Detecting Topic Evolution in Scientific Literature: How Can Citations Help?

Qi He[†] Bi Chen[†] Jian Pei[‡] Baojun Qiu[§] Prasenjit Mitra[†,§] C. Lee Giles[†,§]
[†] College of Information Sciences and Technology, The Pennsylvania State University,
University Park, PA 16802
[‡] School of Computing Science, Simon Fraser University, Burnaby, BC Canada V5A 1S6
[§] Department of Computer Science and Engineering, The Pennsylvania State University,
University Park, PA 16802
[†]{qhe, bchen, pmitra, giles}@ist.psu.edu, [‡]jpei@cs.sfu.ca, [§]bqiu@cse.psu.edu

ABSTRACT

Understanding how topics in scientific literature evolve is an interesting and important problem. Previous work simply models each paper as a bag of words and also considers the impact of authors. However, the impact of one document on another as captured by citations, one important inherent element in scientific literature, has not been considered. In this paper, we address the problem of understanding topic evolution by leveraging citations, and develop citation-aware approaches. We propose an iterative topic evolution learning framework by adapting the Latent Dirichlet Allocation model to the citation network and develop a novel inheritance topic model. We evaluate the effectiveness and efficiency of our approaches and compare with the state of the art approaches on a large collection of more than 650,000 research papers in the last 16 years and the citation network enabled by CiteSeerX. The results clearly show that citations can help to understand topic evolution better.

Categories and Subject Descriptors

H.1.m [**Information Systems Models and Principles**]: Miscellaneous

General Terms

Algorithms, Experimentation, Human Factors

Keywords

topic evolution, citations, Inheritance Topic Model

1. INTRODUCTION

"Dwarfs standing on the shoulders of giants." In scientific research, many new topics and new principles evolve from existing ones. Our knowledge, as well as the development of our knowledge, have been largely recorded in detail by a huge amount of archived scientific literature in the last several hundred years. Detailed research papers can be summarized by topics. Can we understand how topics evolve over time by mining the archived scientific literature?

Topic evolution in scientific literature shows how research on one topic influenced research on another and helps us understand the lineage of topics. Understanding such topic evolution is an important problem with a few interesting applications. For example, in sociology of science, topic evolution analysis can help us understand and objectively evaluate the contribution of a scientist or an article. Moreover, topic evolution analysis may lead to information retrieval tools that can recommend citations for scientific researchers.

Due to its importance and great application potential, topic evolution has recently attracted fast growing interest in the information retrieval community [23, 19, 20, 28, 27, 33, 22]. Existing approaches [33, 6, 7, 13] for topic evolution in scientific literature model a paper as a bag of words, and detect topics on documents in different time periods. Then, topic evolution is analyzed by comparing the changes of topics over time as well as the number of documents of different topics. Some recent work further tries to analyze the roles of social network analysis (i.e., the co-authorship [29, 33, 21] or direction-sensitive messages sent between authors [18]), annotated data [4], named entities [25] and ontologies [22] in topic detection and evolution. Section 2 reviews those existing methods briefly.

A research paper contains more information than just a bag of words. Particularly, for topic evolution, citations, the important inherent elements in scientific literature, naturally indicate linkages between topics. Surprisingly, citations have not been considered by most of the existing methods for topic evolution. Bolelli *et al.* [6, 7] propose a segmented author-topic model to identify topic evolution by simply using citations to identify and boost the weight for the top "topic-bearing" words in documents. To the best of our knowledge, no existing work directly infers citations in the Bayesian framework and fits topic evolution to the temporal development of citations. One of the key advantages of a Bayesian framework for modeling citations is that the uncertainty associated with the citation parameters (e.g., influential weights on citing papers) can be quantified. We propose such a Bayesian model to identify the evolution of topics.

Can citations be easily used in topic evolution? One challenge is that the impact of citations cannot be captured by casting them in a straightforward manner into a bag of words. Intuitively, when a paper A cites another paper B, more often than not, A wants to use some content of B to extend the content of A. Therefore, the topics in B should have some impact on the topics in A. Without considering such impact, we may miss some topics related to A.

