitting of paths. From T2 to T3, "bayesian network" split and merge to "machine learning". From T3 to T4, "machine learning" is integrated into the "bayesian network" partly, it also shows that these two research areas are closely related. The same situation also exists between the DS/XD and SW/DL.

We also searched for information on the development of artificial intelligence and learned some facts. In 1995, Corinna proposed Support Vector Machine. In 1997, (a) the computer "dark blue" defeated Kasparov in chess. (b) Long Short-Term Memory (LSTM) was first proposed by Sepp Hochreiter and Jürgen Schmidhuber. (c) And AdaBoost was also proposed and used to achieve the effect of a strong classifier. In 1998, Tim Berners Lee proposed the semantic web and previous research is more biased towards "descriptive logic". In 2001, Conditional Random Field (CRF) was proposed by Lafferty et al. Based on these critical events, and we can better understand and believe that the above findings are in line with the facts, and also find that these significant scientific advances have promoted the emergence of new paths in T3 and T4.

Similarly, in the sub-figure (b) in Fig. 7, we can analyze the technology evolution path mapped using HDP. HDP has successfully identified 10 types of topics and has also obtained some effective technological evolution paths. However, we found that many paths identified by HDP lack actual meaning. In other words, the paths tracked using HDP are less effective than CIHDP.

The technology path mapped by CIHDP has been analyzed in detail above, so here is only a brief analysis of the path traced by HDP. It can be seen from the figure based on HDP that there is a little path information identified between T1, T2, and T3. There are many unexplained correlations and evolution paths in subgraph (b). Especially T3–T4, there are many-to-many associations. For example, data mining, web mining, information retrieval, network services, and statistical machine translation are integrated into the database system, which covers almost all the topics of T3. It is difficult to judge the core evolution path. Except for the evolution process from T3 to T4, it is almost difficult to find useful information.

Combined with the above analysis, we found that the technology evolution path we identified is consistent with the facts, indicating that the proposed method is valid in mapping the technology evolution path. We use CIHDP and HDP to make a comparative analysis of the paths and find that the former can find a more complete and detailed path evolution information. Therefore, it also proves that CIHDP is better at mapping the path of technological evolution.

Conclusion

In this study, we developed a method of mapping the technology evolution path that uses a novel non-parametric topic model (CIHDP) to achieve better dynamic topic detection and tracking of scientific literature. We performed a visual analysis of the evolution path based on D3.js. By incorporating literature citation information into the topic modeling process, the combination of textual and citation information is achieved. By using CIHDP, we have successfully completed the mapping of technological evolution paths, and obtained more detailed and complete path splitting and fusion information.

The method proposed in this paper is universal and suitable for technology path mapping. In principle, CIHDP is designed to mine and analyze data containing textual information (such as the title and abstract of the literature data) and citation information, including commonly used paper or patent data. Therefore, it is also feasible to use CIHDP to process patent data. Considering the availability and standardization of the data, we selected the paper data for technology path analysis in this study.

It is worth mentioning that the method and process in this paper can also help to solve general technology management problems, such as (1) analyzing the overall development trend of technology in a field, (2) determining the mainstream or emerging technology in the process of technology evolution, (3) or technology life cycle evaluation, and so on.

The key findings and contributions were as follows. First, this paper proposes a technology path mapping method based on an improved topic model and compares CIHDP with some traditional methods (such as LDA and HDP). For evaluating the proposed method, we used three data sets to verify our model. Through the comparison of algorithms and case studies, we found that the proposed method can find more detailed and complete technical evolution path information, and the identified evolution path is more interpretable than HDP.

Second, CIHDP makes full use of the information in the literature data, taking into account both textual semantic information and citation information, and mines the literature data from a more comprehensive perspective. This method can identify the path of technological evolution, and the experimental results also indicated that the method in this paper effectively avoids the lack of information caused by a single perspective analysis.

Third, few previous studies have mentioned how to set the parameters of non-parametric Bayesian models (such as HDP). In this paper, a large number of parameter orthogonal experiments were carried out separately on three different data sets. It provides reference values or process recommendations for users of HDP and CIHDP models to set optimal parameters. Besides, the traditional evaluation indexes (such as perplexity index) of the topic modeling algorithm are not enough to explain the pros and cons of the model, and actual case verification should be carried out.

Finally, this study conducted a visual analysis of the technology evolution path based on D3.js. We found that this visual method is suitable for analyzing complex evolution paths. Data mining combined with visual analysis can find the path splitting and fusion evolution process more efficient.

However, there are still limitations and future work here. First of all, considering the workload and time, the case study uses the core literature of the AI research field organized by the Aminer team, and the data set is not large. Later, we will consider changing to another field and use data sets with large data volume for further verification. Second, this paper has conducted in-depth data mining on scientific and technological literature. To discover more and more complete path evolution information, data fusion and analysis of different types of literature data may be performed in the future. Third, the scientific

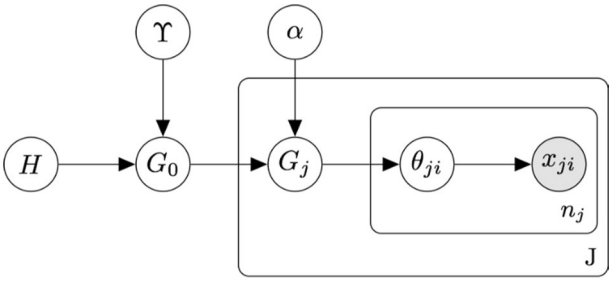


Fig. 8 The graphical representation for HDP

literature is rich in information, including not only textual and citation relationships, but also co-occurrence of authors, the quality of journal literature, and other factors. Including these factors in this model also has the potential to mine more accurate path information.

Funding This work was supported by the National Natural Science Foundation of China (Nos. 71974107, 91646102, L1824043, L1924058, L1824039, L1724034), the MOE (Ministry of Education in China) Project of Humanities and Social Sciences (16JDGC011), the Construction Project of China Knowledge Center for Engineering Sciences and Technology (No. CKCEST-2020-2-5), the UK–China Industry Academia Partnership Program (UK-CIAPP/260) and Tsinghua University Project of Volvo-supported Green Economy and Sustainable Development (20153000181). The findings and observations contained in this paper are those of the authors and do not necessarily reflect the views of the National Natural Science Foundation.

Appendix A: Hierarchical Dirichlet Process

The Hierarchical Dirichlet Process (HDP) is a non-parametric Bayesian topic model that assumes that there are infinite topics in the corpus. To understand what HDP is, we need to start with what is the Dirichlet Process (DP).

The DP is a stochastic process that generates probability distributions, parameterized by a scaling parameter λ and a base probability measure H. We denote it by $G_0 \sim DP(\gamma, H)$, A perspective on the Dirichlet process is provided by the Chinese restaurant process (CRP) (Aldous 1985). A sequence of variables $\theta_1, \theta_2, \ldots$ are independent and identically distributed according to G_0. In this metaphor, take θ_i to be a customer entering a restaurant with infinitely many tables, each serving a unique dish ϕ_k. Each arriving customer chooses a table, in proportion to the number of customers already sitting at that table, denoted as m_k. With some positive probability proportional to γ, the customer chooses a new, previously unoccupied table. The above equation can be expressed as follows (Blackwell and Macqueen 1973):

$$\theta_i|\theta_1, \ldots, \theta_{i-1}, \gamma, H \sim \sum_{k=1}^{K} \frac{m_k}{i-1+\gamma}\delta_{\theta_{\phi_k}} + \frac{\gamma}{i-1+\gamma}H \tag{8}$$

The Dirichlet process can be used to model group data, and the HDP is used to link group-specific Dirichlet processes, which can share clusters among groups of data. The graphical model is shown in Fig. 8.

The HDP has two-level DP structures. The G_j is distributed as a DP corresponding group, j, with a concentration parameter, α, and a base distribution, G_0. G_0 is also distributed as a DP with a concentration parameter, γ, and a base distribution, H. Moreover, J is the number of observed groups, n_j is the number of observed variables in group j, x_{ji} is the i_{th} observed variable in group j, θ_{ji} is the factor of x_{ji},. And $F(\theta_{ji})$ is the distribution of the x_{ji} given θ_{ji}. The generative model for HDP is as follows:

$$G_0|\gamma, H \sim DP(\gamma, H), \tag{9}$$

$$G_j|\alpha, G_0 \sim DP(\alpha, G_0), \tag{10}$$

$$\theta_{ji}|G_j \sim G_j, \tag{11}$$

$$x_{ji}|\theta_{ji} \sim F(\theta_{ji}). \tag{12}$$

To better understand HDP, we will combine the topic model process over documents to explain the HDP model. In the HDP, the sampling order of θ_{ji} is exchangeable, and so is G_j. H is taken as a Dirichlet distribution whose dimension is the size of the vocabulary, i.e., it is the distribution over an uncountable number of term distributions. Moreover, G_0 is a distribution over a countable but infinite number of topic-word distributions. For each document j, G_j is a distribution over a countable but infinite number of categorical term distributions, i.e., topic distributions of the document. Here, θ_{ji} is a categorical distribution over terms, i.e., a topic, and x_{ji} represents observed variables.

We can use the Chinese Restaurant Franchise (Teh et al. 2006) to understand HDP. In this metaphor, there a lot of restaurants that shares the same menu. There is an unlimited number of dishes in this menu. Additionally, each restaurant has an unlimited number of tables. Each table can serve an unlimited number of customers. However, each table only has one dish. All of the dishes will be chosen after customers of all the restaurants have chosen a table to sit. Like customers in CRP, when the i_{th} customer θ_{ji} enters the j_{th} restaurant, the customer will choose an occupied table or a new table according to the equation:

$$\theta_{ji}|\theta_{j1}, \ldots, \theta_{ji-1}, \alpha, G_0 \sim \sum_{t=1}^{m_{j\cdot}} \frac{n_{jt\cdot}}{i-1+\alpha}\delta_{\psi_{jt}} + \frac{\alpha}{i-1+\alpha}G_0, \tag{13}$$

where $m_{j\cdot}$ is the number of occupied tables of the j_{th} restaurant; $n_{jt\cdot}$ is the number of the customers of the t_{th} table of the j_{th} restaurant; ψ_{jt} is the dish index of the t_{th} table of the j_{th} restaurant; and $\delta_{\psi_{jt}}$ is a probability measure concentrated at ψ_{jt}.

After all of the customers of all of the restaurants have chosen a table to sit, the customers of each table will pick one dish for each table in turn. In each selection process, customers do not know which dish is good. Thus, the customers will consider the number of different dishes that have been selected. The more times a dish is selected, the more likely it is that the customers will select the dish. At the same time, there is a certain probability of choosing new dishes. This process is following the equation:

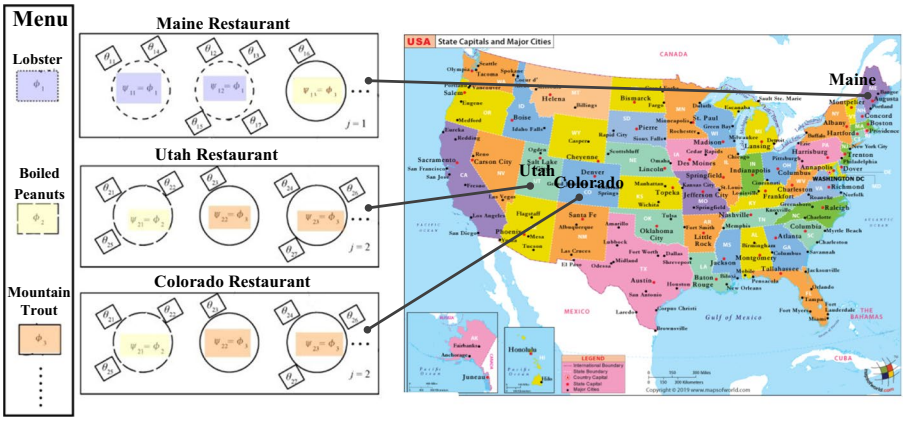


Fig. 9 A Depiction of Generation process of USA Local Specialties Restaurant Franchise. (the USA map quoted from: www.mapsofworld.com/usa/)

$$\psi_{jt}|\psi_{11},\psi_{12},\ldots,\psi_{21},\ldots,\psi_{jt-1},\gamma,H \sim \sum_{k=1}^{K}\frac{m_{.k}}{m_{..}+\gamma}\delta_{\phi_k}+\frac{\gamma}{m_{..}+\gamma}H \tag{14}$$

where $m_{.k}$ is the number of the tables serving the k_{th} dish in all the restaurant, $m_{..}$ is the number of all the occupied tables of all the restaurants, and ϕ_k is the k_{th} dish.

In this metaphor, each dish ϕ corresponds to a global topic, and each restaurant corresponds to a document. Because each document's topics are drawn from the same measure G0, global topics could be shared by all documents. Moreover, because of the two levels of DP, each document contains words from different topics.

It is easy to find that, when choosing a dish (topic) for table t in restaurant j, all the numbers of the different dishes in different restaurant are taken into considered in the same degree. That is to say, customers who choose a new dish in the restaurant j are under the same influence of all the restaurants. Corresponding to the topic model process, the topic distribution of a document was influenced to the same degree by all the documents. This is not reasonable. We think that integrating the citation information into the topic model can handle this problem.

