Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

Topic discovery and evolution in scientific literature based on content and citations*

Hou-kui ZHOU[1,3,4], Hui-min YU[†‡1,2], Roland HU[1]

([1]*College of Information Science & Electronic Engineering, Zhejiang University, Hangzhou 310027, China*)

([2]*State Key Lab of CAD & CG, Zhejiang University, Hangzhou 310027, China*)

([3]*School of Information Engineering, Zhejiang A&F University, Lin'an 311300, China*)

([4]*Zhejiang Provincial Key Laboratory of Forestry Intelligent Monitoring and Information Technology, Lin'an 311300, China*)

[†]E-mail: yhm2005@zju.edu.cn

Received Apr. 1, 2016; Revision accepted June 30, 2016; Crosschecked Sept. 22, 2017

Abstract: Researchers across the globe have been increasingly interested in the manner in which important research topics evolve over time within the corpus of scientific literature. In a dataset of scientific articles, each document can be considered to comprise both the words of the document itself and its citations of other documents. In this paper, we propose a citation-content-latent Dirichlet allocation (LDA) topic discovery method that accounts for both document citation relations and the content of the document itself via a probabilistic generative model. The citation-content-LDA topic model exploits a two-level topic model that includes the citation information for 'father' topics and text information for sub-topics. The model parameters are estimated by a collapsed Gibbs sampling algorithm. We also propose a topic evolution algorithm that runs in two steps: topic segmentation and topic dependency relation calculation. We have tested the proposed citation-content-LDA model and topic evolution algorithm on two online datasets, *IEEE Transactions on Pattern Analysis and Machine Intelligence* (PAMI) and IEEE Computer Society (CS), to demonstrate that our algorithm effectively discovers important topics and reflects the topic evolution of important research themes. According to our evaluation metrics, citation-content-LDA outperforms both content-LDA and citation-LDA.

Key words: Topic extraction; Topic evolution; Evaluation method
https://doi.org/10.1631/FITEE.1601125 **CLC number:** TP391

1 Introduction

Scientific literature records and supports the continuing progress of research within a wide variety of domains. As a researcher approaches a new research area, he/she benefits from a thorough knowledge of hot topics related to that field and the evolution of those topics throughout the body of literature, particularly when this knowledge can be gathered quickly and conveniently. The body of scientific literature is exceedingly difficult to navigate; however, as an increasing number of documents (even documents of dubious veracity) are readily available through digital databases and other online sources, automatic topic discovery and evolution represent an interesting and very potentially beneficial solution to this problem.

In recent decades, topic models have provided a simple way to analyze large volumes of unlabeled document collections. Among these models, probabilistic latent semantic analysis (PLSA) (Hofmann, 2001) and latent Dirichlet allocation (LDA) (Blei

‡ Corresponding author

* Project supported by the National Basic Research Program (973) of China (No. 2012CB316400)

Electronic supplementary materials: The online version of this article (https://doi.org/10.1631/FITEE.1601125) contains supplementary materials, which are available to authorized users

ORCID: Hou-kui ZHOU, http://orcid.org/0000-0001-7915-8684

et al., 2003) are the most popular. Both of them are based fairly simply on probability analysis and graphical models, exploiting the 'bag of words' concept to discover topics from documents. The topic model combined with the timestamp information can reflect the evolution of the topic over time, i.e., providing topic evolution analysis. Topic discovery and evolution analysis for scientific literature have attracted a considerable amount of research interest in terms of data mining and discovery. At present, there are two roadmaps through the corpus of scientific research: one uses only the text information of scientific documents for topic discovery and evolution research (Blei and Lafferty, 2006; Wang and McCallum, 2006; Ahmed and Xing, 2010), and the other combines the text information with additional information, such as similarities in citations among papers, to detect topics and track topic evolution over time. The latter model clearly uses more information from the literature to detect topics and their evolution, and thus tends to return better results. The inheritance topic model developed by He *et al.* (2009), for example, uses the citation information among papers to represent topic evolution in scientific literature. The model works on the principle that ideas and techniques inherited among papers are reflected by the citations shared among them. Nallapati *et al.* (2008) and Guo *et al.* (2014) have combined the content and citations of scientific documents for the purposes of topic modeling frameworks. Lu *et al.* (2014) proposed a topic model which uses authorship, published venues, and citation relations among scientific documents to detect topics and identify the most notable works in the corpus. Though effective in certain regards, all of these methods focus on only one or two aspects of topic discovery and evolution research. Wang *et al.* (2013) explored several aspects of topic and theme evolution by modeling document citations with a probability generative model. Their technique is the inspiration for the existing study, and we take it one step further by building a comprehensive topic modeling framework that fully combines text and citation information to represent topics and topic evolution in the corpus.

Unlike the first-published citation-LDA topic detection model, the model we propose here explores topics in a scientific corpus in terms of both text (i.e., the document itself) and citations in the documents through a probability generative model. The main idea is that a collection of scientific documents can be considered not only a 'bag of words' but also a 'bag of citations'. Extracting topics and identifying their evolution through both aspects of the documents return more accurate results. We develop the proposed method to extract research topics as accurately as possible, and to analyze topic evolution in the literature without discretizing the corpus in advance. Similar to the method proposed by Wang *et al.* (2013), our model can extract research topics, particularly groundbreaking (i.e., 'milestone') papers, and topic keywords through a two-level (text and citation) generative topic model. The contributions of this work can be summarized as follows:

1. We propose a new probability generative model which uses a two-level LDA based on text and citation information. The model exploits both the bag of words and the bag of citations; thus, any document in the corpus can be viewed as a collection of words or a collection of citations, as required.

2. We address the topic evolution analysis problem by building an algorithm that combines topic segmentation and topic dependency relation calculations. The most notable characteristics of the proposed algorithm are that the topics extracted from the entire corpus are projected into different time slots, and topic dependency relations are established by a random walk process.

2 Related work

In this section, we will review previous work related to topic extraction and evolution for scientific documents based on both text information and link network (i.e., citation) information.

Probabilistic hypertext-induced topic selection (PHITS) (Cohn and Chang, 2000) was the first topic model developed to merge the term-based PLSA model and citation-based analysis into a joint probabilistic model. It assumes that documents and topics are all generated by a probability distribution of both words and citations, where the two distributions share the same document–topic (or 'doc–topic') mixing proportions. A similar model called the 'link LDA model' was developed by Erosheva *et al.* (2004) by adopting a mixed membership model for words and

citations but treating the membership scores as random Dirichlet realizations in the documents. Later, Dietz *et al.* (2007) proposed the citation influence model, which integrates text and links into a probabilistic model and infers the influences of citations on the topic distribution. Nallapati and Cohen (2008) established the link-PLSA-LDA model, which is a very scalable LDA model for topic modeling and link prediction. Chang and Blei (2009) proposed the relational topic model by describing the links among documents via a binary random variable, abandoning the exchangeability assumption of documents in the LDA topic model. The NetPLSA model proposed by Mei *et al.* (2008) combines other statistical topic models under discrete regularization based on a graph structure. It can be applied to text mining tasks, such as author–topic analysis, community discovery, and spatial text mining. The collective topic model (CTM) proposed by Lu *et al.* (2014) simultaneously discovers topics and related milestone papers in the corpus by modeling papers, authors, and published venues as a bag of citations based on the PLSA model.