How can we capture such impact in an effective way? We need to develop a comprehensive model to integrate both the main body

CIKM'09, November 2–6, 2009, Hong Kong, China.
Copyright 2009 ACM 978-1-60558-512-3/09/11 ...$10.00.

of a document, modeled as a bag of words, and citations. Another challenge is that there exists a huge amount of literature. Understanding topic evolution on a large number of research papers demands high efficiency and scalability in the underlying models and analysis methods.

In this paper, we tackle the problem of topic evolution analysis on scientific literature by leveraging citations. When detecting topics in a collection $D(t)$ of new papers, in addition to those papers in $D(t)$, we also consider the papers not in $D(t)$ but cited by the papers in $D(t)$. To the best of our knowledge, we are the first to tackle the problem in this manner. We make the following contributions.

First, we present a simple yet effective model for topic evolution analysis. We quantify the similarity of topics and measure the relationship between two topics in different types of topic evolution.

Second, we develop effective and efficient methods for topic evolution analysis systematically. We explore two steps. Since the Latent Dirichlet Allocation (LDA) model [3] has been extensively adopted in information retrieval [4, 12, 25, 29, 33, 18, 22, 24], as the first step we extend LDA for topic evolution analysis. For each unit time period, we generate the topics *independently*, and then compare the topics with the previous topic space to track topic changes. The temporal order of documents within the same time period has not been considered. As the next step, we successively release the constraints of this simple solution one by one. Topics depend on documents not only in the current time period, but also from previous time periods. We further propose a novel inheritance topic model that conceptually captures how citations can be used to analyze topic evolution in an explicit way. In this model, citations are explicitly modeled as topic inheritance. The temporal order of documents even at the same time period has been considered and must respect the partial order of citation graph.

Last, we conduct an extensive empirical study using a real dataset of more than 650,000 research papers in the last 16 years and the citation network enabled by CiteSeerX, a scientific literature digital library and search engine focusing primarily on the literature in computer and information science. The results clearly show that citations can help to understand topic evolution better, and our methods are effective and efficient.

2. RELATED WORK

In this section, we briefly review previous work in topic models and their applications on topic detection and topic evolution. Other previous work not using topic models but solving similar problems is discussed as well.

2.1 Probabilistic Topic Models

The LDA model on the bag of words [3] was extended to model 1) the impact of authors [29, 33]; 2) the impact of the direction-sensitive messages sent between social entities (e.g., persons) [18]; 3) the impact of one type of annotation on another type of annotation at the topic-level for annotated data [4]; 4) the impact of named entities [25]; and 5) the impact of ontologies [22].

None of the above has modeled the impact of citations while generating topics. Recent work including [10, 12, 24] respectively used pLSA model, LDA model and a combination of them to predict the citations between documents by modeling topicality of citations. If the topic distribution of a citation can be generated by two documents with a high probability, this citation was recommended to link these two documents. Instead of predicting citations, we explicitly use citations to enhance topic evolutions in our novel inheritance topic model, which is partially motivated by [11], where the citations were also modeled as inheritances. However, the latter can only handle the simple bipartite citation graph, not the complex citation network as ours does. Moreover, in [11], any paper that cites and is cited by other papers needs to be cloned, with one cloned version being treated as a citing paper and another as a cited paper. The cloning operation adds another difficulty, as the topics associated with the two clones are statistically unrelated.

2.2 Topic Detection

Topic detection was defined to generate the topics from a document stream, which has been extensively studied by the topic detection and tracking (TDT) community in the past [1]. The main task of TDT does not include topic evolution as we target at in this paper. The pair-wise topic relations are crucial in our problem, which cannot be solved by existing TDT techniques.

Moreover, no existing work on topic detection has considered the citations between documents, except for Jo *et al.* [16], which attempted to combine citations and text for topic detection. Heuristically, if a term (2-gram words) is relevant to a topic, the sub-citation graph consisting of those documents containing this term has a denser connectivity than any sub-citation graph consisting of documents randomly selected. In other words, citations are only used to model the topical similarity between documents.