Appendix B: Citation involved Hierarchical Dirichlet Process

In the CIHDP model, $D=\{j,j_2,\ldots\}$ is a collection of scientific literature, J is the number of scientific literature works, $\boldsymbol{x}_j$ consists of a series of words chosen from a vocabulary V as $\boldsymbol{x}_j=\{x_{j1},x_{j2},\ldots\}$, and $G=(V,E)$ represents the graph of the citation network. Moreover, each element $v\in V$ is an index of document j, and each element directed edge $(u,v)\in \mathrm{E}$ represents a citation. In our model, the citation directed graph G was not directly used. Instead, node2vec and cosine similarity were used to calculate the similarity of each pair of documents based on the citation directed graph G. A similarity matrix, M, was generated in this process. The element $sim(j_a,j_b)$ in the a_{th} row and b_{th} column of M was the similarity of document a and document b.

Similar to the HDP, there were two level Dirichlet Processes in our model. G_0, as a topic, was generated as a Dirichlet Process with a base distribution, H, and a concentration

parameter, α. Additionally, a set of measures, $\{G_j\}_{j=1}^{J}$, was drawn from the DP with a base distribution, $G_{0,}$ and a concentration parameter, γ. However, the value of ψ was influenced by the similarity matrix, *M,* except in the case of other parameters related to ψ in the HDP.

That is, in the process of choosing dishes for each table, we took into consideration the differences in customer preferences caed by differences in the geographical locations of different restaurants. This metaphor was named the USA Local Specialties Restaurant Franchise. Similar to the Chinese Restaurant Franchise (CRF) metaphor, some restaurants were sharing the same menu. The number of dishes, the number of tables in each restaurant, and the number of customers that each table could sit were all infinite. Each customer chose a table to sit in the same way as in the CRF metaphor. Nevertheless, these restaurants were located all over USA. Moreover, in the process of picking dishes, the customer not only considered the number of different dishes that had been selected, but also the geographical locations of each restaurant.

As shown in Fig. 9, all the blocks of ϕ_1 in purple represent lobster, all the blocks of ϕ_2 in yellow represent the boiled peanuts, and all the blocks of ϕ_3 in orange represent the mountain trout. We assumed that all the tables in the restaurant located in Utah had not been ordered yet, and the franchise only had three restaurants. Except for the Utah restaurant, the lobster had been ordered twice, the mountain trout had been ordered twice, and the boiled peanuts also had been ordered twice. Thus, according to the CRF's ordering principle, these three dishes would be ordered with the same probability, which was clearly unreasonable. It is known that Utah people like to eat mountain trout and have similar tastes as people in Colorado. And between boiled peanuts and lobsters, Utah people are more likely to eat boiled peanuts.

However, in the CIHDP model, the geographical locations of each restaurant were taken into consideration. Because Maine is far from Utah, the dishes ordered in Maine restaurant would have less impact on the dishes ordered by customers at Utah restaurants. On the contrary, because the Colorado is adjacent to the Utah, the dishes ordered in Colorado would have a greater impact on the dishes ordered by customers at Utah restaurant. Of course, the dishes ordered in the Utah restaurant would have the greatest impact on themselves. Different geographical location will affect the distribution of dishes on the table.

In detail, for the restaurant farthest from their restaurant, the number of different dishes that were selected would have a weak influence on their selection of dish. For the restaurant near their restaurant, the influence would be strong. It was obvious that the influence of the restaurant itself was the biggest. The following equations confirmed the whole process:

$$\theta_{ji}|\theta_{j1},\ldots,\theta_{ji-1},\alpha,G_0 \sim \sum_{t=1}^{m_{j.}} \frac{n_{jt.}}{i-1+\alpha}\delta_{\psi_{jt}} + \frac{\alpha}{i-1+\alpha}G_0, \tag{15}$$

$$\psi_{jt}|\psi_{11},\psi_{12},\ldots,\psi_{21},\ldots,\psi_{jt-1},\gamma,H \sim \sum_{k=1}^{K} \frac{m^*_{jk}}{m^*_{j.}+\gamma}\delta_{\phi_k} + \frac{\gamma}{m^*_{j.}+\gamma}H \tag{16}$$

$$m^*_{jk} = \sum_{l\in D,j=1}^{J} m_{jk} sim\left(j_{l_1}, j_{l_2}\right) \tag{17}$$

Table 6 Generation process of the USA Local Specialties Restaurant Franchise

Algorithm 1: Generation process of the USA Local Specialties Restaurant Franchise

for each restaurant $\boldsymbol{j}$ do

  for each customer $\boldsymbol{\theta}$ in restaurant $\boldsymbol{j}$ do

    choose an existing table t to sit $\boldsymbol{t} \propto \frac{n_{jt\cdot}}{i-1+\alpha}$

    choose a new table $\boldsymbol{t^{new}} \propto \frac{\alpha}{i-1+\alpha}$

    if a new table $\boldsymbol{t^{new}}$ is chosen do

      choose a dish for this table:

        choose an existing dish in menu $\boldsymbol{k} \propto \frac{m^*_{jk}}{m^*_{j\cdot}+\gamma}$

        develop a new dish $\boldsymbol{k^{new}} \propto \frac{\gamma}{m^*_{j\cdot}+\gamma}$

    end if

end for

end for

$$m^*_{j\cdot} = \sum_{k=1}^{K} m^*_{jk} \tag{18}$$

where j_l is the l_{th} restaurant.

In the process of topic modeling by CIHDP, geography corresponded to the distance between the two documents in the citation network. Moreover, the degree of the impact was determined by the similarity of the two documents calculated by node2vec and cosine similarity based on the citation network. The generation of the words for each document is described in Table 6.

We used the Gibbs sampling method to obtain the index variables, $t_{ji,}$ (associating tables with customers/words) and k_{jt} (associating tables with dishes/topics). Given these variables, we could reconstruct the distribution over topics for each document and the distribution over words for each topic. The sampling probability of table t_{ji} was as follows:

$$p\left(t_{ji}|\mathbf{t}^{\neg ji}, k\right) \propto \begin{cases} n_{jt}^{\neg ji} f_{k_{jt}}^{\neg x_{ji}}\left(x_{ji}\right), & t \text{ is an existing table} \\ \alpha p\left(x_{ji}|\mathbf{t}^{\neg \mathbf{ji}}, t_{ji} = t^{new}, \mathbf{k}\right), & t \text{ is a new table} \end{cases}, \tag{19}$$

where $n_{jt}^{\neg ji}$ is a count of customers at table t in restaurant j; $\neg ji$ denotes the counter calculated without considering the customer i in restaurant j; and $f_k^{\neg x_{ji}}\left(x_{ji}\right)$ is the likelihood of generating x_{ji} for existing table t, which could be calculated by:

$$f_k^{\neg x_{ji}}\left(x_{ji}\right) = \frac{\int f\left(x_{ji}|\phi_k\right) \Pi_{j'i' \neq ji, z_{j'i'}=k} f\left(x_{j'i'}|\phi_k\right) h\left(\phi_k\right) d\left(\phi_k\right)}{\int \Pi_{j'i' \neq ji, z_{j'i'}=k} f\left(x_{j'i'}|\phi_k\right) h\left(\phi_k\right) d\left(\phi_k\right)}, \tag{20}$$

Table 7 Factors and levels of parameter orthogonal test (for CIHDP and HDP)

Level	β	γ	α
1	0.1	0.1	0.3
2	0.2	0.3	0.5
3	0.3	0.5	1
4	0.4	0.7	1.5
5	0.5	0.9	2

where $k = k_{jt}$ is the dish served at table t in restaurant j. In addition, $p(x_{ji}|t_{\neg ji}, t_{ji} = t^{\text{new}}, k)$ is the conditional distribution of x_{ji} for $t_{ji} = t^{\text{new}}$, which could be calculated by integrating the possible values of $k_{jt^{\text{new}}}$ as follows:

$$p\left(x_{ji}|\mathbf{t}^{\neg ji}, t_{ji} = t^{\text{new}}, K\right) = \sum_{k=1}^{K} \frac{m_{jk}^{*}}{m_{j.}^{*} + \gamma} f_{k}^{\neg x_{ji}}\left(x_{ji}\right) + \frac{\gamma}{m_{j.}^{*} + \gamma} f_{k^{\text{new}}}^{\neg x_{ji}}\left(x_{ji}\right), \tag{21}$$

where m_{jk}^{*} is the influenced number of tables assigned to dish k for restaurant j.and$m_{j.}^{*}$ is the total influenced number of tables for restaurant $\boldsymbol{j}$.

Also, $f_{\neg x_{ji}}^{k^{\text{new}}}\left(x_{ji}\right) = \int f\left(x_{ji}|\phi\right) h(\phi) d(\phi)$ is the prior density of x_{ji}. The prior probability that the new table t^{new} served a new dish k^{new} was proportional to γ. If the sample value of t_{ji} was equal to t^{new}, we could obtain a sample of $k_{jt^{\text{new}}}$ by sampling as follows:

$$p(k_{jt^{\text{new}}} = k|\mathbf{t}, \mathbf{k}^{\neg jt^{\text{new}}}) \propto \begin{cases} m_{jk}^{*} f_{k}^{\neg x_{ji}}\left(x_{ji}\right), & k \text{ is an existing dish} \\ \gamma f_{k^{new}}^{\neg x_{ji}}\left(x_{ji}\right), & k \text{ is a new dish} \end{cases} \tag{22}$$

In the process of Gibbs sampling, if the value of n_{jt} was reduced to 0, the probability that the next customer would choose table t_{ji} was 0. Thus, we needed to delete the k_{jt} corresponding to t_{ji}. Moreover, if dish k did not correspond to a table at this time, k_{jt} had to be updated as follows:

$$p(k_{ji} = k|\mathbf{t}, \mathbf{k}^{\neg jt}) \propto \begin{cases} m_{jk}^{*\neg jt} f_{k}^{\neg \mathbf{x}_{\mathbf{jt}}}\left(\mathbf{x}_{jt}\right), & k \text{ is an existing dish} \\ \gamma f_{k^{new}}^{\mathbf{x}_{jt}}\left(\mathbf{x}_{jt}\right), & k \text{ is a new dish} \end{cases} \tag{23}$$

Appendix C: Parameter setting and orthogonal test

Both CIHDP and HDP are non-parametric Bayesian models. Although it is not necessary to set the number of topics in advance like LDA, it is also necessary to provide the parameter values of the prior distribution of the model. The parameters of the two models are the same, both β, γ, and α. In this study, three-factor and five-level orthogonal

Table 8 Parameter setting using orthogonal experiment for CIHDP

Condition	β	γ	α	Cora		Citeseer		Aminer	
				Perplexity	K	Perplexity	K	Perplexity	K
1	0.1	0.1	0.3	443.889	12.	991.776	10.	369.171	13.333
2	0.1	0.3	0.5	404.353	15.333	937.566	14.667	343.342	14.333
3	0.1	0.5	1	386.967	17.333	884.724	16.	349.427	13.667
4	0.1	0.7	1.5	392.396	16.	869.951	18.	347.426	12.667
5	0.1	0.9	2	380.891	18.333	857.654	18.333	345.830	13.667
6	0.2	0.1	0.5	467.330	7.667	1146.719	4.667	375.512	7.667
7	0.2	0.3	1	443.791	10.	1003.565	9.	370.188	9.
8	0.2	0.5	1.5	437.541	10.667	1026.976	9.667	366.359	9.667
9	0.2	0.7	2	443.986	10.333	967.986	13.	352.189	11.
10	0.2	0.9	0.3	451.636	9.333	1009.860	11.333	351.597	12.333
11	0.3	0.1	1	509.588	5.333	1187.428	4.667	405.518	4.333
12	0.3	0.3	1.5	477.325	7.	1099.758	6.667	382.612	6.667
13	0.3	0.5	2	470.026	8.	1022.945	10.333	383.617	6.667
14	0.3	0.7	0.3	491.719	7.333	1108.175	8.333	372.295	9.
15	0.3	0.9	0.5	471.284	8.	1068.991	7.	367.097	9.333
16	0.4	0.1	1.5	549.292	4.333	1180.325	4.333	414.750	4.333
17	0.4	0.3	2	523.036	5.	1116.887	5.333	398.329	5.667
18	0.4	0.5	0.3	530.002	6.	1197.933	4.667	387.566	7.
19	0.4	0.7	0.5	504.876	6.	1140.858	5.667	385.537	6.667
20	0.4	0.9	1	496.726	6.	1117.367	6.	377.025	8.667
21	0.5	0.1	2	535.189	4.	1282.257	2.667	422.308	3.667
22	0.5	0.3	0.3	533.201	5.	1281.657	3.667	398.238	5.
23	0.5	0.5	0.5	529.989	5.333	1162.613	4.333	397.653	6.
24	0.5	0.7	1	506.205	5.667	1145.107	5.333	399.803	5.333
25	0.5	0.9	1.5	510.587	5.667	1120.929	6.333	395.204	6.333

experiments were performed on the parameters to determine the optimal experimental parameter combination (Table 7).