Topic discovery and evolution analysis are often solved together within a unified framework, and topics are often modeled by combining text and citation information. This integrated approach to topic modeling has made it easier and altogether more effective to identify topic evolution in scientific literature (Nallapati *et al.*, 2008). He *et al.* (2009) developed a particularly interesting citation-aware approach to the topic evolution problem by proposing the inhertance topic model (ITM), which applies the LDA model to the citation network and uses an iterative topic evolution learning framework. Wang *et al.* (2013) proposed a novel topic evolution model called the 'citation-LDA' model, which likewise jointly analyzes text and citations in scientific documents. Citation-LDA was designed to identify milestone papers and evaluate topic dependency via citation information to return more accurate topic evolution results and construct better evolution graphs than the traditional LDA model.

To the best of our knowledge, the approaches mentioned above can solve only part of the questions discussd here, i.e., the discovery of research topics (including milestone papers and key words of topics) and the construction of a topic evolution graph. In this study, we jointly model the generation of the citation and content information to address all the problems discussed above.

3 Probabilistic modeling of literature content and citations

The original LDA model, which is based on the co-occurrence of certain words, can also be called the 'content-LDA model'. There are a handful of issues inherent to the use of the content-LDA model to extract topics from scientific literature. First, the model is not suitable for short documents, as it must synthesize a large number of topics to perform well. Shorter papers may provide only a title and an abstract effectively. In addition, any large amount of noise in scientific documents due to background information or topic-irrelevant words impedes the performance of the content-LDA model. The citation-LDA model, conversely, uses only citation link relations to generate topics in a probability framework, similar to the LDA model. This method has two distinct advantages: (1) In a scientific corpus, link information contains less noise than content information. Although citation information may indeed contain some 'noise' (MacRoberts and MacRoberts, 1989), i.e., when an author takes advantage of another's work, he or she has been influenced by that work. (2) Compared with using content-LDA, computational complexity is greatly reduced using citation-LDA, because there are far fewer citations than words in the literature. However, the citation relations among scientific papers are fairly sparse. In a typical corpus, there are fewer authors cited and citing in any document. Also, more recent papers (or much older papers) may not be cited (or have citations) at all. Certain topics extracted via the citation-LDA model thus lose generality and representativeness. Citation information of papers will help find important topics and key papers of topics. In our model, the citation-LDA process is similar to a clustering process, which is responsible for the discovery of father topics. The number of topics is usually significantly smaller than the number of citation papers. The sparse citation information of papers will not influence the discovery of topics. On the contrary, some citation-unrelated papers that are not included in topics will be discharged as noise information.

We consider the above characteristics of citation-LDA and content-LDA carefully in developing the topic model we propose, aptly named 'citation-content-LDA'. Our model combines the text and citation information from the literature to identify both 'father' topics and sub-topics. Fig. 1 shows the diagram of the proposed model.

3.1 Citation-content-LDA topic model

In a corpus of scientific literature, any document can be considered not only a bag of words but also a bag of citations. We can use citation information to extract father topics at the first level of the document (Wang *et al.*, 2013). In this step, the extracted topics represent a cluster of documents, making the first part of the first step of our proposed method essentially a document clustering process.

A content-LDA model is then applied to extract sub-topics from each of the father topics. Each document d in a corpus cites a group of other documents $\{c_t\}$ $(t=1, 2, \ldots)$; thus, similar to the LDA model, we can suppose that each document d obeys a probability distribution $D_{\text{doc–topic}}$ over latent variables z (i.e., topics), while each topic from $\{z_k\}$ $(k=1, 2, \ldots)$ obeys a probability distribution $D_{\text{topic–doc}}$ over a group of documents cited by document d. We can also suppose that document–topic distribution $D_{\text{doc–topic}}$ and topic–doc distribution $D_{\text{topic–doc}}$ are multinomial distributions with Dirichlet parameters α and β drawn priori. Here, we use θ_d and φ_z to denote $D_{\text{doc–topic}}(;\, d)$ and $D_{\text{topic–doc}}(;\, z)$, respectively, where θ_d~Dir(α) and φ_z~Dir(β). An inference is necessary for obtaining model parameters θ_d and φ_z via the collapsed Gibbs sampling algorithm (Griffiths and Steyvers, 2004).

Through this first level of topic modeling, we obtain topic–doc probability distribution $\{\varphi_{k,\,j}\}$, i.e., the model-wide father topic. This topic–doc distribution indicates the importance of a single paper document d_j in terms of father topic z_k. In the second level, we apply a word-based LDA model to extract sub-topics from each father topic generated in the first level. As opposed to a standard word-based LDA model, documents in father topics from the first level of our model are not equiprobable—each document instead has a different probability over the father topic.

In the second level, each document gathered based on the father topic in the first level is assumed to obey a multinomial probability distribution with parameter θ'_d over sub-topic variables. Each sub-topic also obeys a multinomial probability distribution over words with parameter β. Once the collapsed Gibbs sampling algorithm is applied to infer the model parameters of doc–topic probability distribution $p(z'|d)$ and topic–word probability distribution $p(w|z')$, final sub-topics $p(z'|d)$ in the second level are acquired. Briefly, our topic generative process for documents in the corpus is shown in Algorithm 1.

Algorithm 1 Topic generative process for documents in the corpus

1 Sample a topic z_k~Multi(θ_i);
2 Sample a document to cite c_{d_j}~Multi(φ_{z_k});
3 **for** word w_n of document d in topic z_k
4 Choose a sub-topic $z'_{d,\,n}$~Multinomial(θ'_d);
5 Choose a word $w_{d,\,n}$ from $p(w_{d,\,n}|z'_{d,\,n}, \varphi_{z_k}, \beta)$, a multinomial probability conditioned on topic z_k and sub-topic $z'_{d,\,n}$;
6 **end for**

A graphical representation of this generative process for each document in the corpus is shown in Fig. 2.

3.2 Parameter estimation and inference

The parameters of the proposed model can be inferred, by the collapsed Gibbs sampling algorithm.