In fact, the previous work on topic detection did not consider the relations among topics thoroughly, though some of them [8, 9] tried to implicitly link correlated topics by understanding documents using natural language processing techniques.

In this paper, we propose a generative topic model for topic detection. Our topic detection model is designed to detect topics for the main task of topic evolution. Moreover, our generative model explicitly uses citations between documents to model topics, so that the connection between new topics and old topics can be more easily captured compared to *citation-unaware* topic models.

2.3 Topic Evolution

The main task of topic evolution is to discover how and what topics change over time.

2.3.1 Discriminative Approaches

Changes of topics are monitored by treating each topic as a distribution over words or a mixture over documents. Morinaga and Yamanishi [23] used a finite mixture model to represent documents at each discrete time. Their algorithm detects topic changes on certain documents if the topic mixtures drift significantly from the previous ones. Mei and Zhai [19] conducted clustering sequentially and then correlated clusters via a temporal graph model, which was in turn used to represent the topic evolutions in a document stream. Mei *et al.* [20] used a probabilistic approach to detect spatiotemporal theme patterns and then observed the evolution of theme patterns by comparing the theme life cycles and theme snapshots. Spiliopoulou *et al.* [28] detected and tracked changes in clusters based on the content of the underlying data stream. Schult and Spiliopoulou [27] used a clustering approach to find out the ontology/taxonomy evolution for documents.

2.3.2 Generative Approaches

Recently, many studies used generative topic models to observe topic evolution on document streams. Zhou *et al.* [33] used the LDA model to observe temporal topic evolution over scientific literature. Specifically, a k-component LDA model is constructed over the whole dataset to generate k global topics. For each topic, the trend is obtained by simply counting the number of papers belonging to the topic year by year. The author information is also used to explain why some topics tend to decline yet some others expand.

Blei and Lafferty [5] developed a dynamic topic model (DTM) by assuming that topic models evolve gradually in time and are distributed normally. Specifically, a k-component LDA analysis is conducted at each time slice t. Each topic is modeled as a Gaussian process centered upon the previous value. Similar to [33], the topic is global and the topic trend is obtained by counting the number of papers. The dynamic topic model assumes that all papers at time t are correlated to all papers at time $t-1$. In our work, only cited papers at time $t-1$ are related to their citing papers at time t. Wang *et al.* [31] further extended this discrete DTM to a continuous version.

Morchen *et al.* [22] used probabilistic topic models to annotate recent articles with the most likely ontology terms. They also proposed a solution for automatically determining how new ontology terms can evolve from old terms. AlSumait *et al.* [2] extended the LDA model to an online version by incrementally updating the current model for new data and claimed that this model has certain ability of capturing the dynamic changes of topics. Gohr and Hinneburg [13] used latent variables to index new words while deleted those outdated words within a sliding window for a stream of documents. Those indexed new words were used to portray the topic changes for the information retrieval domain.

All the above work on topic evolution models a paper as a bag of words without considering the citations at all. More recently, Bolelli *et al.* [6, 7] proposed a generative author topic model that integrated the temporal ordering of the documents to model topic trends sequentially, where the discovered topics at an early time were propagated to influence the topics generated later. They use citations to identify "topic-bearing" words whose weights should be doubled. Mann *et al.* [17] used an n-gram topic model to identify the influence of one topic on another. However, this approach modeled citations indirectly in the topic model, and the resulting topic influence is also time irrelevant.

Our work is distinguished from the previous work on topic evolution in three ways. First, we consider both content and citations in a full-generative inheritance topic model. Second, we infer citations directly in the Bayesian framework. Third, we use a citation network analysis approach to explicitly emphasize the relationship between topics.

3. TOPIC EVOLUTION

In this section, we first describe the problem of topic evolution. Then, we present two citation-unaware Latent Dirichlet Allocation approaches.