To reduce the randomness of the test results, we conducted 3 repeated tests for each condition (that is, each set of parameters). The perplexity and the number of topics (K) in the experiment are the averages of the results of three repeated experiments. In the experiment, the number of model iterations is set to 150. Under this condition, the perplexity of the model can be reduced to the convergence. The results of the parameter orthogonal experiment are shown in Tables 8 and 9.

Appendix D: Topics in the overall time span

This section gives the experimental results of topic detection in the overall period in the case study. Among them, "topic tag" is a label that is calibrated for the topic after reading 25 high-frequency keywords. At the same time, the probability value and frequency

Table 9 Parameter setting using orthogonal experiment for HDP

Condition	β	γ	α	Cora		Citeseer		Aminer	
				Perplexity	K	Perplexity	K	Perplexity	K
1	0.1	0.1	0.3	436.946	10.	1036.216	8.667	373.653	10.333
2	0.1	0.3	0.5	410.220	14.333	950.905	13.	356.497	11.667
3	0.1	0.5	1	406.885	14.333	916.103	13.	341.565	12.667
4	0.1	0.7	1.5	406.226	13.667	887.200	14.333	353.480	10.333
5	0.1	0.9	2	408.243	14.	884.853	16.	344.405	13.333
6	0.2	0.1	0.5	480.046	7.333	1129.312	5.	369.397	8.333
7	0.2	0.3	1	471.605	7.333	993.338	8.	395.456	5.333
8	0.2	0.5	1.5	459.279	8.333	1024.254	8.333	377.334	8.
9	0.2	0.7	2	470.374	8.333	1004.016	9.	381.410	7.333
10	0.2	0.9	0.3	457.433	9.667	1007.059	8.667	365.448	10.
11	0.3	0.1	1	498.280	5.667	1131.996	4.333	421.889	3.667
12	0.3	0.3	1.5	498.229	5.333	1118.208	5.333	388.429	6.
13	0.3	0.5	2	492.475	6.333	1095.069	6.667	392.596	6.
14	0.3	0.7	0.3	494.636	7.	1110.199	5.667	361.881	9.667
15	0.3	0.9	0.5	468.405	8.	1049.479	7.667	367.250	9.667
16	0.4	0.1	1.5	533.065	4.667	1240.606	3.	430.033	3.
17	0.4	0.3	2	531.901	4.667	1167.213	4.333	416.268	4.333
18	0.4	0.5	0.3	512.341	5.333	1187.029	4.667	376.999	7.333
19	0.4	0.7	0.5	498.064	7.	1116.782	4.667	380.086	6.667
20	0.4	0.9	1	497.798	6.333	1112.521	6.	374.177	8.333
21	0.5	0.1	2	544.722	4.	1234.937	3.667	426.039	3.667
22	0.5	0.3	0.3	545.243	4.333	1236.789	3.	431.074	4.333
23	0.5	0.5	0.5	529.598	4.667	1169.052	4.667	394.328	5.667
24	0.5	0.7	1	523.124	4.667	1135.793	6.	392.471	6.333
25	0.5	0.9	1.5	525.971	5.667	1130.510	6.333	399.661	5.667

of occurrence of words in the topic are given behind the words (word frequency is given in parentheses). The words marked in red are representative words of the corresponding topic (Tables 10, 11).

Appendix E: Topics in each period

This section gives the experimental results of topic tracking in the case study. The topic information for each period is given below. Among them, "topic tag" is a label that is calibrated for the topic after reading 25 high-frequency keywords. At the same time, the probability value and frequency of occurrence of words in the topic are given behind the words (word frequency is given in parentheses). The words marked in red are representative words of the corresponding topic (Tables 12, 13).

Table 10 Topic extracted from Aminer by CIHDP

Ndex	Topic tag	Top 25 words of each topic
1	Data Mining/Association Rules	method: 0.01986 (133); data: 0.01703 (114); classifi: 0.01211 (81); classif: 0.01017 (68); applic: 0.01017 (68); pattern: 0.01002 (67); compar: 0.00972 (65); improv: 0.00942 (63); combin: 0.00928 (62); analysi: 0.00928 (62); larg: 0.00913 (61); learn: 0.00913 (61); effici: 0.00883 (59); rule: 0.00853 (57); approach: 0.00853 (57); studi: 0.00853 (57); mine: 0.00838 (56); techniqu: 0.00838 (56); experiment: 0.00793 (53); inform: 0.00778 (52); set: 0.00778 (52); evalu: 0.00778 (52); experi: 0.00749 (50); select: 0.00719 (48); class: 0.00689 (46)
2	Database Systems/XML Data	time: 0.0143 (65); data: 0.01276 (58); effici: 0.01167 (53); number: 0.01123 (51); requir: 0.01123 (51); effect: 0.01079 (49); distribut: 0.01057 (48); approach: 0.01057 (48); evalu: 0.01035 (47); reduc: 0.01013 (46); techniqu: 0.00969 (44); process: 0.00925 (42); implement: 0.00904 (41); improv: 0.00882 (40); control: 0.00772 (35); parallel: 0.00772 (35); studi: 0.0075 (34); develop: 0.0075 (34); larg: 0.0075 (34); simul: 0.00728 (33); memori: 0.00728 (33); oper: 0.00706 (32); achiev: 0.00706 (32); index: 0.00706 (32); space: 0.00684 (31)
3	Web Mining/Information Fusion	describ: 0.02084 (74); inform: 0.01607 (57); process: 0.01551 (55); develop: 0.01466 (52); applic: 0.01466 (52); design: 0.01326 (47); user: 0.01185 (42); data: 0.01045 (37); implement: 0.01017 (36); web: 0.00989 (35); specif: 0.00989 (35); research: 0.00904 (32); discuss: 0.00848 (30); servic: 0.00848 (30); object: 0.0082 (29); approach: 0.0082 (29); requir: 0.00792 (28); larg: 0.00736 (26); tool: 0.00736 (26); manag: 0.0068 (24); sourc: 0.0068 (24); compon: 0.0068 (24); goal: 0.0068 (24); framework: 0.0068 (24); interfac: 0.00652 (23)
4	Information Retrieval	approach: 0.01924 (64); inform: 0.01355 (45); object: 0.01325 (44); semant: 0.01235 (41); integr: 0.01055 (35); reason: 0.00995 (33); repres: 0.00995 (33); retriev: 0.00995 (33); valu: 0.00995 (33); data: 0.00965 (32); defin: 0.00965 (32); knowledg: 0.00965 (32); develop: 0.00905 (30); represent: 0.00875 (29); formal: 0.00845 (28); order: 0.00845 (28); attribut: 0.00845 (28); level: 0.00815 (27); logic: 0.00815 (27); introduc: 0.00815 (27); languag: 0.00815 (27); support: 0.00755 (25); theori: 0.00725 (24); framework: 0.00725 (24); descript: 0.00695 (23)
5	Semantic Web/Description Logics	express: 0.01866 (53); languag: 0.01796 (51); logic: 0.01726 (49); descript: 0.01375 (39); constraint: 0.01165 (33); comput: 0.01129 (32); complex: 0.01059 (30); reason: 0.01024 (29); formal: 0.01024 (29); form: 0.00989 (28); program: 0.00989 (28); class: 0.00954 (27); set: 0.00954 (27); interest: 0.00884 (25); represent: 0.00884 (25); semant: 0.00849 (24); complet: 0.00849 (24); function: 0.00849 (24); exist: 0.00814 (23); includ: 0.00814 (23); studi: 0.00779 (22); context: 0.00779 (22); integr: 0.00744 (21); order: 0.00744 (21); defin: 0.00744 (21)

Table 10 (continued)

Ndex	Topic tag	Top 25 words of each topic
6	Machine Learning	learn: 0.01791 (43); method: 0.01459 (35); function: 0.01335 (32); number: 0.01294 (31); bound: 0.01294 (31); class: 0.01086 (26); optim: 0.01045 (25); paramet: 0.01003 (24); comput: 0.0092 (22); error: 0.0092 (22); discuss: 0.00879 (21); sampl: 0.00879 (21); set: 0.00837 (20); deriv: 0.00837 (20); demonstr: 0.00837 (20); estim: 0.00796 (19); case: 0.00796 (19); linear: 0.00755 (18); obtain: 0.00755 (18); depend: 0.00755 (18); theoret: 0.00755 (18); space: 0.00755 (18); train: 0.00755 (18); prove: 0.00755 (18); data: 0.00755 (18)
7	Pattern recognition/Image analysis	imag: 0.02519 (56); reserv: 0.02205 (49); method: 0.01533 (34); object: 0.01398 (31); comput: 0.01174 (26); detect: 0.0113 (25); point: 0.0113 (25); dimension: 0.01085 (24); extract: 0.01085 (24); experi: 0.0104 (23); featur: 0.0104 (23); approach: 0.0095 (21); transform: 0.0095 (21); techniqu: 0.0095 (21); experiment: 0.00861 (19); segment: 0.00861 (19); digit: 0.00861 (19); appli: 0.00816 (18); 1999: 0.00816 (18); compar: 0.00816 (18); distanc: 0.00816 (18); map: 0.00816 (18); repres: 0.00771 (17); index: 0.00771 (17); introduc: 0.00771 (17)
8	NLS[a]/Machine Translation	word: 0.02459 (28); languag: 0.0211 (24); text: 0.02023 (23); sentenc: 0.01935 (22); automat: 0.01848 (21); lexic: 0.01674 (19); inform: 0.01412 (16); corpu: 0.01412 (16); document: 0.01325 (15); knowledg: 0.01325 (15); pars: 0.01238 (14); linguist: 0.01238 (14); test: 0.01151 (13); argu: 0.01064 (12); syntact: 0.01064 (12); sourc: 0.01064 (12); describ: 0.01064 (12); machin: 0.00976 (11); evalu: 0.00976 (11); semant: 0.00976 (11); suggest: 0.00889 (10); statist: 0.00889 (10); research: 0.00889 (10); grammar: 0.00889 (10); method: 0.00802 (9)
9	Bayesian Networks/Belief function	network: 0.02867 (25); learn: 0.02071 (18); probabl: 0.01615 (14); independ: 0.01388 (12); statist: 0.01274 (11); unit: 0.01274 (11); markov: 0.01274 (11); comput: 0.0116 (10); bayesian: 0.0116 (10); hidden: 0.0116 (10); expect: 0.01047 (9); infer: 0.01047 (9); variabl: 0.01047 (9); observ: 0.01047 (9); artifici: 0.00933 (8); neural: 0.00933 (8); condit: 0.00933 (8); estim: 0.00933 (8); minim: 0.00933 (8); local: 0.00933 (8); probabilist: 0.00819 (7); commun: 0.00819 (7); maxim: 0.00819 (7); graphic: 0.00819 (7); connect: 0.00819 (7)
10	Web Services	access: 0.0216 (9); buffer: 0.0216 (9); probabl: 0.0169 (7); size: 0.0169 (7); larg: 0.01455 (6); analysi: 0.01455 (6); page: 0.01221 (5); simul: 0.01221 (5); partit: 0.01221 (5); refer: 0.01221 (5); author: 0.00986 (4); polici: 0.00986 (4); predict: 0.00986 (4); workload: 0.00986 (4); share: 0.00986 (4); small: 0.00986 (4); number: 0.00986 (4); transact: 0.00986 (4); reason: 0.00986 (4); node: 0.00986 (4); frequenc: 0.00986 (4); examin: 0.00986 (4); overhead: 0.00986 (4); valid: 0.00986 (4); equal: 0.00751 (3)