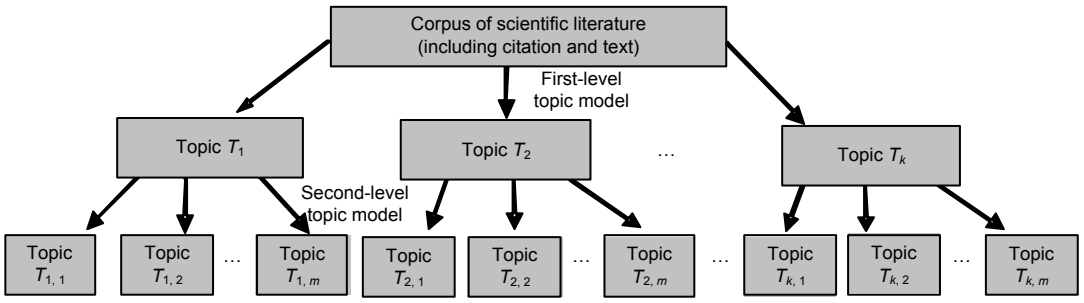

Fig. 1 Framework of a two-level citation-content-LDA topic model (LDA: latent Dirichlet allocation)

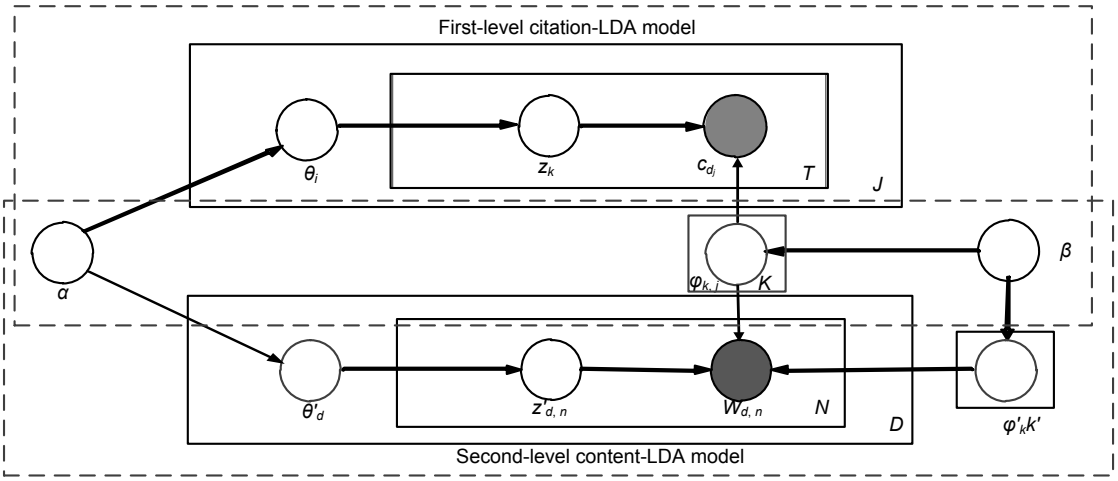

Fig. 2 Graphic of our citation-content-LDA model

We first sample the topic–doc distribution of the father topics (citation-LDA) at the first level, and then sample the topic–word distribution of the sub-topics for each father topic (content-LDA) at the second level. The sampling algorithms are then initialized by assigning random topic labels $\{z_k\}$, and then updating each of them repetitively. To be precise, for each document d_i and its tth citation of document d_j in the corpus, the topic–doc probability is computed as follows:

$$p\left(z_k \mid c_{i,t}=d_j, Z_{\neg(i,t)}, C_{\neg(i,t)}\right) \propto \left(\alpha_k + n_{z_k\neg(d_i)}^{(d_i)}\right)\frac{\beta_j + n_{z_k\neg(i,t)}^{(c_{d_j})}}{\sum_j\left(\beta_j + n_{z_k\neg(i,t)}^{(c_{d_j})}\right)}, \quad (1)$$

where $\neg(i,t)$ indicates that tokens i and t are excluded from the corresponding documents and citations, and $n_{z_k}^{(d_i)}$ refers to the number of citations linked to document d_i belonging to topic z_k. The probability converges to a stationary state of probability distribution after the burn-in stage. Posterior expectations of $\theta_{i,k}$ and $\varphi_{k,j}$ are given by

$$\theta_{i,k} = \frac{n_{d_i}^{(z_k)} + \alpha_k}{\sum_k\left(n_{d_i}^{(z_k)} + \alpha_k\right)}, \quad (2)$$

$$\varphi_{k,j} = \frac{n_{z_k}^{(c_{d_j})} + \beta_j}{\sum_j\left(n_{z_k}^{(c_{d_j})} + \beta_j\right)}. \quad (3)$$

For the second-level topic model, for each word $\{w_t\}$ (t=1, 2, …, V) and each sub-topic $\{z'_k\}$ (k=1, 2, …, K) in each corpus composed of documents from a father topic of topic–document distribution $\{\varphi_{k,m}\}$ (k=1, 2, …, K; m=1, 2, …, M), the sub-topic assignment is computed as follows:

$$p(z'_i = k' \mid z'_{\neg i}, \varphi_{k,\mathrm{m}}, w') \propto \frac{n_{k',\neg i}^{t} + \beta_t}{\sum_{t=1}^{V}\left(n_{k',\neg i}^{t} + \beta_t\right)} \cdot \frac{n_{m,\neg i}^{k'} \cdot \varphi_{k,m} + \alpha_{k'}}{\left[\sum_{k'=1}^{K}\left(n_{m,\neg i}^{k'} \cdot \varphi_{k,m} + \alpha\right)\right] - 1}, \quad (4)$$

where counts $n_{*,\neg i}^{(*)}$ indicate that token i is excluded from the corresponding document or topic.

Again, the sampling converges to a stationary state of probability distribution after the burn-in stage. Finally, multinomial parameter sets $\varphi'_{k',t}$ and $\theta'_{m,k'}$ are obtained by

$$\varphi'_{k',t} = \frac{n_{k'}^{(t)} + \beta_t}{\sum_{t=1}^{V}\left(n_{k'}^{(t)} + \beta_t\right)}, \quad (5)$$

$$\theta'_{m,k'} = \frac{n_{m}^{(k')} + \alpha_{k'}}{\sum_{k'=1}^{K'}\left(n_{m}^{(k')} + \alpha_{k'}\right)}, \quad (6)$$

where $n_{k'}^{(t)}$ denotes the number of times that term t has been observed with topic $z_{k'}$, and $n_{m}^{(k')}$ refers to the number of times topic $z_{k'}$ has been observed with a word of document m.

4 Construction of the topic evolution graph

In this section, we discuss the topic evolution graph which is constructed differently from those of previous studies. The distinguishing characteristic of our method is that topic extraction and the construction of the topic evolution graph take place in two separate processes. We consider the time information of the documents not during topic extraction, but when constructing the topic evolution graph. Graph construction is also a two-step process: first comes topic segmentation, and then establishment of the topic evolution graph.

4.1 Topic segmentation

In Section 3, we discuss topic extraction from the corpus without considering the time information of the documents. Each topic covers a time range throughout the corpus, and segmenting the topics according to the time information is necessary to ensure accurate end results. We used the collapsed Gibbs sampling algorithm to obtain the distribution of topics, including topic–doc distribution $p(z|d)$: $\theta'_{m,k'}$ and topic–word distribution $p(w|z)$: $\varphi'_{k',t}$, but neither of them contains any time-related information.

We can identify milestone papers and calculate the probability that each topic fits within these topic distribution parameters by

$$p(d\mid z)=\frac{p(d,z)}{p(z)}=\frac{p(z\mid d)\cdot p(d)}{p(z)} \propto p(z\mid d)\cdot p(d)\propto \theta'_{m,k'}\cdot n_d, \tag{7}$$

where $p(z)$ refers to the prior topic probability distribution over documents of the whole corpus, and $p(d)$ refers to the prior probability of document d. The former is the same for all documents and the latter can be represented by n_d which refers to the word count of document d.