3.1 Problem Definition

Let $\mathbb{W} = \{w_1, \ldots, w_V\}$ be a *vocabulary set*. A *(probabilistic) vocabulary distribution* on $\mathbb{W}$ is a point in the $V-1$ dimensional simplex, functioned as $\mathbf{f} : \mathbb{W} \to [0,1]$ such that $\sum_{w \in \mathbb{W}} \mathbf{f}(w) = 1$. A vocabulary distribution $\mathbf{f}$ can also be written as a vector $\mathbf{f} = \langle w_1 : \mathbf{f}(w_1), \ldots, w_V : \mathbf{f}(w_V) \rangle$. For two vocabulary distributions $\mathbf{f}$ and $\mathbf{g}$, the similarity between them is modeled as the cosine similarity[1]: $sim(\mathbf{f}, \mathbf{g}) = \frac{\mathbf{f} \cdot \mathbf{g}}{\|\mathbf{f}\| \|\mathbf{g}\|}$.

Let $D = \{d_1, \ldots, d_m\}$ be a set of scientific publication corpus in question. A *document* d consists of a vocabulary distribution, a citation set L_d, and a timestamp.

A *topic* $\mathbf{z}$ is a vocabulary distribution. Intuitively, a topic is popular if it is similar to many documents in D. Imagine that we virtually combine all documents in D into a single long document d'. We can get a word vector $\mathbf{w}$ for d'. Each element of $\mathbf{w}$ is a word from d'. If a word w appears n times in d', then there are n duplicates of w in $\mathbf{w}$. We call $\mathbf{w}$ the *word sampling space*. By conducting a Bernoulli trial (appear or not appear) for each element in the word sampling space, we can generate a vocabulary distribution, which is a candidate topic. Fixing the number of topics (e.g., k), the task of a *topic detection method* T is to generate k topics maximizing the likelihood of the observed data.

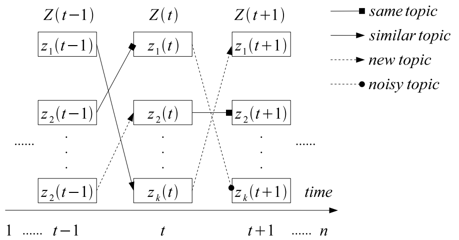


Figure 1: The topic evolution bipartite over time. Each rectangle represents a topic and the arc between 2 rectangles indicates various types of topic evolution.

To conduct topic evolution analysis, we divide the document corpus D into exclusive temporal subsets $D(1), \ldots, D(n)$ according to the timestamps of the documents such that $D = \cup_{t=1}^{n} D(t)$. Let $Z(t)$ be the k topics generated by T from $D(t)$. The problem of *topic evolution analysis* at time t is to analyze the relationship between the topics in $Z(t)$ and those in $Z(t-1)$.

Concretely, we need to specify the pairwise relationship between topics in $Z(t-1)$ and $Z(t)$. For two topics $\mathbf{z}_i(t-1) \in Z(t-1)$ and $\mathbf{z}_j(t) \in Z(t)$, we have

$$p(\mathbf{z}_j(t) | \mathbf{z}_i(t-1)) \propto sim(\mathbf{z}_i(t-1), \mathbf{z}_j(t)).$$

We simply use the raw similarity rather than computing the true conditional probability. This is a design decision because given an existing topic $\mathbf{z}_i(t-1)$, we never know the whole topic space which could evolve from it. If we simply assume that k topics in $Z(t)$ consist of the candidate set (each has a uniform prior $1/k$), then probabilities conditioned on different previous topics are incomparable to each other. Fortunately, the raw similarity does not take any topic as the reference object and thus affords a fair measure for all pairs of topics in comparison. The raw similarity is also constrained within the unit range $[0, 1]$, making the fair comparison practical by setting some global parameters.