[a]NLS: Abbreviation for natural language system

Table 11 Topic extracted from Aminer data set by HDP

Index	Topic tag	Top 25 words of each topic
1	Pattern recognition/Image analysis	method: 0.01971 (133); approach: 0.0135 (91); compar: 0.0129 (87); classif: 0.01157 (78); reserv: 0.01113 (75); classifi: 0.01098 (74); data: 0.01098 (74); improv: 0.01083 (73); experi: 0.01054 (71); applic: 0.00994 (67); combin: 0.0095 (64); featur: 0.0095 (64); imag: 0.0092 (62); obtain: 0.00906 (61); number: 0.00906 (61); experiment: 0.00891 (60); techniqu: 0.00891 (60); set: 0.00876 (59); learn: 0.00861 (58); evalu: 0.00817 (55); class: 0.00758 (51); recognit: 0.00713 (48); real: 0.00698 (47); measur: 0.00684 (46); point: 0.00684 (46)
2	Semantic Web/Description Logics	languag: 0.01574 (96); logic: 0.01394 (85); express: 0.01361 (83); applic: 0.01165 (71); semant: 0.01116 (68); integr: 0.01083 (66); reason: 0.01067 (65); formal: 0.01067 (65); object: 0.0105 (64); approach: 0.01018 (62); data: 0.00952 (58); class: 0.00919 (56); descript: 0.00919 (56); repres: 0.00903 (55); knowledg: 0.00854 (52); concept: 0.00854 (52); order: 0.00838 (51); complex: 0.00821 (50); defin: 0.00788 (48); constraint: 0.00772 (47); rule: 0.00756 (46); framework: 0.00739 (45); form: 0.00739 (45); theori: 0.0069 (42); introduc: 0.0069 (42)
3	Database Systems/XML Data	data: 0.01623 (95); effici: 0.01504 (88); larg: 0.01452 (85); time: 0.0135 (79); number: 0.01231 (72); effect: 0.01009 (59); approach: 0.00992 (58); requir: 0.00941 (55); techniqu: 0.00941 (55); process: 0.0089 (52); implement: 0.00873 (51); reduc: 0.00822 (48); space: 0.00822 (48); studi: 0.00719 (42); distribut: 0.00719 (42); compar: 0.00702 (41); evalu: 0.00702 (41); access: 0.00685 (40); simul: 0.00685 (40); improv: 0.00685 (40); index: 0.00651 (38); size: 0.00617 (36); develop: 0.00617 (36); order: 0.00617 (36); comput: 0.006 (35)
4	Web Mining/Information Fusion	inform: 0.01728 (66); describ: 0.01597 (61); applic: 0.01545 (59); design: 0.01545 (59); develop: 0.01545 (59); research: 0.01415 (54); data: 0.01389 (53); user: 0.01284 (49); discuss: 0.0118 (45); web: 0.0118 (45); process: 0.01049 (40); approach: 0.01023 (39); specif: 0.00997 (38); compon: 0.00997 (38); tool: 0.00867 (33); framework: 0.00841 (32); purpos: 0.00814 (31); issu: 0.00788 (30); servic: 0.00788 (30); support: 0.00762 (29); implement: 0.00762 (29); larg: 0.00762 (29); integr: 0.00736 (28); access: 0.00684 (26); knowledg: 0.00658 (25)
5	Machine Learning	bound: 0.01833 (41); learn: 0.01699 (38); estim: 0.01521 (34); function: 0.01343 (30); case: 0.01299 (29); optim: 0.01254 (28); sampl: 0.0121 (27); method: 0.01165 (26); number: 0.01121 (25); comput: 0.01077 (24); class: 0.01032 (23); approxim: 0.00988 (22); independ: 0.00988 (22); space: 0.00943 (21); complex: 0.00943 (21); train: 0.00899 (20); prove: 0.00854 (19); exampl: 0.00854 (19); time: 0.0081 (18); error: 0.0081 (18); finit: 0.00765 (17); probabl: 0.00765 (17); depend: 0.00765 (17); variabl: 0.00721 (16); theoret: 0.00721 (16)

Table 11 (continued)

Index	Topic tag	Top 25 words of each topic
6	Information Retrieval	document: 0.02924 (47); retriev: 0.02491 (40); inform: 0.02429 (39); text: 0.02429 (39); word: 0.01871 (30); probabilist: 0.01561 (25); method: 0.01561 (25); languag: 0.01561 (25); statist: 0.01499 (24); combin: 0.01437 (23); evalu: 0.01375 (22); sentenc: 0.0119 (19); corpu: 0.01128 (18); lexic: 0.01128 (18); term: 0.01066 (17); test: 0.01066 (17); pars: 0.01004 (16); collect: 0.01004 (16); techniqu: 0.00942 (15); approach: 0.00942 (15); automat: 0.00942 (15); linguist: 0.0088 (14); weight: 0.00818 (13); relev: 0.00818 (13); knowledg: 0.00818 (13)
7	Data Mining/Association Rules	mine: 0.02064 (32); data: 0.02 (31); discoveri: 0.01936 (30); user: 0.01872 (29); pattern: 0.01808 (28); knowledg: 0.01808 (28); discov: 0.01744 (27); web: 0.01615 (25); page: 0.01551 (24); search: 0.01487 (23); interest: 0.01423 (22); number: 0.01359 (21); studi: 0.01359 (21); analysi: 0.01231 (19); content: 0.01038 (16); defin: 0.01038 (16); effici: 0.01038 (16); method: 0.00974 (15); larg: 0.00974 (15); log: 0.00846 (13); access: 0.00846 (13); make: 0.00846 (13); rule: 0.00846 (13); link: 0.00782 (12); experi: 0.00782 (12)
8	NLS/Statistical Machine Translation	manag: 0.02453 (31); transact: 0.0206 (26); process: 0.01981 (25); execut: 0.01824 (23); control: 0.01745 (22); correct: 0.01431 (18); workflow: 0.01352 (17); oper: 0.01352 (17); support: 0.01274 (16); exist: 0.01274 (16); applic: 0.01195 (15); ensur: 0.01116 (14); data: 0.01116 (14); recoveri: 0.01038 (13); flexibl: 0.01038 (13); concurr: 0.01038 (13); activ: 0.01038 (13); environ: 0.00959 (12); requir: 0.00959 (12); distribut: 0.00959 (12); busi: 0.00959 (12); constraint: 0.00959 (12); log: 0.00881 (11); mechan: 0.00881 (11); make: 0.00802 (10)
9	Bayesian Networks/Belief function	network: 0.02867 (24); learn: 0.02038 (17); hidden: 0.01801 (15); statist: 0.01682 (14); probabilist: 0.01564 (13); analysi: 0.01564 (13); bayesian: 0.01564 (13); markov: 0.01564 (13); deriv: 0.01209 (10); probabl: 0.0109 (9); neural: 0.0109 (9); field: 0.0109 (9); uncertainti: 0.00972 (8); recognit: 0.00972 (8); variabl: 0.00853 (7); intellig: 0.00853 (7); local: 0.00853 (7); version: 0.00853 (7); commun: 0.00853 (7); artifici: 0.00853 (7); error: 0.00735 (6); report: 0.00735 (6); activ: 0.00735 (6); shown: 0.00735 (6); observ: 0.00735 (6)
10	Web Services	action: 0.02434 (9); describ: 0.01905 (7); observ: 0.0164 (6); agent: 0.0164 (6); occur: 0.01376 (5); attempt: 0.01376 (5); behavior: 0.01376 (5); event: 0.01376 (5); domain: 0.01111 (4); defin: 0.01111 (4); expect: 0.01111 (4); chang: 0.01111 (4); collect: 0.01111 (4); belief: 0.01111 (4); initi: 0.01111 (4); reliabl: 0.00847 (3); respons: 0.00847 (3); rel: 0.00847 (3); learn: 0.00847 (3); produc: 0.00847 (3); calcul: 0.00847 (3); frequent: 0.00847 (3); condit: 0.00847 (3); graph: 0.00847 (3); discoveri: 0.00847 (3)

Table 12 Topics in each period (Topic extracted from Aminer by CIHDP)

Period	Index	Topic tag	Top 25 words of each topic
T1	1	NLS/Statistical Machine Translation	languag: 0.02013 (37); semant: 0.01797 (33); approach: 0.0131 (24); inform: 0.01255 (23); author: 0.01255 (23); integr: 0.01147 (21); object: 0.01093 (20); rule: 0.01039 (19); oper: 0.00985 (18); type: 0.00985 (18); defin: 0.00985 (18); repres: 0.00931 (17); comput: 0.00931 (17); applic: 0.00877 (16); logic: 0.00877 (16); natur: 0.00877 (16); extens: 0.00823 (15); formal: 0.00823 (15); descript: 0.00823 (15); requir: 0.00768 (14); framework: 0.00768 (14); data: 0.00768 (14); depend: 0.00768 (14); class: 0.00768 (14); knowledg: 0.00768 (14)
	2	Database Systems/XML Data	process: 0.01809 (25); data: 0.01665 (23); number: 0.0145 (20); implement: 0.01378 (19); effici: 0.01378 (19); distribut: 0.01235 (17); join: 0.01163 (16); parallel: 0.01163 (16); time: 0.01091 (15); transact: 0.01091 (15); author: 0.01091 (15); effect: 0.01091 (15); execut: 0.01091 (15); design: 0.00948 (13); oper: 0.00948 (13); dynam: 0.00948 (13); approach: 0.00948 (13); optim: 0.00876 (12); simul: 0.00876 (12); environ: 0.00876 (12); develop: 0.00876 (12); requir: 0.00876 (12); control: 0.00876 (12); achiev: 0.00876 (12); support: 0.00804 (11)
	3	Information Retrieval	data: 0.01888 (21); retriev: 0.01621 (18); index: 0.01621 (18); method: 0.01621 (18); probabilist: 0.01443 (16); evalu: 0.01443 (16); describ: 0.01354 (15); appli: 0.01264 (14); text: 0.01264 (14); inform: 0.01175 (13); document: 0.01175 (13); techniqu: 0.01086 (12); requir: 0.01086 (12); store: 0.01086 (12); analysi: 0.00997 (11); singl: 0.00997 (11); larg: 0.00908 (10); collect: 0.00908 (10); specif: 0.00908 (10); experi: 0.00819 (9); approach: 0.00819 (9); infer: 0.00819 (9); attribut: 0.00819 (9); framework: 0.0073 (8); comparison: 0.0073 (8)
	4	Semantic Web/Description Logics	properti: 0.01719 (14); express: 0.01477 (12); constraint: 0.01477 (12); logic: 0.01235 (10); function: 0.01235 (10); effici: 0.01114 (9); exist: 0.01114 (9); construct: 0.01114 (9); studi: 0.01114 (9); respect: 0.01114 (9); finit: 0.00993 (8); techniqu: 0.00993 (8); updat: 0.00872 (7); possibl: 0.00872 (7); shown: 0.00872 (7); independ: 0.00872 (7); prove: 0.00872 (7); complet: 0.00751 (6); access: 0.00751 (6); number: 0.00751 (6); recurs: 0.00751 (6); practic: 0.00751 (6); linear: 0.00751 (6); comput: 0.00751 (6); transit: 0.00751 (6)
	5	Bayesian Networks/Belief function	measur: 0.01763 (8); adapt: 0.01548 (7); method: 0.01548 (7); learn: 0.01333 (6); knowledg: 0.01333 (6); autom: 0.01333 (6); domain: 0.01333 (6); address: 0.01333 (6); case: 0.01118 (5); handl: 0.01118 (5); form: 0.00903 (4); task: 0.00903 (4); outlin: 0.00903 (4); belief: 0.00903 (4); real: 0.00903 (4); demonstr: 0.00903 (4); desir: 0.00903 (4); methodolog: 0.00903 (4); unit: 0.00903 (4); simpl: 0.00688 (3); classif: 0.00688 (3); optim: 0.00688 (3); induct: 0.00688 (3); network: 0.00688 (3); discoveri: 0.00688 (3)

Table 12 (continued)

Period	Index	Topic tag	Top 25 words of each topic
T2	1	Database Systems/XML Data	data: 0.02371 (45); object: 0.01584 (30); design: 0.01375 (26); support: 0.01375 (26); implement: 0.01322 (25); applic: 0.01217 (23); effici: 0.01217 (23); inform: 0.01217 (23); describ: 0.01165 (22); integr: 0.01165 (22); develop: 0.0106 (20); number: 0.0106 (20); schema: 0.0106 (20); orient: 0.0106 (20); research: 0.01007 (19); techniqu: 0.01007 (19); process: 0.01007 (19); oper: 0.00955 (18); access: 0.00955 (18); interfac: 0.00955 (18); exist: 0.00955 (18); approach: 0.00955 (18); includ: 0.00902 (17); effect: 0.00902 (17); issu: 0.0085 (16)
	2	Semantic Web/Description Logics	languag: 0.01268 (21); express: 0.01268 (21); logic: 0.01268 (21); formal: 0.01208 (20); semant: 0.01089 (18); constraint: 0.01089 (18); complex: 0.01089 (18); applic: 0.01029 (17); reason: 0.01029 (17); class: 0.00969 (16); approach: 0.00969 (16); order: 0.00909 (15); complet: 0.00909 (15); descript: 0.00909 (15); represent: 0.00909 (15); defin: 0.00909 (15); form: 0.00849 (14); concept: 0.00849 (14); extend: 0.00849 (14); featur: 0.00789 (13); knowledg: 0.00789 (13); notion: 0.00789 (13); determin: 0.00789 (13); rule: 0.00789 (13); call: 0.0073 (12)
	3	Information Retrieval	retriev: 0.02693 (28); document: 0.0212 (22); inform: 0.02025 (21); method: 0.01738 (18); search: 0.01738 (18); time: 0.01738 (18); index: 0.01643 (17); collect: 0.01643 (17); techniqu: 0.01547 (16); evalu: 0.01452 (15); text: 0.01356 (14); experi: 0.01356 (14); pattern: 0.01165 (12); effect: 0.0107 (11); larg: 0.0107 (11); automat: 0.00974 (10); term: 0.00974 (10); reduc: 0.00974 (10); scheme: 0.00879 (9); improv: 0.00879 (9); word: 0.00879 (9); test: 0.00879 (9); match: 0.00879 (9); produc: 0.00783 (8); space: 0.00783 (8)
	4	Data Mining/Association Rules	data: 0.02275 (23); rule: 0.02078 (21); larg: 0.01882 (19); mine: 0.01882 (19); knowledg: 0.01882 (19); discoveri: 0.01784 (18); discov: 0.01686 (17); studi: 0.01686 (17); analysi: 0.01392 (14); develop: 0.01392 (14); user: 0.01294 (13); set: 0.01294 (13); associ: 0.01196 (12); discuss: 0.01196 (12); relationship: 0.01098 (11); process: 0.01098 (11); applic: 0.01098 (11); effici: 0.01 (10); method: 0.01 (10); level: 0.00902 (9); interest: 0.00902 (9); field: 0.00902 (9); pattern: 0.00706 (7); simpl: 0.00706 (7); class: 0.00706 (7)
	5	Bayesian Networks/Belief function	method: 0.02373 (19); network: 0.02002 (16); approach: 0.01755 (14); compar: 0.01755 (14); learn: 0.01632 (13); probabl: 0.01384 (11); train: 0.01261 (10); repres: 0.01137 (9); error: 0.01137 (9); final: 0.01137 (9); appli: 0.01014 (8); function: 0.01014 (8); probabilist: 0.01014 (8); techniqu: 0.01014 (8); neural: 0.01014 (8); illustr: 0.0089 (7); inform: 0.0089 (7); requir: 0.0089 (7); bayesian: 0.0089 (7); correct: 0.0089 (7); distribut: 0.00766 (6); interpret: 0.00766 (6); suitabl: 0.00766 (6); number: 0.00766 (6); class: 0.00766 (6)