Once doc–topic probability $p(d|z)$ is identified, we can identify not only the milestone papers of each topic according to the order of probability, but also the time information for these milestone papers. In other words, each topic can also be viewed as a bag-of-documents model and each topic can be divided based on the time information of its documents. We propose the simple topic segmentation scheme (Fig. 3), which divides the documents of each topic into sub-topics in each time slot. The start and end times of the corpus are denoted as s_0 and s_n, respectively, and the number of time slots as Q; thus, the time interval for each time slot is seg=$(s_n-s_0+1)/Q$. Supposing the number of topics is k, we can acquire $K\cdot Q$ sub-topics after topic segmentation.

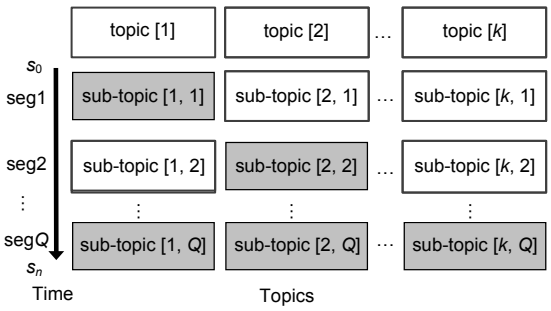

Fig. 3 A simple topic segmentation scheme

4.2 Topic evolution graph establishment

To establish the topic evolution graph, the relation between two sub-topics in adjacent time slots should be measured first. Kullback–Leibler-divergence (KL-divergence) is a typical approach used to measure the topic transition distance in topic evolution models (Mei and Zhai, 2005). It is in effect a measure of topic dependency relations. We borrowed from the KL-divergence approach here, because the relation between two topics is considered to be the influence of one topic on another.

There are several methods for measuring influence via directed-weighted graphs (e.g., web link analysis and social network analysis). Influences in these methods are assumed to propagate through the edges of the graph. Among these methods, PageRank, which employs the random walk concept (Brin and Page, 1998), is applied most often to link analysis in web page ranking applications. Similar to PageRank, we adopt the random walk concept to calculate the influence between topic pairs.

There are no actual edges between any two topics, however. In our topic model, topics are composed of words. Thus, we can establish the relation between two topics according to the co-occurrence of words. First, we can build a bipartite directed graph $G=(V, E)$, where V represents the set of vertices (which correspond to words or topics), and E represents the set of edges. For each word w belonging to topic z, we

add both edges (w, z) and (z, w). In the simple graph shown in Fig. 4, there are two topics (indicated by squares) and four words (indicated by circles). Edge weights α_{ij} between topic z_i and word w_j represent the topic–word distribution probability $p(w_j|z_i)$.

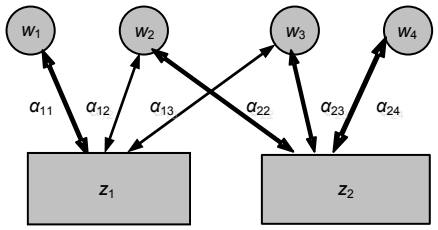

Fig. 4 Bipartite-directed graph of topics and words

Here, we define the relation between two topics as relation(z_i, $z_j|w$), which should be high if the two topics are related closely, and w plays a key role in this relation. The bipartite-directed graph of topics and words can be considered a representation of the topic–word relation graph (Fig. 4). Intuitively, if the two topics are related, a short path starting from z_i should reach z_j frequently. We calculate the probability distribution for random walks starting from z_i to complete the graph, where the probability distribution is the proportion of the time the walker spends on each node:

$$p_{z_i}(r) = \mu \cdot \delta_r(z_i) + (1-\mu) \cdot \sum_{(s,r)\in E} p_{z_i}(s) \cdot p(r|s), \quad (8)$$

where μ controls the restart probability of one random walk, $p(r|s)$ is the probability of reaching r from s, and $\delta_r(z_i)$ is denoted as follows:

$$\delta_r(z_i) = \begin{cases} 1, & r = z_i, \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

Next, consider the effect of w on these walks in the bipartite-directed graph. If we change w into a sink node, $p^w(r|s)$ has the same probability as $p(r|s)$; however, there is no way out of node w. The probability distribution for this new graph is $p^w_{z_j}(r)$. The relation between z_i and z_j with regard to w is defined as the difference between these two probability distributions, $p_{z_i}(z_j) - p^w_{z_i}(z_j)$. Thus, we have

$$\text{relation}(z_i, z_j|w) = p_{z_i}(z_j) - p^w_{z_i}(z_j). \quad (10)$$

Thus, the total relation between z_i and z_j is computed as follows:

$$\text{relation}(z_i, z_j) = \sum_w \text{relation}(z_i, z_j|w). \quad (11)$$

Briefly, constructing the topic evolution graph proceeds in the following steps: (1) calculate the strength of relation between any two topics of adjacent time slots via Eq. (11); (2) connect topic pairs with a strength of relation equal to or greater than the given threshold τ in advance. The evolution graph of the dataset is then complete.

5 Experiments and results

We first formally described the two datasets (CS and PAMI) we used to demonstrate our citation-content-LDA technique, then evaluated the results of our topic discovery and evolution method in detail, and compared our citation-content-LDA with conventional content-LDA and citation-LDA baselines according to two evaluation metrics: perplexity and symmetric KL (sKL) divergence.

5.1 Datasets

We used two scientific literature datasets (both are available online) to test our method: the IEEE Computer Society (CS) scientific literature dataset, which contains documents related to artificial intelligence, pattern recognition, image processing, data mining, and other computer science fields; the PAMI dataset, which contains work from the journal of *IEEE Transactions on Pattern Analysis and Machine Intelligence*. The CS dataset comprises a total of 42 213 papers published between 1967 and 2006, including citing and cited papers, from 67 venues with 33 961 citations (including only cited works also within the dataset). The PAMI dataset contains all papers published in *IEEE Transactions on Pattern Analysis and Machine Intelligence* from 1995 to September, 2012, with a total of 2719 papers and 6284 citations (again, including only cited works within the dataset). All papers in both datasets include the paper ID, title, venue, publication year, and

citation information. After text pre-processing for the title (including stopword removal and filtering words which appear fewer than five times in the dataset), we acquired 5704 unique words for the CS dataset and 886 for the PAMI dataset.

5.2 Scientific topic detection results

It is challenging to select the number of topics appropriately for topic modeling. In our experiments, we ran citation-content-LDA with 100 topics in CS and 30 in PAMI, first according to the characteristics of the dataset itself (topic number selection process is further discussed in Section 5.4). Other parameters in our topic detection model included: the number of topics in the first-level citation-LDA model (10 topics for both the CS and PAMI datasets) and hyper-parameters α=0.5 and β=0.01.