Using two user-specified parameters ϵ_1 and ϵ_2 such that $1 \geq \epsilon_1 > \epsilon_2 > 1/k$, we define three types of relationships between $\mathbf{z}_i(t-1) \in Z(t-1)$ and $\mathbf{z}_j(t) \in Z(t)$:

Same topic: $\mathbf{z}_j(t)$ and $\mathbf{z}_i(t-1)$ are very similar. Specifically, $p(\mathbf{z}_j(t) | \mathbf{z}_i(t-1)) \geq \epsilon_1$;

Similar topic: $\mathbf{z}_j(t)$ are similar to $\mathbf{z}_i(t-1)$, that is, $\epsilon_1 > p(\mathbf{z}_j(t) | \mathbf{z}_i(t-1)) \geq \epsilon_2$; and

New topic: $\mathbf{z}_j(t)$ looks new compared to $\mathbf{z}_i(t-1)$, that is, $p(\mathbf{z}_j(t) | \mathbf{z}_i(t-1)) < \epsilon_2$.

The two threshold parameters ϵ_1 and ϵ_2 may be determined experimentally. A user may also judge whether a topic is meaningful. We thus set up the fourth type, noisy topic, which means such a topic does not correspond to any meaningful topic, i.e., it contains mainly stopwords that are always present.

For simplicity, in this paper, the number of topics for each discrete time is fixed as k. It can be easily extended to any dynamic number of topics using algorithms such as Hierarchical Dirichlet Process [30].

Based on the above four types of topic evolution, we can generate a *topic evolution bipartite* over time for the whole document corpus D, as elaborated in Figure 1. An arc from one topic $\mathbf{z}_i(t-1)$ to another one $\mathbf{z}_j(t)$ indicates that within $Z(t-1)$, $\mathbf{z}_i(t-1)$ has the maximum conditional probability to $\mathbf{z}_j(t)$.

3.2 Citation-Unaware Approaches

Let us consider two simple approaches for topic evolution.

[1] A point in the $V-1$ dimensional simplex can be easily mapped to the V dimensional Euclidean space using the natural parameterization [5], so that the cosine similarity can be calculated properly.

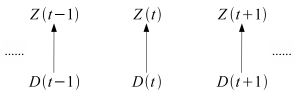


Figure 2: The independent topic evolution learning.

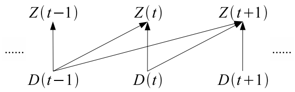


Figure 3: The accumulative topic evolution learning.

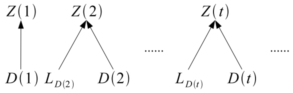


Figure 4: The *citation-aware* topic evolution learning.

Given the current time t, the *independent topic evolution learning* method detects topics only from $D(t)$. In other words, $Z(t)$ is independent from $Z(t-1)$, as illustrated in Figure 2. The learning process is defined as follows.

$$Z(t) = \arg\max_{Z(t)} \prod_{d \in D(t)} p(d|Z(t)), \quad (1)$$

where $p(d|Z(t))$ is the likelihood of document d given $Z(t)$ by assuming all documents in $D(t)$ are equally important for $Z(t)$.

Can we consider the dependence of the topics in $Z(t)$ on the documents at time instant t and before? The *accumulative topic evolution learning* method, as elaborated in Figure 3, learns the current topic space $Z(t)$ from all papers published at time t and before, i.e., from document set $\cup_{i=1}^{t} D(i)$. The learning process is

$$Z(t) = \arg\max_{Z(t)} \prod_{d \in \cup_{i=1}^{t} D(i)} p(d|Z(t)), \quad (2)$$

assuming all documents in $\cup_{i=1}^{t} D(i)$ are equally important for $Z(t)$.

Both methods are *citation-unaware* since they do not consider the citations. The independent topic evolution learning method tends to generate a large number of isolated new topics irrelevant to existing topics. In the accumulative topic evolution learning method, the existing topics tend to dominate the topic space as time goes by.

To learn topic spaces in the two citation-unaware methods, i.e., maximizing the likelihood of the data, any traditional topic models can be applied. Here, we use one of the most popular models in machine learning and information retrieval, the Latent Dirichlet Allocation (LDA) [3] framework, to generate topics. Collapsed Gibbs sampler can be used to infer the LDA posterior probabilities [14]. We denote by *i-LDA* the Gibbs sampling algorithm of independent topic evolution learning, and by *a-LDA* the Gibbs sampling algorithm of accumulative topic evolution learning.