Table 12 (continued)

Period	Index	Topic tag	Top 25 words of each topic
	6	Pattern recognition/Image analysis	imag: 0.01802 (12); transform: 0.01507 (10); estim: 0.01507 (10); point: 0.01211 (8); discuss: 0.01211 (8); space: 0.01211 (8); demonstr: 0.01211 (8); featur: 0.01211 (8); properti: 0.01211 (8); match: 0.01064 (7); digit: 0.01064 (7); set: 0.01064 (7); consist: 0.01064 (7); extract: 0.00916 (6); simpl: 0.00916 (6); experi: 0.00916 (6); process: 0.00916 (6); optim: 0.00916 (6); dimension: 0.00916 (6); select: 0.00916 (6); project: 0.00768 (5); chang: 0.00768 (5); cluster: 0.00768 (5); effici: 0.00768 (5); scale: 0.00768 (5)
T3	1	Machine Learning	learn: 0.01745 (53); method: 0.01319 (40); data: 0.01188 (36); combin: 0.01155 (35); optim: 0.01056 (32); classif: 0.01056 (32); improv: 0.01024 (31); classifi: 0.00991 (30); error: 0.00958 (29); select: 0.00827 (25); class: 0.00827 (25); function: 0.00794 (24); set: 0.00794 (24); number: 0.00761 (23); distribut: 0.00728 (22); addit: 0.00728 (22); estim: 0.00728 (22); train: 0.00728 (22); discuss: 0.00728 (22); real: 0.00696 (21); compar: 0.00696 (21); bound: 0.00663 (20); applic: 0.00663 (20); experi: 0.00663 (20); linear: 0.00663 (20)
	2	Web Mining/Information Fusion	web: 0.01793 (35); inform: 0.01793 (35); describ: 0.01589 (31); approach: 0.01182 (23); develop: 0.01131 (22); process: 0.01131 (22); larg: 0.01131 (22); data: 0.0108 (21); framework: 0.01029 (20); design: 0.01029 (20); applic: 0.00978 (19); user: 0.00825 (16); complex: 0.00774 (15); requir: 0.00774 (15); site: 0.00723 (14)); defin: 0.00723 (14); introduc: 0.00723 (14); final: 0.00723 (14); current: 0.00723 (14); compon: 0.00723 (14); engin: 0.00672 (13); observ: 0.00621 (12); content: 0.00621 (12); identifi: 0.00621 (12); set: 0.00621 (12)
	3	Pattern recognition/Image analysis	reserv: 0.03175 (36); imag: 0.02035 (23); 1999: 0.01947 (22); object: 0.0186 (21); compar: 0.01509 (17); approach: 0.01333 (15); comput: 0.01333 (15); method: 0.01246 (14); experi: 0.01246 (14); number: 0.01158 (13); obtain: 0.01158 (13); approxim: 0.01158 (13); point: 0.0107 (12); scheme: 0.00982 (11); requir: 0.00982 (11); function: 0.00895 (10); retriev: 0.00895 (10); dimension: 0.00895 (10); repres: 0.00895 (10); featur: 0.00895 (10); extract: 0.00895 (10); local: 0.00895 (10); detect: 0.00895 (10); improv: 0.00807 (9); appli: 0.00807 (9)
	4	Data Mining/Association Rules	effici: 0.02008 (21); number: 0.01913 (20); applic: 0.01818 (19); search: 0.01723 (18); data: 0.01723 (18); techniqu: 0.01534 (16); method: 0.01534 (16); process: 0.0125 (13); time: 0.01155 (12); similar: 0.01155 (12); comput: 0.01155 (12); implement: 0.01061 (11); store: 0.01061 (11); cluster: 0.01061 (11); mine: 0.00966 (10); tree: 0.00966 (10); studi: 0.00966 (10); find: 0.00966 (10); explor: 0.00871 (9); space: 0.00871 (9); interest: 0.00871 (9); index: 0.00871 (9); discuss: 0.00871 (9); discoveri: 0.00871 (9); rule: 0.00777 (8)

Table 12 (continued)

Period	Index	Topic tag	Top 25 words of each topic
	5	Semantic Web/Description Logics	express: 0.0245 (17); descript: 0.02165 (15); languag: 0.01738 (12); relationship: 0.01595 (11); knowledg: 0.01595 (11); logic: 0.01595 (11); integr: 0.01453 (10); repres: 0.01453 (10); defin: 0.01453 (10); inform: 0.01453 (10); represent: 0.01311 (9); semant: 0.01311 (9); introduc: 0.01168 (8); linguist: 0.01168 (8); task: 0.01168 (8); class: 0.01168 (8); partial: 0.01026 (7); make: 0.01026 (7); domain: 0.01026 (7); case: 0.00883 (6); reason: 0.00883 (6); comput: 0.00883 (6); version: 0.00883 (6); interest: 0.00883 (6); ontolog: 0.00883 (6)
	6	NLS/Statistical Machine Translation	text: 0.02085 (12); evalu: 0.01915 (11); automat: 0.01573 (9); research: 0.01573 (9); method: 0.01573 (9); success: 0.01402 (8); analysi: 0.01402 (8); demonstr: 0.0106 (6); complet: 0.0106 (6); studi: 0.0106 (6); approach: 0.0106 (6); describ: 0.00889 (5); achiev: 0.00889 (5); definit: 0.00889 (5); rate: 0.00889 (5); corpu: 0.00889 (5); normal: 0.00889 (5); test: 0.00889 (5); lexic: 0.00889 (5); task: 0.00889 (5); explor: 0.00889 (5); statist: 0.00718 (4); larg: 0.00718 (4); resolut: 0.00718 (4); assess: 0.00718 (4)
	7	Database Systems/XML Data	approach: 0.0267 (12); design: 0.02013 (9); compar: 0.01357 (6); guarante: 0.01357 (6); distribut: 0.01357 (6); elimin: 0.01357 (6); small: 0.01138 (5); singl: 0.01138 (5); assign: 0.00919 (4); attempt: 0.00919 (4); solut: 0.00919 (4); updat: 0.00919 (4); method: 0.00919 (4); simpl: 0.00919 (4); oper: 0.00919 (4); adapt: 0.00919 (4); consist: 0.00919 (4); replic: 0.00919 (4); transact: 0.00919 (4); superior: 0.00919 (4); progress: 0.00919 (4); creat: 0.007 (3); cluster: 0.007 (3); connect: 0.007 (3); applic: 0.007 (3)
	8	Bayesian Networks/Belief function	state: 0.02 (6); markov: 0.01677 (5); suggest: 0.01677 (5); cluster: 0.01355 (4); infer: 0.01355 (4); hidden: 0.01355 (4); global: 0.01355 (4); dynam: 0.01032 (3); field: 0.01032 (3); question: 0.01032 (3); user: 0.01032 (3); collect: 0.01032 (3); network: 0.01032 (3); fix: 0.01032 (3); length: 0.01032 (3); interfac: 0.01032 (3); strong: 0.01032 (3); approxim: 0.01032 (3); challeng: 0.0071 (2); context: 0.0071 (2); exact: 0.0071 (2); neural: 0.0071 (2); empir: 0.0071 (2); answer: 0.0071 (2); deriv: 0.0071 (2)
T4	1	Machine Learning	class: 0.01439 (21); combin: 0.01439 (21); classifi: 0.01303 (19); error: 0.01303 (19); learn: 0.01303 (19); method: 0.01236 (18); experi: 0.01236 (18); classif: 0.01168 (17); estim: 0.01168 (17); data: 0.01168 (17); train: 0.011 (16); compar: 0.011 (16); addit: 0.01032 (15); improv: 0.01032 (15); approach: 0.00964 (14); sampl: 0.00964 (14); appli: 0.00964 (14); real: 0.00896 (13); set: 0.00896 (13); distribut: 0.00896 (13); experiment: 0.00896 (13); condit: 0.00896 (13); select: 0.00828 (12); statist: 0.00828 (12); work: 0.00828 (12)

Table 12 (continued)

Period	Index	Topic tag	Top 25 words of each topic
	2	Pattern recognition/Image analysis	approach: 0.02492 (24); discuss: 0.02183 (21); represent: 0.01565 (15); reserv: 0.01565 (15); comput: 0.01565 (15); method: 0.01565 (15); evalu: 0.01565 (15); number: 0.01256 (12); inform: 0.01256 (12); develop: 0.01256 (12); effici: 0.01256 (12); introduc: 0.01153 (11); applic: 0.01153 (11); recognit: 0.01153 (11); pattern: 0.0105 (10); multipl: 0.0105 (10); imag: 0.00947 (9); extens: 0.00947 (9); featur: 0.00947 (9); techniqu: 0.00844 (8); exist: 0.00844 (8); compar: 0.00844 (8); complex: 0.00844 (8); detect: 0.00844 (8); time: 0.00742 (7)
	3	Web Mining/Information Fusion	pattern: 0.02139 (18); web: 0.02021 (17); inform: 0.01551 (13); user: 0.01434 (12); process: 0.01199 (10); mine: 0.01199 (10); page: 0.01199 (10); improv: 0.01081 (9); discov: 0.01081 (9); discoveri: 0.01081 (9); design: 0.01081 (9); log: 0.00964 (8); research: 0.00964 (8); domain: 0.00964 (8); analysi: 0.00964 (8); time: 0.00964 (8); analyz: 0.00846 (7); shown: 0.00846 (7); studi: 0.00846 (7); execut: 0.00846 (7); content: 0.00846 (7); issu: 0.00846 (7); complet: 0.00846 (7); activ: 0.00846 (7); experiment: 0.00846 (7)
	4	Semantic Web/Description Logics	languag: 0.01924 (12); servic: 0.01924 (12); semant: 0.01767 (11); web: 0.01451 (9); logic: 0.01451 (9); describ: 0.01451 (9); reason: 0.01293 (8); repres: 0.01136 (7); ontolog: 0.01136 (7); busi: 0.01136 (7); develop: 0.01136 (7); complex: 0.01136 (7); process: 0.01136 (7); support: 0.00978 (6); specif: 0.00978 (6); concept: 0.0082 (5); includ: 0.0082 (5); definit: 0.0082 (5); interact: 0.0082 (5); set: 0.0082 (5); involv: 0.0082 (5); implement: 0.0082 (5); formal: 0.0082 (5); approach: 0.0082 (5); collect: 0.0082 (5)
	5	NLS/Statistical Machine Translation	research: 0.01898 (8); text: 0.01898 (8); evalu: 0.01667 (7); data: 0.01435 (6); analysi: 0.01435 (6); machin: 0.01435 (6); introduc: 0.01204 (5); interest: 0.01204 (5); term: 0.00972 (4); document: 0.00972 (4); attent: 0.00972 (4); specif: 0.00972 (4); address: 0.00972 (4); review: 0.00972 (4); complex: 0.00972 (4); sentenc: 0.00972 (4); task: 0.00972 (4); focu: 0.00972 (4); languag: 0.00972 (4); measur: 0.00972 (4); articl: 0.00972 (4); techniqu: 0.00972 (4); automat: 0.00972 (4); benchmark: 0.00741 (3); origin: 0.00741 (3)
	6	Web Services	alloc: 0.01311 (3); implement: 0.01311 (3); time: 0.00902 (2); support: 0.00902 (2); creat: 0.00902 (2); main: 0.00902 (2); experiment: 0.00902 (2); requir: 0.00902 (2); resourc: 0.00902 (2); reserv: 0.00902 (2); schedul: 0.00902 (2); integr: 0.00902 (2); studi: 0.00902 (2); simpl: 0.00902 (2); high: 0.00902 (2); control: 0.00902 (2); contribut: 0.00902 (2); compon: 0.00902 (2); desir: 0.00902 (2); bound: 0.00902 (2); framework: 0.00902 (2); defin: 0.00902 (2); lexic: 0.00492 (1); intellig: 0.00492 (1); current: 0.00492 (1)

Table 13 Topics in each period (Topic extracted from Aminer by HDP)