5.2.1 Finding milestone papers

Milestone papers for two typical topics, 'handwritten character recognition' from PAMI and '3D reconstruction' from CS, are presented in Tables 1 and 2, respectively. Topic–doc probability φ_{kj} for these two topics and the venue/journal sources for the second topic are also included. Here, we consider milestone papers to be as typical of the topic as possible and as widely accepted by the academic community as possible with respect to the topic. In Table 1, all milestone papers listed belong to the 'handwritten character recognition' topic from PAMI, except the first and the fifth papers. The first paper contains 'street name recognition' within one line of text and the fifth paper involves 'writing identification' from uppercase western script. These two papers still

Table 1 Top 10 highest impact papers on topic 'handwritten character recognition' (topic 10, PAMI)

Topic–doc probability	Paper title
0.014 975	A statistical approach for phrase location and recognition within a text line: an application to street name recognition
0.013 723	Offline recognition of unconstrained handwritten texts using HMMs and statistical language models
0.013 041	Lexicon-driven segmentation and recognition of handwritten character strings for Japanese address reading
0.012 983	Off-line handwritten Chinese character recognition as a compound Bayes decision problem
0.012 788	Automatic writer identification using connected-component contours and edge-based features of uppercase western script
0.012 230	Improving offline handwritten text recognition with hybrid HMM/ANN models
0.012 151	An HMM-based approach for off-line unconstrained handwritten word modeling and recognition
0.012 146	Statistical character structure modeling and its application to handwritten Chinese character recognition
0.012 117	A discrete contextual stochastic model for the off-line recognition of handwritten Chinese characters
0.011 934	Off-line recognition of totally unconstrained handwritten numerals using multilayer cluster neural network

PAMI: *IEEE Transactions on Pattern Analysis and Machine Intelligence*

Table 2 Top 10 highest impact papers on topic '3D reconstruction' (topic 29, CS)

Topic–doc probability	Venue	Paper title
0.007 036	ICCV, 1999	Inherent two-way ambiguity in 2D projective reconstruction from three uncalibrated 1D images
0.006 719	PAMI, 2001	Two-way ambiguity in 2D projective reconstruction from three uncalibrated 1D images
0.005 850	CVPR, 1999	Trajectory triangulation of lines: reconstruction of a 3D point moving along a line from a monocular image sequence
0.005 794	PAMI, 1995	Invariants of six points and projective reconstruction from three uncalibrated images
0.004 687	CVPR, 1997	Uncalibrated 1D projective camera and 3D affine reconstruction of lines
0.004 350	IJCAI, 1991	Combining stereo and monocular information to compute dense depth maps that preserve depth discontinuities
0.004 305	ECCV, 1998	A factorization method for projective and Euclidean reconstruction from multiple perspective views via iterative depth estimation
0.004 140	ECCV, 2000	Homography tensors: on algebraic entities that represent three views of static or moving planar points
0.004 084	CVPR, 1997	Critical motion sequences for monocular self-calibration and uncalibrated Euclidean reconstruction
0.003 963	CVPR, 1999	Efficient iterative solution to M-view projective reconstruction problem

CS: IEEE Computer Society; PAMI: *IEEE Transactions on Pattern Analysis and Machine Intelligence*; ICCV: IEEE International Conference on Computer Vision; CVPR: IEEE Conference on Computer Vision and Pattern Recognition; IJCAI: International Joint Conferences on Artificial Intelligence; ECCV: European Conference on Computer Vision

belong to the character recognition research area. All the milestone papers listed in Table 2 are related closely to topic '3D reconstruction' from CS. The venue/journal sources of all the papers belong mainly to the image processing and computer vision fields, e.g., PAMI, IEEE International Conference on Computer Vision (ICCV), IEEE Conference on Computer Vision and Pattern Recognition (CVPR), and European Conference on Computer Vision (ECCV).

5.2.2 Extracting topic keywords

The representative 10 topics identified in the PAMI and CS datasets are listed in Tables 3 and 4, respectively. For each topic (which will be discussed in detail later in Section 5.3, the six words with the highest probabilities are singled out. The extracted keywords are related to the task, problem, methodology, and model of the topics. For topic 15 in PAMI, the 'face recognition' problem, most models involve a 'learning' process and use 'dimensionality reduction' methods. For topic 37 in CS, 'classification problems', methods include 'support vector machines', 'kernel', and 'learning' techniques. In general, it is easy to summarize the research problems or specific methods for each topic through the extracted key words, as different methods/models which apply to the same topic are represented in the extracted keywords. For example, topics 5 and 6 in PAMI are related to the research theme 'image segmentation', but the key words differentiate them: 'Markov random field modeling' (topic 5) versus 'texture modeling' (topic 6).

5.3 Topic evolution analysis results

5.3.1 Topic segmentation by time range

After applying the citation-content-LDA model to extract topics from the PAMI and CS datasets, we tracked the topic evolution graph for these topics through two-step topic segmentation and topic evolution graph generation processes described above. The parameters for topic segmentation are the start year and end year (1995 and 2012, respectively for PAMI, and 1985 and 2004, respectively for CS) and the number of segments (six for PAMI, five for CS). The topic segmentation results for topics 15 (PAMI) and 29 (CS) are presented in Tables 5 and 6, respectively. In each table, we list the sequence number and the time range of each segment, and the top key words

Table 3 Representative 10 topics in the PAMI dataset (30 topics)

Topic	Weight	Top key words
15	0.052 477	Face; Recognition; Dimensionality; Reduction; Object; Learning
6	0.042 572	Image; Segmentation; Texture; Modeling; Local; Detection
18	0.041 959	Affine; Invariant; Shape; Recognition; Wavelet; Local
13	0.041 887	Texture; Classification; Image classification; Performance; Feature
5	0.040 208	Markov; Random; Field; Image segmentation; Textured
10	0.040 099	Handwritten; Character; Recognition; Off-line; HMM; Features
16	0.039 933	Graph; Matching; Recognition; Shape; Algorithm; Detection
20	0.032 613	Visual; Tracking; Object; Visual; 3D; Learning
3	0.031 624	Classification; Feature; Selection; Learning; Nearest; Neighbor
9	0.033 609	Stereo; Matching; Shape; Detection; Object; Robust

PAMI: *IEEE Transactions on Pattern Analysis and Machine Intelligence*

Table 4 Representative 10 topics in the CS dataset (100 topics)

Topic	Weight	Top key words
29	0.014 080	3D; Reconstruction; Motion; Images; Stereo; Projective
59	0.013 223	Reinforcement; Learning; Function; Algorithm; Decision; Concept
32	0.012 706	Decision; Tree; Learning; Classifiers; Induction; Estimation
3	0.012 546	Explanation-based; Learning; Control; Planning; Search; Reasoning
37	0.012 156	Support; Vector; Machines; Kernel; Learning; Classification
6	0.012 041	Constraint; Satisfaction; Problem; Satisfaction; Search; Solving
40	0.011 527	Neural; Networks; Learning; Analysis; Text; Classification
91	0.010 915	Information; Retrieval; Language; Query; Text; Document
48	0.010 622	Belief; Networks; Learning; Inference; Bayesian; Probabilistic
45	0.009 727	Reasoning; Logic; Default; Learning; Nonmonotonic; Networks

CS: IEEE Computer Society

and key papers (only paper IDs are listed here to save space; the paper titles are listed in Tables S1 and S2 in the supplementary) for each topic in each segment. Other topic segmentation results are omitted here to save space. Table 5 shows that any topic has different research focuses in different time segments, reflecting the topic evolution over time. For example, topic 15 (PAMI) is mainly about research on face recognition; however, in different segments there are different emphases, including, from the first to the sixth segments: image matching, support vector machines, neural networks, dimensionality reduction (at both fourth and fifth segments), and sparse representation.