4. CITATION-AWARE APPROACHES

In this section, we extend the LDA approaches in Section 3 by taking citations into account. We also develop an approach explicitly modeling citations as inheritance in documents.

4.1 Frameworks

To be citation aware, the current topic space $Z(t)$ should be generated not only from $D(t)$, but also from $L_{D(t)}$, the set of papers in $\{D(t') : t' < t\}$ cited by papers in $D(t)$, as illustrated in Figure 4.

At time t, a simple citation aware method to compute $Z(t)$ is

$$Z(t) = \arg\max_{Z(t)} \prod_{d \in D(t) \cup L_{D(t)}} p(d|Z(t)), \quad (3)$$

assuming all documents in $D(t) \cup L_{D(t)}$ are equally important for $Z(t)$. Again, we can use LDA in topic generation. We denote by *c-LDA* the algorithm of Eq. 3 using LDA.

Is *c-LDA* a good solution to balance new topics and existing topics via citations? There are two problems. First, not all citations are equally important. Among all papers cited by a document d, typically only a small subset is topic-related to d. Therefore, treating all citations equally may dilute the truly important topics. Second, due to the sheer number of historical papers, some out-of-date topics may be resurrected by citations solely if the citations are not properly associated with the current topics. We call such topics *ghost topics*.

To address the above concerns, we propose a learning method based on Dirichlet prior smoothing [32]. At the given time t, we learn the topic space by,

$$Z(t) = \arg\max_{Z(t)} \prod_{d \in D(t) \cup L_{D(t)}} p'(d|Z(t)), \quad (4)$$

where

$$p'(d|Z(t)) = \lambda \cdot p(d|Z(t)) + (1-\lambda) \cdot \sum_{d_j \in L_d} \gamma_{d_j} \cdot p'(d_j|Z(t))$$

is the likelihood of document d given $Z(t)$ linearly combining two factors: the language models for both citing and cited documents, weighted by λ, and the topical influence from all cited documents, individually weighted by the vector γ.

Eq. 4 actually defines an iterative learning process, where the topic models of both the citing paper and all its cited papers are learnt using the same procedure. To reduce the number of tuning parameters, we assume Dirichlet priors α_λ and α_γ for λ and γ, respectively. Here, λ is drawn from a Beta distribution $Beta(\alpha_\lambda)$, while γ is drawn from a Dirichlet distribution $Dirichlet(\alpha_\gamma)$. Plugging in the distributions, we have

$$\begin{aligned} p'(d|Z(t)) = p(d|\alpha_\lambda, \alpha_\gamma) = \prod_{w \in d} p(w|\alpha_\lambda, \alpha_\gamma) = \int p(\lambda|\alpha_\lambda) \\ \{\prod_{w \in d} \sum_{s \in \{0,1\}} p(s|\lambda)[s \cdot p(z|d)p(w|z) + (1-s) \cdot \int p(\gamma|\alpha_\gamma) \\ \Big(\sum_{c \in L_d} p(c = d_j|\gamma)p'(d_j|Z(t))p(z'|d_j)p(w|z')\Big)d\gamma]\}d\lambda, \quad (5) \end{aligned}$$

where the indicator variable s denotes whether a word is sampled from the cited papers ($s = 0$) with topic assignment z or from the citing paper ($s = 1$) with topic assignment z', and is drawn from a Bernoulli distribution $Bernoulli(\lambda)$, the variable c indicates for a word sampled from the cited papers ($s = 0$) which cited paper should be sampled from the citation list L_d, and is drawn from a multinomial distribution $Multi(\gamma)$, and $p(c = d_j|\gamma)$ represents the topical influence of the cited paper d_j on the citing paper d.

To the best of our knowledge, no existing topic model is able to support the iterative learning process defined in Eq. 5. We therefore have to develop our own topic model.