Period	Index	Topic tag	Top 25 words of each topic
T1	1	Semantic Web/Description Logics	languag: 0.01746 (32); semant: 0.01692 (31); logic: 0.01367 (25); rule: 0.01312 (24); object: 0.01204 (22); constraint: 0.0115 (21); author: 0.01095 (20); repres: 0.01095 (20); oper: 0.01095 (20); comput: 0.01041 (19); construct: 0.00987 (18); class: 0.00933 (17); extens: 0.00933 (17); descript: 0.00933 (17); defin: 0.00933 (17); function: 0.00879 (16); express: 0.00879 (16); properti: 0.00879 (16); represent: 0.00879 (16); framework: 0.00879 (16); type: 0.00824 (15); consist: 0.00824 (15); featur: 0.00824 (15); reason: 0.00824 (15); formal: 0.00824 (15)
	2	Database Systems/XML Data	effici: 0.02469 (28); develop: 0.01944 (22); effect: 0.01769 (20); optim: 0.01594 (18); join: 0.01594 (18); time: 0.01419 (16); data: 0.01419 (16); author: 0.01331 (15); distribut: 0.01331 (15); number: 0.01331 (15); process: 0.01243 (14); requir: 0.01068 (12); execut: 0.01068 (12); evalu: 0.00981 (11); techniqu: 0.00981 (11); parallel: 0.00981 (11); dynam: 0.00981 (11); implement: 0.00893 (10); analysi: 0.00893 (10); studi: 0.00893 (10); improv: 0.00893 (10); determin: 0.00893 (10); applic: 0.00806 (9); scheme: 0.00806 (9); simul: 0.00806 (9)
	3	Information Retrieval	inform: 0.02399 (23); retriev: 0.01986 (19); probabilist: 0.01572 (15); text: 0.01572 (15); knowledg: 0.01365 (13); term: 0.01262 (12); document: 0.01262 (12); varieti: 0.01262 (12); index: 0.01158 (11); statist: 0.01158 (11); data: 0.01158 (11); automat: 0.01055 (10); integr: 0.01055 (10); word: 0.01055 (10); specif: 0.01055 (10); describ: 0.01055 (10); techniqu: 0.01055 (10); estim: 0.00951 (9); evalu: 0.00951 (9); method: 0.00951 (9); store: 0.00951 (9); collect: 0.00848 (8); determin: 0.00848 (8); larg: 0.00848 (8); approach: 0.00848 (8)
	4	Machine Learning	method: 0.0214 (9); test: 0.01674 (7); case: 0.01674 (7); desir: 0.01442 (6); learn: 0.01209 (5); final: 0.01209 (5); analysi: 0.01209 (5); autom: 0.01209 (5); measur: 0.01209 (5); previou: 0.00977 (4); simpl: 0.00977 (4); extend: 0.00977 (4); induct: 0.00977 (4); evalu: 0.00977 (4); theoret: 0.00977 (4); captur: 0.00977 (4); content: 0.00977 (4); adapt: 0.00744 (3); error: 0.00744 (3); real: 0.00744 (3); classic: 0.00744 (3); conclud: 0.00744 (3); basic: 0.00744 (3); hybrid: 0.00744 (3); optim: 0.00744 (3)
T2	1	Semantic Web/Description Logics	applic: 0.01386 (42); data: 0.01288 (39); semant: 0.01091 (33); object: 0.01058 (32); integr: 0.01058 (32); languag: 0.01058 (32); class: 0.01058 (32); approach: 0.01025 (31); constraint: 0.00992 (30); defin: 0.00926 (28); express: 0.00926 (28); order: 0.00926 (28); concept: 0.00926 (28); logic: 0.00894 (27); represent: 0.00894 (27); implement: 0.00861 (26); support: 0.00828 (25); design: 0.00795 (24); extens: 0.00795 (24); framework: 0.00795 (24); featur: 0.00795 (24); specif: 0.00795 (24); type: 0.00762 (23); schema: 0.00762 (23); includ: 0.00762 (23)

Table 13 (continued)

Period	Index	Topic tag	Top 25 words of each topic
	2	Information Retrieval	retriev: 0.02241 (35); data: 0.01922 (30); larg: 0.01922 (30); method: 0.01731 (27); knowledg: 0.0154 (24); document: 0.01413 (22); discoveri: 0.01349 (21); techniqu: 0.01349 (21); effici: 0.01349 (21); describ: 0.01349 (21); process: 0.01349 (21); mine: 0.01222 (19); discuss: 0.01158 (18); inform: 0.01158 (18); develop: 0.01095 (17); discov: 0.01095 (17); search: 0.01095 (17); studi: 0.01095 (17); index: 0.01031 (16); user: 0.01031 (16); effect: 0.01031 (16); rule: 0.01031 (16); associ: 0.00968 (15); evalu: 0.00904 (14); pattern: 0.00904 (14)
	3	Database Systems/XML Data	number: 0.01474 (22); effect: 0.01474 (22); time: 0.01408 (21); execut: 0.01408 (21); data: 0.01208 (18); share: 0.01142 (17); reduc: 0.01076 (16); achiev: 0.01076 (16); oper: 0.01076 (16); simul: 0.01076 (16); environ: 0.01009 (15); architectur: 0.01009 (15); approach: 0.01009 (15); order: 0.00943 (14); control: 0.00943 (14); parallel: 0.00943 (14); requir: 0.00943 (14); effici: 0.00943 (14); process: 0.00943 (14); improv: 0.00943 (14); exist: 0.00943 (14); increas: 0.00876 (13); join: 0.00876 (13); implement: 0.00876 (13); distribut: 0.0081 (12)
	4	NLS/Statistical Machine Translation	method: 0.02285 (24); network: 0.01624 (17); techniqu: 0.01341 (14); inform: 0.01246 (13); probabilist: 0.01246 (13); simpl: 0.01058 (11); languag: 0.01058 (11); probabl: 0.01058 (11); train: 0.01058 (11); analysi: 0.01058 (11); compar: 0.00963 (10); learn: 0.00963 (10); statist: 0.00963 (10); interpret: 0.00869 (9); effici: 0.00869 (9); knowledg: 0.00869 (9); process: 0.00869 (9); valu: 0.00869 (9); rule: 0.00774 (8); complex: 0.00774 (8); comput: 0.00774 (8); basi: 0.00774 (8); author: 0.00774 (8); neural: 0.00774 (8); number: 0.00774 (8)
	5	Pattern recognition/Image analysis	demonstr: 0.01537 (16); imag: 0.01537 (16); set: 0.01442 (15); comput: 0.01442 (15); transform: 0.01442 (15); space: 0.01442 (15); reduc: 0.01252 (13); select: 0.01252 (13); estim: 0.01252 (13); optim: 0.01157 (12); effici: 0.01157 (12); method: 0.01063 (11); experiment: 0.01063 (11); obtain: 0.01063 (11); solut: 0.00968 (10); point: 0.00968 (10); time: 0.00968 (10); improv: 0.00873 (9); consist: 0.00873 (9); paramet: 0.00873 (9); function: 0.00873 (9); match: 0.00873 (9); compar: 0.00778 (8); linear: 0.00778 (8); dimension: 0.00778 (8)
T3	1	Information Retrieval	data: 0.02251 (57); effici: 0.01818 (46); larg: 0.01818 (46); method: 0.01661 (42); techniqu: 0.01582 (40); mine: 0.01267 (32); retriev: 0.01149 (29); inform: 0.0111 (28); index: 0.0107 (27); number: 0.01031 (26); pattern: 0.00992 (25); applic: 0.00992 (25); evalu: 0.00952 (24); addit: 0.00952 (24); studi: 0.00952 (24); implement: 0.00834 (21); select: 0.00834 (21); associ: 0.00795 (20); rule: 0.00795 (20); improv: 0.00795 (20); demonstr: 0.00795 (20); amount: 0.00756 (19); describ: 0.00716 (18); search: 0.00716 (18); knowledg: 0.00716 (18)

Table 13 (continued)

Period	Index	Topic tag	Top 25 words of each topic
	2	Semantic Web/Description Logics	express: 0.0149 (28); languag: 0.01437 (27); descript: 0.01226 (23); constraint: 0.01226 (23); logic: 0.01173 (22); complex: 0.01173 (22); knowledg: 0.0112 (21); data: 0.01067 (20); concept: 0.01067 (20); inform: 0.01014 (19); defin: 0.01014 (19); reason: 0.01014 (19); represent: 0.01014 (19); order: 0.00961 (18); introduc: 0.00961 (18); form: 0.00909 (17); semant: 0.00909 (17); approach: 0.00856 (16); formal: 0.00803 (15); extend: 0.00803 (15); theori: 0.00803 (15); valu: 0.00803 (15); power: 0.00803 (15); framework: 0.0075 (14); class: 0.0075 (14)
	3	Database Systems/XML Data	applic: 0.02145 (36); data: 0.01967 (33); design: 0.01967 (33); process: 0.01967 (33); approach: 0.01611 (27); implement: 0.01434 (24); develop: 0.01374 (23); user: 0.01197 (20); integr: 0.01197 (20); support: 0.01078 (18); framework: 0.01078 (18); object: 0.01019 (17); view: 0.01019 (17); level: 0.01019 (17); schema: 0.0096 (16); issu: 0.0096 (16); behavior: 0.0096 (16); specif: 0.0096 (16); requir: 0.0096 (16); introduc: 0.009 (15); discuss: 0.009 (15); sourc: 0.009 (15); environ: 0.00841 (14); address: 0.00841 (14); manag: 0.00841 (14)
	4	Pattern recognition/Image analysis	method: 0.02094 (35); imag: 0.02035 (34); reserv: 0.01737 (29); featur: 0.01321 (22); experi: 0.01261 (21); 1999: 0.01202 (20); experiment: 0.01202 (20); applic: 0.01202 (20); recognit: 0.01083 (18); extract: 0.01083 (18); comput: 0.01083 (18); requir: 0.01083 (18); obtain: 0.01023 (17); detect: 0.01023 (17); approach: 0.00964 (16); transform: 0.00964 (16); dimension: 0.00904 (15); effect: 0.00845 (14); discuss: 0.00845 (14); network: 0.00785 (13); real: 0.00785 (13); paramet: 0.00785 (13); function: 0.00785 (13); compar: 0.00785 (13); high: 0.00785 (13)
	5	Machine Learning	learn: 0.02459 (28); network: 0.01848 (21); analysi: 0.01761 (20); train: 0.01587 (18); number: 0.01325 (15); sampl: 0.01325 (15); paramet: 0.01151 (13); case: 0.01151 (13); observ: 0.01064 (12); method: 0.01064 (12); deriv: 0.01064 (12); obtain: 0.00976 (11); comput: 0.00976 (11); error: 0.00976 (11); hidden: 0.00976 (11); length: 0.00976 (11); weight: 0.00889 (10); probabl: 0.00889 (10); estim: 0.00889 (10); function: 0.00889 (10); distribut: 0.00802 (9); random: 0.00802 (9); approxim: 0.00802 (9); statist: 0.00802 (9); variabl: 0.00802 (9)
	6	Web Services	web: 0.03412 (30); user: 0.02056 (18); describ: 0.01831 (16); search: 0.01718 (15); engin: 0.01718 (15); number: 0.01492 (13); page: 0.01492 (13); content: 0.01492 (13); automat: 0.01492 (13); collect: 0.01266 (11); access: 0.01266 (11); design: 0.01266 (11); specif: 0.01266 (11); publish: 0.01153 (10); site: 0.0104 (9); resourc: 0.0104 (9); wide: 0.0104 (9); graph: 0.0104 (9); link: 0.0104 (9); improv: 0.00927 (8); purpos: 0.00927 (8); servic: 0.00927 (8); call: 0.00927 (8); topic: 0.00814 (7); organ: 0.00814 (7)

Table 13 (continued)