5.3.2 Topic evolution graph

After topic segmentation, we constructed the topic evolution graph for both datasets via the algorithm discussed in Section 4.2. We set τ=0.2 and μ=0.8. Fig. 5 shows the resulting topic evolution graph for the PAMI dataset, and the graph for the CS dataset is shown in Fig. 6 (we divide it into two graphs, because it is fairly large).

The topic evolution graph for the PAMI dataset is divisible into four components corresponding to four research directions: image segmentation, face recognition, handwritten character recognition, and tracking. Numbers in the graph indicate the number of topics, and edges indicate the evolutionary paths between topics. For each research direction of the four components, there are one or several evolutionary paths: image segmentation has seven different paths and focuses on topics 4, 5, 6, 7, and 17; face

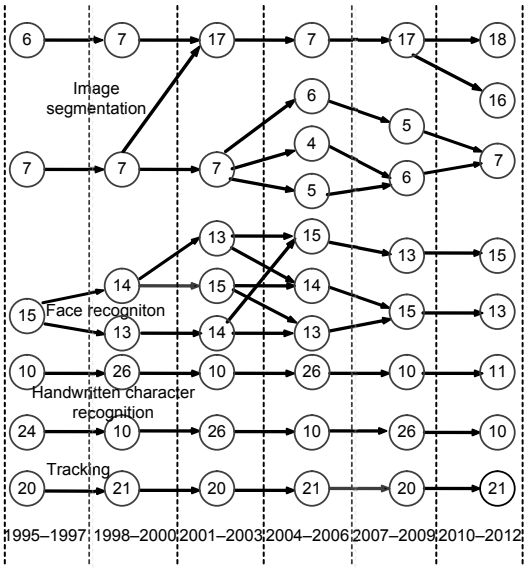

Fig. 5 Topic evolution graph for the PAMI dataset
PAMI: *IEEE Transactions on Pattern Analysis and Machine Intelligence*

Table 5 Topic segmentation results of topic 'face recognition' (topic 15, PAMI)

Segment	Time range	Top key words	Key papers (paper ID)
1	1995–1997	Recognition; Object; Image; Matching; Visual; Face	349, 351, 356, 364, 352, 348
2	1998–2000	Recognition; Face; Object; Support; Vector; Illumination	737, 428, 424, 612, 782, 637
3	2001–2003	Recognition; Face; Neural; Networks; Feature; Lighting	862, 814, 1180, 1189, 1108, 1222
4	2004–2006	Recognition; Face; Linear; Dimensionality; Reduction; PCA	1362, 1628, 1242, 1404, 1282, 1448
5	2007–2009	Recognition; Face; Reduction; Dimensionality; Object; Extraction	1880, 2007, 2099, 1763, 1959, 1979
6	2010–2012	Recognition; Face; Illumination; Alignment; Sparse; Representation	2596, 2391, 2297, 2385, 2450, 2445

PAMI: *IEEE Transactions on Pattern Analysis and Machine Intelligence*

Table 6 Topic segmentation results of topic '3D reconstruction' (topic 29, CS)

Segment	Time range	Top key words	Key papers (paper ID)
1	1985–1988	3D; Vision; Image; Motion; Calibration; Stereo	16 549 332, 13 920 339, 11 693 906, 12 744 126, 19 254 858, 16 611 145
2	1989–1992	Stereo; Motion; 3D; Estimation; Reconstruction; Uncalibrated	9 791 025, 12 618 997, 12 838 003, 10 764 090, 11 046 635, 13 830 823
3	1993–1996	Motion; Reconstruction; 3D; Structure; Geometry; Projective	9 968 838, 16 700 632, 12 657 226, 13 382 699, 12 134 266, 11 360 979
4	1997–2000	Reconstruction; 3D; Motion; Projective; Stereo; Structure	11 337 071, 15 582 579, 16 222 326, 13 908 159, 18 818 276, 18 338 412
5	2001–2004	3D; Reconstruction; Motion; Image; Stereo; Projective	14 227 596, 11 931 025, 11 078 465, 10 088 014, 17 623 078, 15 639 317