Period	Index	Topic tag	Top 25 words of each topic
	7	Web Mining/Information Fusion	data: 0.0156 (11); simul: 0.01421 (10); approach: 0.01421 (10); server: 0.01281 (9); distribut: 0.01281 (9); reduc: 0.01142 (8); load: 0.01142 (8); time: 0.01142 (8); local: 0.01142 (8); larg: 0.01003 (7); scalabl: 0.01003 (7); main: 0.01003 (7); number: 0.01003 (7); term: 0.00864 (6); cach: 0.00864 (6); individu: 0.00864 (6); control: 0.00864 (6); updat: 0.00864 (6); user: 0.00864 (6); record: 0.00864 (6); overhead: 0.00864 (6); detail: 0.00864 (6); set: 0.00864 (6); small: 0.00864 (6); client: 0.00864 (6)
	8	Data Mining/Association Rules	knowledg: 0.02602 (17); inform: 0.02148 (14); data: 0.02148 (14); discoveri: 0.02148 (14); mine: 0.01846 (12); research: 0.01694 (11); field: 0.01543 (10); human: 0.01392 (9); applic: 0.01392 (9); identifi: 0.01392 (9); machin: 0.01392 (9); area: 0.01241 (8); analysi: 0.01241 (8); direct: 0.01089 (7); current: 0.01089 (7); tool: 0.01089 (7); set: 0.01089 (7); studi: 0.00938 (6); world: 0.00938 (6); articl: 0.00938 (6); statist: 0.00938 (6); author: 0.00938 (6); classif: 0.00787 (5); explor: 0.00787 (5); intellig: 0.00787 (5)
	9	NLS/Statistical Machine Translation	studi: 0.01909 (9); differ: 0.01494 (7); rule: 0.01494 (7); sentenc: 0.01286 (6); construct: 0.01286 (6); context: 0.01286 (6); linguist: 0.01286 (6); varieti: 0.01079 (5); part: 0.01079 (5); includ: 0.01079 (5); framework: 0.01079 (5); languag: 0.01079 (5); evalu: 0.00871 (4); tradit: 0.00871 (4); contrast: 0.00871 (4); parallel: 0.00871 (4); wide: 0.00871 (4); examin: 0.00871 (4); make: 0.00871 (4); syntact: 0.00664 (3); discuss: 0.00664 (3); carri: 0.00664 (3); research: 0.00664 (3); robust: 0.00664 (3); program: 0.00664 (3)
T4	1	Machine Learning	method: 0.01778 (57); combin: 0.01374 (44); data: 0.01374 (44); learn: 0.01342 (43); number: 0.01156 (37); classif: 0.01094 (35); optim: 0.01094 (35); classifi: 0.01001 (32); function: 0.0097 (31); compar: 0.0097 (31); class: 0.00876 (28); set: 0.00876 (28); real: 0.00814 (26); studi: 0.00783 (25); error: 0.00783 (25); obtain: 0.00752 (24); approach: 0.00752 (24); select: 0.00752 (24); construct: 0.00721 (23); improv: 0.0069 (22); case: 0.00659 (21); effici: 0.00659 (21); effect: 0.00659 (21); network: 0.00628 (20); featur: 0.00628 (20)
	2	Database Systems/XML Data	inform: 0.01566 (38); applic: 0.01484 (36); data: 0.01443 (35); process: 0.01402 (34); web: 0.01402 (34); describ: 0.0132 (32); develop: 0.01197 (29); user: 0.01115 (27); larg: 0.01115 (27); approach: 0.01033 (25); design: 0.01033 (25); framework: 0.00869 (21); improv: 0.00869 (21); discuss: 0.00869 (21); introduc: 0.00869 (21); time: 0.00828 (20); requir: 0.00787 (19); implement: 0.00787 (19); effici: 0.00746 (18); techniqu: 0.00705 (17); space: 0.00664 (16); sourc: 0.00664 (16); number: 0.00664 (16); studi: 0.00623 (15); content: 0.00623 (15)

Table 13 (continued)

Period	Index	Topic tag	Top 25 words of each topic
	3	Pattern recognition/Image analysis	reserv: 0.03615 (37); imag: 0.02255 (23); object: 0.02157 (22); 1999: 0.01963 (20); approach: 0.01672 (17); comput: 0.01672 (17); experi: 0.01574 (16); compar: 0.01477 (15); method: 0.01186 (12); point: 0.01088 (11); improv: 0.01088 (11); index: 0.01088 (11); dimension: 0.01088 (11); requir: 0.00991 (10); featur: 0.00991 (10); approxim: 0.00991 (10); consist: 0.00894 (9); transform: 0.00894 (9); retriev: 0.00894 (9); techniqu: 0.00894 (9); shape: 0.00797 (8); represent: 0.00797 (8); explor: 0.00797 (8); scheme: 0.00797 (8); detect: 0.00797 (8)
	4	Semantic Web/Description Logics	express: 0.03293 (24); repres: 0.02204 (16); languag: 0.01932 (14); descript: 0.01932 (14); logic: 0.0166 (12); complex: 0.0166 (12); object: 0.01388 (10); represent: 0.01388 (10); comput: 0.01252 (9); defin: 0.01252 (9); inform: 0.01252 (9); integr: 0.01252 (9); specif: 0.01116 (8); constraint: 0.0098 (7); formal: 0.0098 (7); part: 0.00844 (6); call: 0.00844 (6); data: 0.00844 (6); address: 0.00844 (6); limit: 0.00844 (6); equival: 0.00844 (6); introduc: 0.00844 (6); task: 0.00844 (6); version: 0.00844 (6); includ: 0.00844 (6)
	5	Machine Learning	deriv: 0.02371 (15); approxim: 0.02371 (15); comput: 0.02059 (13); iter: 0.01591 (10); approach: 0.01591 (10); learn: 0.01435 (9); term: 0.01435 (9); markov: 0.01279 (8); method: 0.01279 (8); state: 0.01279 (8); maxim: 0.01279 (8); estim: 0.01123 (7); distribut: 0.01123 (7); rule: 0.01123 (7); network: 0.01123 (7); suggest: 0.00967 (6); process: 0.00967 (6); infer: 0.00967 (6); neural: 0.00967 (6); connect: 0.00967 (6); extend: 0.00967 (6); sequenti: 0.00967 (6); simpl: 0.00811 (5); origin: 0.00811 (5); exact: 0.00811 (5)
	6	NLS/Statistical Machine Translation	evalu: 0.02387 (13); text: 0.02206 (12); automat: 0.01483 (8); document: 0.01483 (8); specif: 0.01302 (7); techniqu: 0.01302 (7); knowledg: 0.01302 (7); domain: 0.01121 (6); sentenc: 0.01121 (6); analysi: 0.01121 (6); addit: 0.0094 (5); studi: 0.0094 (5); word: 0.0094 (5); linguist: 0.0094 (5); articl: 0.0094 (5); typic: 0.0094 (5); lexic: 0.0094 (5); achiev: 0.0094 (5); statist: 0.0094 (5); question: 0.0094 (5); program: 0.0094 (5); predic: 0.00759 (4); respect: 0.00759 (4); criteria: 0.00759 (4); context: 0.00759 (4)
	7	Web Services	specif: 0.01582 (6); chang: 0.01582 (6); correct: 0.01327 (5); analyz: 0.01327 (5); workflow: 0.01327 (5); question: 0.01071 (4); local: 0.01071 (4); point: 0.01071 (4); address: 0.01071 (4); singl: 0.01071 (4); busi: 0.01071 (4); global: 0.01071 (4); guarante: 0.01071 (4); dynam: 0.01071 (4); properti: 0.01071 (4); benefit: 0.00816 (3); workload: 0.00816 (3); process: 0.00816 (3); mathemat: 0.00816 (3); affect: 0.00816 (3); initi: 0.00816 (3); comprehens: 0.00816 (3); suitabl: 0.00816 (3); major: 0.00816 (3); rel: 0.00816 (3)

Table 13 (continued)

Period	Index	Topic tag	Top 25 words of each topic
	8	Web Mining/Information Fusion	key: 0.01362 (3); complet: 0.01362 (3); decid: 0.00936 (2); smaller: 0.00936 (2); search: 0.00936 (2); possibl: 0.00936 (2); design: 0.00936 (2); interv: 0.00936 (2); length: 0.00936 (2); join: 0.00936 (2); comput: 0.00936 (2); long: 0.00936 (2); research: 0.00936 (2); approach: 0.00936 (2); valu: 0.00936 (2); aris: 0.00936 (2); adapt: 0.00936 (2); divid: 0.00936 (2); criteria: 0.00936 (2); fundament: 0.00511 (1); factor: 0.00511 (1); time: 0.00511 (1); function: 0.00511 (1); highli: 0.00511 (1); shown: 0.00511 (1)
T5	1	Machine Learning	method: 0.01632 (30); class: 0.01578 (29); combin: 0.01416 (26); learn: 0.01416 (26); experi: 0.01362 (25); number: 0.012 (22); classifi: 0.012 (22); compar: 0.01092 (20); approach: 0.01092 (20); error: 0.01092 (20); function: 0.01038 (19); classif: 0.00984 (18); data: 0.00984 (18); applic: 0.00984 (18); improv: 0.0093 (17); set: 0.00876 (16); studi: 0.00876 (16); train: 0.00876 (16); estim: 0.00822 (15); evalu: 0.00822 (15); featur: 0.00822 (15); optim: 0.00768 (14); real: 0.00768 (14); comparison: 0.00768 (14); sampl: 0.00768 (14)
	2	Pattern recognition/Image analysis	effici: 0.02508 (23); comput: 0.01968 (18); discuss: 0.01643 (15); data: 0.01643 (15); method: 0.01535 (14); reserv: 0.01427 (13); represent: 0.01319 (12); time: 0.01319 (12); appli: 0.01211 (11); approach: 0.01103 (10); number: 0.01103 (10); deriv: 0.01103 (10); implement: 0.00995 (9); probabl: 0.00995 (9); order: 0.00995 (9); defin: 0.00995 (9); detect: 0.00995 (9); relationship: 0.00995 (9); object: 0.00886 (8); independ: 0.00886 (8); approxim: 0.00886 (8); extens: 0.00886 (8); complex: 0.00886 (8); decis: 0.00778 (7); iter: 0.00778 (7)
	3	Web Services	process: 0.02211 (17); describ: 0.01697 (13); servic: 0.01697 (13); applic: 0.0144 (11); execut: 0.0144 (11); design: 0.01311 (10); approach: 0.01311 (10); busi: 0.01311 (10); web: 0.01311 (10); develop: 0.01311 (10); inform: 0.01311 (10); workflow: 0.01183 (9); manag: 0.01183 (9); support: 0.01183 (9); tool: 0.01183 (9); time: 0.01054 (8); implement: 0.01054 (8); basi: 0.00925 (7); interact: 0.00925 (7); rule: 0.00797 (6); satisfi: 0.00797 (6); compon: 0.00797 (6); collect: 0.00797 (6); pattern: 0.00797 (6); oper: 0.00797 (6)
	4	Web Mining/Information Fusion	user: 0.02386 (16); web: 0.02091 (14); analysi: 0.02091 (14); pattern: 0.01944 (13); mine: 0.01797 (12); discoveri: 0.01649 (11); knowledg: 0.01502 (10); page: 0.01502 (10); interest: 0.01208 (8); introduc: 0.01208 (8); method: 0.01208 (8); content: 0.0106 (7); improv: 0.0106 (7); discov: 0.0106 (7); associ: 0.0106 (7); discuss: 0.0106 (7); log: 0.00913 (6); site: 0.00913 (6); number: 0.00913 (6); domain: 0.00913 (6); activ: 0.00913 (6); inform: 0.00913 (6); access: 0.00913 (6); address: 0.00766 (5); util: 0.00766 (5)

Table 13 (continued)

Period	Index	Topic tag	Top 25 words of each topic
	5	Semantic Web/Description Logics	context: 0.0225 (14); languag: 0.01933 (12); repres: 0.01775 (11); describ: 0.01775 (11); case: 0.01616 (10); reason: 0.01616 (10); approach: 0.01616 (10); logic: 0.01458 (9); semant: 0.013 (8); theori: 0.013 (8); ontolog: 0.01141 (7); consist: 0.01141 (7); specif: 0.01141 (7); set: 0.01141 (7); integr: 0.01141 (7); foundat: 0.01141 (7); concept: 0.00983 (6); develop: 0.00983 (6); express: 0.00983 (6); exploit: 0.00983 (6); exampl: 0.00983 (6); definit: 0.00983 (6); singl: 0.00983 (6); restrict: 0.00824 (5); flexibl: 0.00824 (5)
	6	NLS/Statistical Machine Translation	research: 0.02044 (10); learn: 0.01443 (7); text: 0.01443 (7); domain: 0.01242 (6); articl: 0.01242 (6); automat: 0.01242 (6); machin: 0.01242 (6); document: 0.01042 (5); natur: 0.01042 (5); larg: 0.01042 (5); evalu: 0.01042 (5); issu: 0.01042 (5); differ: 0.01042 (5); relev: 0.01042 (5); introduc: 0.00842 (4); languag: 0.00842 (4); sentenc: 0.00842 (4); question: 0.00842 (4); make: 0.00842 (4); measur: 0.00842 (4); identifi: 0.00842 (4); human: 0.00842 (4); current: 0.00842 (4); address: 0.00842 (4); construct: 0.00842 (4)

References

Adomavicius, G., Bockstedt, J. C., Gupta, A., & Kauffman, R. J. (2007). Technology roles and paths of influence in an ecosystem model of technology evolution. *Information Technology Management, 8*(2), 185–202.

Aldous, D. J. (1985). Exchangeability and related topics. *Ecole Dete De Probabilites De Saint Flour, 1117*(3), 1–198.

Alsumait, L., Barbará, D., & Domeniconi, C. (2008). On-Line LDA: Adaptive topic models for mining text streams with applications to topic detection and tracking. In: *Eighth IEEE international conference on data mining*.

Amsler, R. A. (1972). Applications of citation-based automatic classification. Linguistics Research Center, University of Texas at Austin.

Blackwell, D., & Macqueen, J. B. (1973). Ferguson distributions via polya urn schemes. *Annals of Statistics, 1*(2), 353–355.

Blei, D. M., & Lafferty, J. D. (2006). Dynamic topic models. In: *Proceedings of the twenty-third international conference machine learning (ICML 2006)*

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3,* 993–1022.

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D^3 data-driven documents. *IEEE Transactions on Visualization Computer Graphics, 17*(12), 2301–2309.