CS: IEEE Computer Society

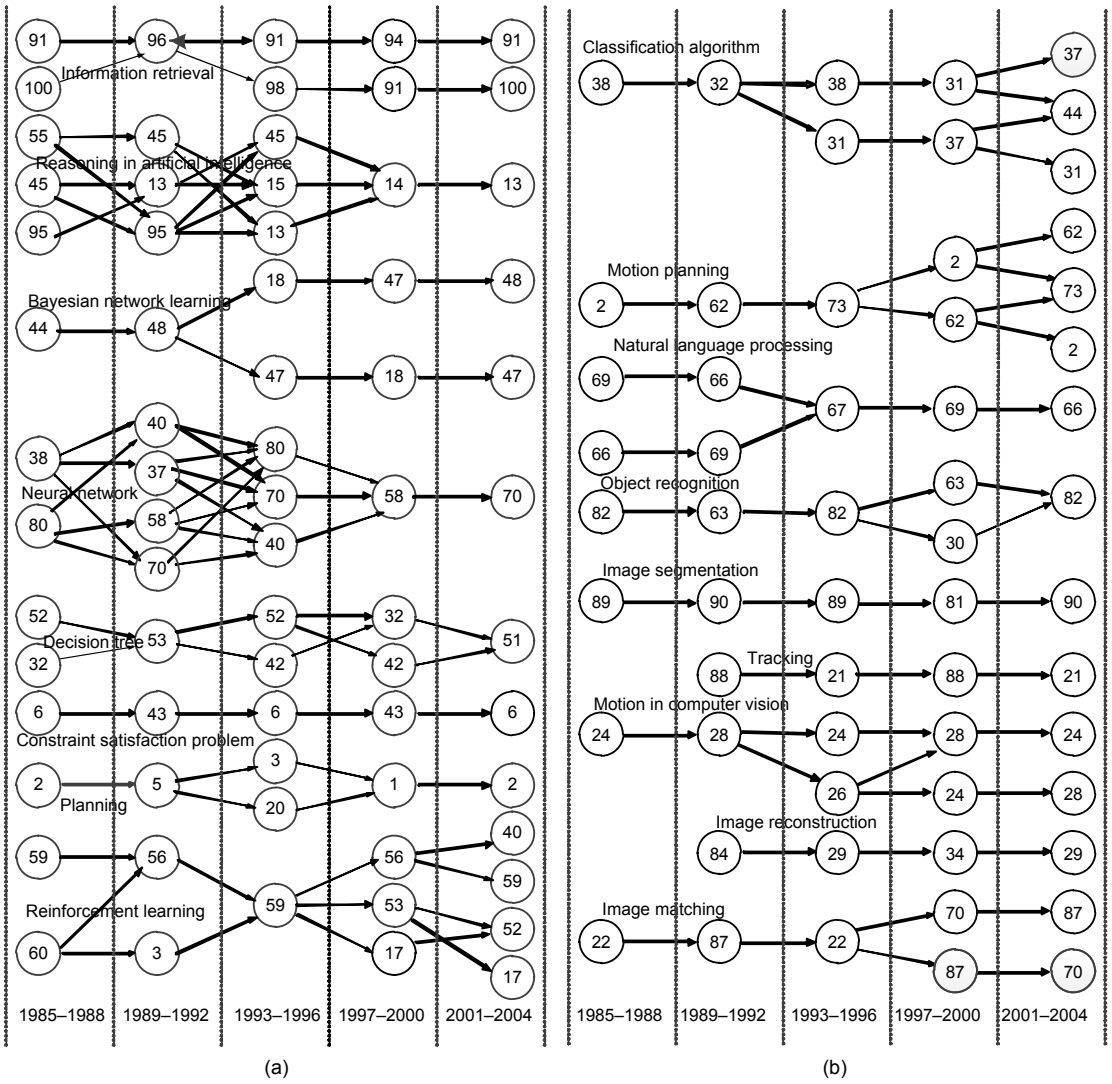

Fig. 6 Topic evolution graph for the CS dataset: (a) part 1; (b) part 2 (CS: IEEE Computer Society)

recognition has six paths and focuses on topics 13, 14, and 15; handwritten character recognition has three paths and focuses on topics 10 and 26; and tracking has only one path and focuses on topics 20 and 21. The number of paths for different components in the graph reflects the importance and complexity of the themes identified in the PAMI dataset. To save space, we include the detailed top key words of each topic in the supplementary (Tables S3–S5).

The first half of the CS dataset graph is shown in Fig. 6a, with eight different components representing eight different research directions in the dataset: information retrieval, reasoning, Bayesian network learning, neural network, decision tree, constraint satisfaction problem, planning, and reinforcement learning. These components represent three computer science domains: information retrieval, artificial intelligence, and machine learning. Most of them relate to artificial intelligence, while reasoning, neural network, and reinforcement learning show more complex evolutionary paths.

The other nine components of the CS dataset are shown in Fig. 6b, including the classification algorithm, motion planning, natural language processing (NLP), object recognition, image segmentation, tracking, motion, image reconstruction, and image matching. These components represent six research domains of computer science: machine learning, robotics and automation, NLP, pattern recognition, image processing, and computer vision. The last two contain more components than the others. Classification algorithm, motion planning, and motion themes

show more complex evolutionary paths than the other themes, and there are no related topics projected into time range 1985–1988 for tracking or image reconstruction themes.

5.4 Model selection and experimental comparisons

Here, we discuss how to select the number of topics for citation-content-LDA and compare its performance with that of content-LDA and citation-LDA based on two metrics, perplexity and sKL divergence. We included the conventional content-LDA and citation-LDA models as our baseline, using the title to represent the papers in both PAMI and CS.

5.4.1 Perplexity evaluation

Perplexity proposed by Blei *et al.* (2003) is an important criterion used to show the generalization power of a model on unseen data and the number of topics. Perplexity is equivalent to the inverse of the geometric mean per-word likelihood. It is a monotonically decreasing function in the likelihood of the test dataset, where a lower perplexity score indicates that the model has a better generalization power.

Generally, for a test set D_{test}, perplexity is defined as follows:

$$\text{perplexity}(D_{\text{test}}) = \exp\left(-\sum_{d=1}^{M}\log p(\boldsymbol{w}_d)\Big/\sum_{d=1}^{M}N_d\right), \tag{12}$$

where N_d represents the number of words in document d and $\boldsymbol{w}_d=(w_{1d}, w_{2d}, \ldots, w_{nd})$ is the vector form of document d.

Figs. 7 and 8 show the comparison of the experimental results in terms of perplexity between the PAMI and CS datasets, respectively. The results show that citation-content-LDA has a better perplexity performance than content-LDA or citation-LDA, and that 30 topics for PAMI and 100 topics for CS are appropriate numbers of topics.

5.4.2 Symmetric Kullback–Leibler divergence evaluation

sKL divergence can also be used to measure the similarity of a pair of topics. It is often employed to estimate the similarity between dual probability distributions, as it effectively represents a natural distance measure for probability distribution (Lin *et al.*, 2007). Given the topic–word distribution of a topic pair, sKL divergence is defined as follows:

$$\text{sKL}(\theta_i,\theta_j) = \sum_{k=2}^{N}\frac{1}{2}\left(\theta_{ik}\log\frac{\theta_{ik}}{\theta_{jk}} + \theta_{jk}\log\frac{\theta_{jk}}{\theta_{ik}}\right). \tag{13}$$

A higher sKL divergence score indicates that the model has more distinct topics. Tables 7 and 8 show the comparison of the experimental results of sKL divergence between the PAMI and CS datasets. We again find that citation-content-LDA outperforms the other two LDA models, and 30 and 100 are good choices for the PAMI and CS dataset topic numbers, respectively.

From Tables 7 and 8, we again observe the following results: (1) Citation-content-LDA performs better in topic detection based on sKL divergence, compared with both content-LDA and citation-LDA; (2) Thirty and 100 topics are good choices for the PAMI and CS datasets, respectively.

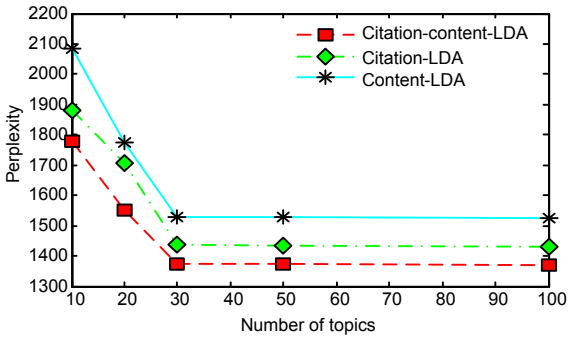

Fig. 7 Comparison of three different models for the PAMI dataset in terms of perplexity
LDA: latent Dirichlet allocation; PAMI: *IEEE Transactions on Pattern Analysis and Machine Intelligence*

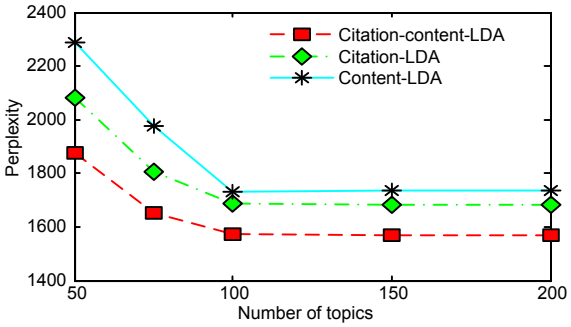

Fig. 8 Comparison of three different models for the CS dataset in terms of perplexity
LDA: latent Dirichlet allocation; CS: IEEE Computer Society