Braun, T., Glänzel, W., & Schubert, A. (2001). Publication and cooperation patterns of the authors of neuroscience journals. *Scientometrics, 51*(3), 499–510.

Calderone, A., & Cesareni, G. (2018). SPV: a javascript signaling pathway visualizer. *Bioinformatics, 34*(15), 2684–2686.

Callon, M., Courtial, J. P., Turner, W. A., & Bauin, S. (1983). From translations to problematic networks: An introduction to co-word analysis. *Social Science Information, 22*(2), 191–235.

Chang, J., & Blei, D. M. (2010). Hierarchical relational models for document networks. *Annals of Applied Statistics, 4*(1), 124–150.

Chaomei, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the Association for Information Science Technology, 57*(3), 359–377.

Chen, J., Zhang, K., Zhou, Y., Chen, Z., Liu, Y., Tang, Z., et al. (2020). A novel topic model for documents by incorporating semantic relations between words. *Soft Computing, 24*(15), 11407–11423.

Chen, S.-H., Huang, M.-H., & Chen, D.-Z. (2013). Exploring technology evolution and transition characteristics of leading countries: A case of fuel cell field. *Advanced Engineering Informatics, 27*(3), 366–377.

Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). BTM: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering, 26*(12), 2928–2941.

Cohn, D., & Hofmann, T. (2000). The missing link: A probabilistic model of document content and hypertext connectivity. In: *International conference on neural information processing systems*

Cui, W., Liu, S., Tan, L., Shi, C., Song, Y., Gao, Z., et al. (2011). Textflow: Towards better understanding of evolving topics in text. *IEEE Transactions on Visualization Computer Graphics, 17*(12), 2412–2421.

Dai, A. M., & Storkey, A. J. (2009). Author disambiguation: A nonparametric topic and co-authorship model. NIPS workshop on applications for topic models text and beyond.

Ding, W., & Chen, C. (2014). Dynamic topic detection and tracking: A comparison of HDP, C-word, and cocitation methods. *Journal of the Association for Information Science Technology, 65*(10), 2084–2097.

Fu, X., Li, J., Yang, K., Cui, L., & Lei, Y. (2016). Dynamic Online HDP model for discovering evolutionary topics from Chinese social texts. *Neurocomputing, 171,* 412–424.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of National Academy of Sciences, 101*(Suppl 1), 5228–5235.

Grover, A., & Leskovec, J. (2016). node2vec: Scalable feature learning for networks. In: *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*.

Guo, Y., Ma, T., Porter, A. L., & Huang, L. (2012). Text mining of information resources to inform forecasting innovation pathways. *Technology Analysis & Strategic Management, 24*(8), 843–861.

Havre, S., Hetzler, E., Whitney, P., & Nowell, L. (2002). Themeriver: Visualizing thematic changes in large document collections. *IEEE Transactions on Visualization Computer Graphics, 8*(1), 9–20.

He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P., & Giles, L. (2009). Detecting topic evolution in scientific literature: how can citations help? In: *Proceedings of the 18th ACM conference on Information and knowledge management*.

Heberle, H., Carazzolle, M. F., Telles, G. P., Meirelles, G. V., & Minghim, R. (2017). CellNetVis: A web tool for visualization of biological networks using force-directed layout constrained by cellular components. *BMC Bioinformatics, 18*(10), 395.

Heinrich, G. (2005). Parameter estimation for text analysis, Technical report.

Hofmann, T. (1999). Probabilistic latent semantic analysis. In: *Fifteenth conference on uncertainty in artificial intelligence*.

Huang, Y., Zhu, F., Guo, Y., Porter, A. L., Zhang, Y., & Zhu, D. (2016). Exploring technology evolution pathways to facilitate technology management: A study of dye-sensitized solar cells (DSSCs). In: *2016 Portland international conference on management of engineering and technology (PICMET)*.

Huang, Y., Zhu, F., Porter, A. L., Zhang, Y., Zhu, D., & Guo, Y. (2020). Exploring technology evolution pathways to facilitate technology management: From a technology life cycle perspective. *IEEE Transactions on Engineering Management, PP*(99), 1–13.

Jeh, G., & Widom, J. (2002). SimRank: A measure of structural-context similarity. In: *Eighth ACM Sigkdd international conference on knowledge discovery & data mining*.

Jeong, D. H., & Min, S. (2014). Time gap analysis by the topic model-based temporal technique. *Journal of Informetrics, 8*(3), 776–790.

Jie, T., Jing, Z., Yao, L., Li, J., Li, Z., & Zhong, S. (2008). ArnetMiner:extraction and mining of academic social networks. In: *ACM Sigkdd intersnational conference on knowledge discovery & data mining*.

Kajikawa, Y., Ohno, J., Takeda, Y., Matsushima, K., & Komiyama, H. (2007). Creating an academic landscape of sustainability science: An analysis of the citation network. *Sustainability Science, 2*(2), 221–231.

Kataria, S., Mitra, P., & Bhatia, S. (2010). Utilizing context in generative Bayesian models for linked corpus. In: *Twenty-fourth AAAI conference on artificial intelligence*.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation, 14*(1), 10–25.

Kim, M., Baek, S. H., & Song, M. (2018). Relation extraction for biological pathway construction using node2vec. *BMC Bioinformatics, 19*(Suppl 8), 206.

Kim, J., & Shin, J. (2018). Mapping extended technological trajectories: Integration of main path, derivative paths, and technology junctures. *Scientometrics, 116*(3), 1439–1459.

Kong, D., Zhou, Y., Liu, Y., & Xue, L. (2017). Using the data mining method to assess the innovation gap: A case of industrial robotics in a catching-up country. *Technological Forecasting & Social Change, 119*.

Li, C., Wang, H., Zhang, Z., Sun, A., & Ma, Z. (2016a). Topic modeling for short texts with auxiliary word embeddings. In: *Proceedings of the 39th international Acm sigir conference on research and development in information retrieval—SIGIR '16*, pp. 165–174

Li, X., Zhou, Y., Xue, L., & Huang, L. (2015). Integrating bibliometrics and roadmapping methods: A case of dye-sensitized solar cell technology-based industry in China. *Technological Forecasting and Social Change, 97*, 205–222.

Li, X., Zhou, Y., Xue, L., & Huang, L. (2016b). Roadmapping for industrial emergence and innovation gaps to catch-up: A patent-based analysis of OLED industry in China. *International Journal of Technology Management, 72*(1/2/3), 105.

Li, Y., Li, Y., Wang, J., & Sherratt, R. S. (2020). Sentiment analysis for E-commerce product reviews in Chinese based on sentiment lexicon and deep learning. *IEEE Access, 8*(1), 23522–23530.

Liu, Y., Wang, J., & Jiang, Y. (2016). PT-LDA: A latent variable model to predict personality traits of social network users. *Neurocomputing, 210,* 155–163.

Liu, Y., Zhou, Y., Liu, X., Dong, F., Wang, C., & Wang, Z. (2019). Wasserstein gan-based small-sample augmentation for new-generation artificial intelligence: A case study of cancer-staging data in biology. *Engineering*, *2019*(5), 156–163.

Malik, S., Smith, A., Hawes, T., Papadatos, P., Li, J., Dunne, C., & Shneiderman, B. (2013). TopicFlow: Visualizing topic alignment of Twitter data over time. In: *Proceedings of the 2013 IEEE/ACM international conference on advances in social networks analysis and mining*.

Mccallum, A., Wang, X., & Corrada-Emmanuel, A. (2007). Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research, 30*(2), 249–272.

Miao, Z., Du, J., Dong, F., Liu, Y., & Wang, X. (2020). Identifying technology evolution pathways using topic variation detection based on patent data: A case study of 3D printing. *Futures, 118*, 102530.

Ming, Y., & Hsu, W. H. (2016). HDPauthor: A new hybrid author-topic model using latent dirichlet allocation and hierarchical dirichlet processes. In: International conference companion on world wide web.

Nallapati, R. M., Ahmed, A., Xing, E. P., & Cohen, W. W. (2008). Joint latent topic models for text and citations. In: *ACM Sigkdd international conference on knowledge discovery & data mining*.

Nordensvard, J., Zhou, Y., & Zhang, X. (2018). Innovation core, innovation semi-periphery and technology transfer: The case of wind energy patents. *Energy Policy*, *120*, 213–227.
Pan, M., Zhou, Y., & Zhou, D. (2019). Comparing the innovation strategies of Chinese and European wind turbine firms through a patent lens. *Environmental Innovation and Societal Transitions, 30*, 6–18.
Perozzi, B., Al-Rfou, R., & Skiena, S. (2014). Deepwalk: Online learning of social representations. In: *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 701–710
Rosen-Zvi, M., Griffiths, T. L., Steyvers, M., & Smyth, P. (2012). The author-topic model for authors and documents. In: *Conference on uncertainty in artificial intelligence.*
Small, H. (1973). Co-citation in the scientific literature: A new measure of the relationship between two documents. *Journal of the American Society for information Science, 24*(4), 265–269.
Steyvers, M., Smyth, P., Rosen-Zvi, M. & Griffiths, T. (2004). Probabilistic author-topic models for information discovery. In: *Tenth Acm Sigkdd international conference on knowledge discovery & data mining*.
Teh, Y. W., Jordan, M. I., Beal, M. J., & Blei, D. M. (2006). Hierarchical dirichlet processes. *Publications of the American Statistical Association, 101*(476), 1566–1581.
Wang, B., Liu, Y., Zhou, Y., & Wen, Z. (2018). Emerging nanogenerator technology in China: A review and forecast using integrating bibliometrics, patent analysis and technology roadmapping methods. *Nano Energy, 46*, 322–330.
Wei, C., Chaoran, L., Chuanyun, L., Lingkai, K., & Zaoli, Y. (2020). Tracing the evolution of 3-D printing technology in China using LDA-based patent abstract mining. *IEEE Transactions on Engineering Management, PP*, 1–14.
Wu, Y., Liu, S., Yan, K., Liu, M., & Wu, F. (2014). Opinionflow: Visual analysis of opinion diffusion on social media. *IEEE Transactions on Visualization Computer Graphics, 20*(12), 1763–1772.
Xiao, Y., Lu, L. Y., Liu, J. S., & Zhou, Z. (2014). Knowledge diffusion path analysis of data quality literature: A main path analysis. *Journal of Informetrics, 8*(3), 594–605.
Xu, H. (2020). Topic-linked innovation paths in science and technology. *Journal of Informetrics, 14*(2), 101014.
Xu, G., Hu, W., Qiao, Y., & Zhou, Y. (2020). Mapping an innovation ecosystem using network clustering and community identification: A multi-layered framework. *Scientometrics, 124*, 2057–2081. https://doi.org/10.1007/s11192-020-03543-0.
Xu, G., Wu, Y., Minshall, T., & Zhou, Y. (2017). Exploring the emerging ecosystem across science, technology and business: A case of 3D printing in China. *Technological Forecasting and Social Change*. https://doi.org/10.1016/j.techfore.2017.06.030.
Yao, Q., Song, Z., & Peng, C. (2011). Research on text categorization based on LDA. *Computer Engineering Applications, 47*(13), 150–153.
Yau, C. K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics, 100*(3), 767–786.
Yu, J. (2011). From 3G to 4G: Technology evolution and path dynamics in China's mobile telecommunication sector. *Technology Analysis Strategic Management, 23*(10), 1079–1093.
Zhang, Y., Zhang, G., Chen, H., Porter, A. L., Zhu, D., & Lu, J. (2016). Topic analysis and forecasting for science, technology and innovation: Methodology with a case study focusing on big data research. *Technological Forecasting Social Change, 105,* 179–191.
Zhao, P., Han, J., & Sun, Y. (2009). P-Rank: A comprehensive structural similarity measure over information networks. In: *ACM conference on information & knowledge management*.
Zhou, Y., & Minshall, T. (2014). Building global products and competing in innovation: The role of Chinese university spin–outs and required innovation capabilities. *International Journal of Technology Management, 64*(2), 180–209.
Zhou, Y., Dong, F., Kong, D., & Liu, Y. (2019b). Unfolding the convergence process of scientific knowledge for the early identification of emerging technologies. *Technological Forecasting and Social Change, 144*(JUL.), 205–220.
Zhou, Y., Dong, F., Liu, Y., Li, Z., Du, J., & Zhang, L. (2020). Forecasting emerging technologies using data augmentation and deep learning. *Scientometrics, 123*(1), 1–29.
Zhou, Y., Li, X., Lema, R., & Urban, F. (2016). Comparing the knowledge bases of wind turbine firms in Asia and Europe: Patent trajectories, networks, and globalisation. *Science and Public Policy*, *43*(4), 476–491. https://doi.org/10.1093/scipol/scv055.
Zhou, Y., Lin, H., Liu, Y., & Ding, W. (2019a). A novel method to identify emerging technologies using a semi-supervised topic clustering model: A case of 3d printing industry. *Scientometrics, 120*, 167.
Zhou, Y., Pan, M., & Urban, F. (2018). Comparing the international knowledge flow of china's wind and solar photovoltaic (pv) industries: Patent analysis and implications for sustainable development. *Sustainability, 10*(6), 1883.