Table 7 Symmetric Kullback–Leibler divergence results for the PAMI dataset

Number of topics	Divergence		
	Content-LDA	Citation-LDA	Citation-content-LDA
10	12.3	13.8	14.1
30	13.2	14.3	15.3
50	12.1	13.5	13.7

LDA: latent Dirichlet allocation; PAMI: *IEEE Transactions on Pattern Analysis and Machine Intelligence*

Table 8 Symmetric Kullback–Leibler divergence results for the CS dataset

Number of topics	Divergence		
	Content-LDA	Citation-LDA	Citation-content-LDA
50	18.6	19.7	21.3
100	19.1	21.4	22.5
150	17.9	18.2	20.4

LDA: latent Dirichlet allocation; CS: IEEE Computer Society

6 Conclusions

In this paper, we presented a two-level latent citation-content-LDA topic extraction model plus a novel topic evolution algorithm for identifying and analyzing topics and thematic evolution within the bodies of scientific literature. The citation-content-LDA model, which exploits both the document itself and its citation information, can extract research topics accurately according to key words and milestone papers. The topic evolution algorithm includes both topic segmentation by timestamp and topic dependency relation calculation. Both processes are necessary to construct a complete topic evolution graph that accurately represents research themes in the corpus. We ran experiments on the PAMI and CS datasets. The results showed that our proposed model, citation-content-LDA, outperforms both content-LDA and citation-LDA models. In effect, citation-content-LDA allows researchers to process scientific literature quickly and effectively.

There are several future directions in which we could take this research, for example, including in the model information such as the authors' names and the journal/venue for the documents. In terms of topic evolution, quantitative analysis of topic dependency relations and probability reasoning in the topic evolution graph are also interesting possible future research directions.

References

Ahmed, A., Xing, E.P., 2010. Timeline: a dynamic hierarchical Dirichlet process model for recovering birth/death and evolution of topics in text stream. Proc. 26th Conf. on Uncertainty in Artificial Intelligence, p.20-29.

Blei, D.M., Lafferty, J.D., 2006. Dynamic topic models. Proc. 23rd ACM Int. Conf. on Machine Learning, p.113-120. https://doi.org/10.1145/1143844.1143859

Blei, D.M., Ng, A.Y., Jordan, M.I., 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.*, **3**:993-1022.

Brin, B.S., Page, L., 1998. The anatomy of a large scale hypertextual web search engine. *Comput. Netw. ISDN Syst.*, **30**(98):107-117. https://doi.org/10.1016/S0169-7552(98)00110-X

Chang, J., Blei, D.M., 2009. Relational topic models for document networks. Proc. 12th Int. Conf. on Artificial Intelligence and Statistics, p.81-88.

Cohn, D., Chang, H., 2000. Learning to probabilistically identify authoritative documents. Proc. 17th Int. Conf. on Machine Learning, p.167-174.

Dietz, L., Bickel, S., Scheffer, T., 2007. Unsupervised prediction of citation influences. Proc. 24th ACM Int. Conf. on Machine Learning, p.233-240. https://doi.org/10.1145/1273496.1273526

Erosheva, E., Fienberg, S., Lafferty, J., 2004. Mixed-membership models of scientific publications. *PNAS*, **101**(Suppl 1):5220-5227. https://doi.org/10.1073/pnas.0307760101

Griffiths, T.L., Steyvers, M., 2004. Finding scientific topics. *PNAS*, **101**(Suppl 1):5228-5235. https://doi.org/10.1073/pnas.0307752101

Guo, Z., Zhang, Z., Zhu, S., *et al.*, 2014. A two-level topic model towards knowledge discovery from citation networks. *IEEE Trans. Knowl. Data Eng.*, **26**(4):780-794. https://doi.org/10.1109/TKDE.2013.56

He, Q., Chen, B., Pei, J., *et al.*, 2009. Detecting topic evolution in scientific literature: how can citations help? Proc. 18th ACM Conf. on Information and Knowledge Management, p.957-966. https://doi.org/10.1145/1645953.1646076

Hofmann, T., 2001. Unsupervised learning by probabilistic latent semantic analysis. *Mach. Learn.*, **42**(1-2):177-196. https://doi.org/10.1023/A:1007617005950

Lin, F.R., Huang, F.M., Liang, C.H., 2007. Individualized storyline-based news topic retrospection. Pacific Asia Conf. on Information Systems, Article 140.

Lu, Z., Mamoulis, N., Cheung, D.W., 2014. A collective topic model for milestone paper discovery. Proc. 37th Int. ACM SIGIR Conf. on Research & Development in Information Retrieval, p.1019-1022. https://doi.org/10.1145/2600428.2609499

Macroberts, M.H., Macroberts, B.R., 1989. Problems of citation analysis: a critical review. *J. Am. Soc. Inform. Sci.*, **40**(5):342-349.

https://doi.org/10.1002/(SICI)1097-4571(198909)40:5<342::AID-ASI7>3.0.CO;2-U

Mei, Q., Zhai, C., 2005. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. Proc. 11th ACM SIGKDD Int. Conf. on Knowledge Discovery in Data Mining, p.198-207. https://doi.org/10.1145/1081870.1081895

Mei, Q., Cai, D., Zhang, D., *et al*., 2008. Topic modeling with network regularization. Proc. 17th Int. Conf. on World Wide Web, p.101-110. https://doi.org/10.1145/1367497.1367512

Nallapati, R., Cohen, W.W., 2008. Link-PLSA-LDA: a new unsupervised model for topics and influence of blogs. Proc. 2nd Int. Conf. on Weblogs and Social Media, p.84-92.

Nallapati, R.M., Ahmed, A., Xing, E.P., *et al*., 2008. Joint latent topic models for text and citations. Proc. 14th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.542-550. https://doi.org/10.1145/1401890.1401957

Wang, X.L., Zhai, C.X., Roth, D., 2013. Understanding evolution of research themes: a probabilistic generative model for citations. Proc. 19th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.1115-1123. https://doi.org/10.1145/2487575.2487698

Wang, X.R., McCallum, A., 2006. Topics over time: a non-Markov continuous-time model of topical trends. Proc. 12th ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining, p.424-433. https://doi.org/10.1145/1150402.1150450

List of supplementary materials

Table S1 Key papers of topic 15 in the PAMI dataset (for Table 5 in Section 5.3)
Table S2 Key papers of topic 29 in the CS dataset (for Table 6 in Section 5.3)
Table S3 Top 10 key words of sub-topics for different topics in the PAMI dataset (for Fig. 5 in Section 5.3)
Table S4 Top 10 key words of sub-topics for different topics in the CS dataset (1) (for Fig. 6a in Section 5.3)
Table S5 Top 10 key words of sub-topics for different topics in the CS dataset (2) (for Fig. 6b in Section 5.3)