4.2 The Inheritance Topic Model

We propose the *Inheritance Topic Model* (ITM for short) in Figure 5 and notation is summarized in Table 1. In our topic model, a paper d is virtually separated into two parts: the *inherited* part d^0 and the *autonomous* part d^1, which are generated independently. The model captures the real world situations where a paper often

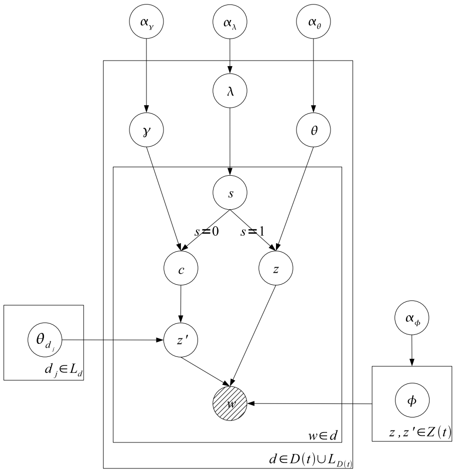


Figure 5: The *Inheritance Topic Model* at time t.

Table 1: Table of symbols for ITM and Gibbs equations.

k	# of topics
V	# of distinct words
α_ϕ	Dirichlet prior for topic-word distribution ϕ
α_θ	Dirichlet prior for document-topic distribution θ
α_λ	Dirichlet prior for inherited and autonomous portions distribution λ
α_γ	Dirichlet prior for document-citation strength distribution γ
θ_d	Topic mix for document d
γ_d	Distribution over the references cited by document d
λ_w	Distribution over two parts (inherited and autonomous) given w
ϕ_z	Distribution over words given z
$n(w, z)$	# times that the word w was assigned to topic z
$n(z)$	# total words assigned to topic z
$auto(d, z)$	# words in the autonomous part of document d that have topic z
$auto(d)$	# words in the autonomous part of document d
$auto(c, z)$	# words in the autonomous part of citation c assigned to topic z
$auto(c)$	# words in the autonomous part of citation c
$inhe(d)$	# words in the inherited part of document d
$inhe(d, c)$	# words in document d inherited from citation c
$inhe(c, z)$	# words inherited from citation c and assigned to topic z over all papers that cited c
$inhe(c)$	# words inherited from citation c over all papers that cited c

reuses ideas and techniques of previous work reflected by the cited papers, simultaneously, contains some new material. Technically, for a paper d, the inherited part d^0 is a mixture of the autonomous parts of all papers in L_d. The detailed generative process of ITM is given as below.

- Draw k multinomials $\phi_z \sim Dirichlet(\alpha_\phi)$, one for each topic.
- For each paper d:
  - Draw a topic distribution $\theta_d \sim Dirichlet(\alpha_\theta)$.
  - Draw $\gamma_d \sim Dirichlet(\alpha_\gamma)$ which measures the strength of cited papers L_d. The dimension of $\gamma_\mathbf{d}$ is $|L_d|$.
  - Draw $\lambda = [\lambda_0, \lambda_1] \sim Beta(\alpha_{\lambda_0}, \alpha_{\lambda_1})$ to weigh d^0 and d^1.
  - For each word $w \in d$:
    - Draw $s \sim Bernoulli(\lambda)$, which determines whether w is drawn from d^1 or d^0.
    - If $s = 1$ (authors' autonomous part):
      - Draw a topic $\mathbf{z} \sim multinomial(\theta_d)$.
      - Draw the word $w \sim multinomial(\phi_z)$.
    - If $s = 0$ (inherited part):
      - Draw the dummy index $c \sim multinomial(\gamma_d)$ of a cited paper d_j. d_j has topic distribution θ_j.
      - Draw a topic $\mathbf{z}' \sim multinomial(\theta_j)$.
      - Draw the word $w \sim multinomial(\phi_{z'})$.

4.3 Collapsed Gibbs Sampler Algorithm for Inferencing Inheritance Topic Model

Similar to LDA, in ITM, we also need to infer the ITM posterior probability. The joint probability of generating the word sampling space $\mathbf{w}$ at a time instant is

$$\begin{aligned} & p(\mathbf{w}, \mathbf{z}, \mathbf{c}, \mathbf{s} | \alpha_\phi, \alpha_\theta, \alpha_\gamma, \alpha_\lambda) \\ & = \int p(\mathbf{w}|\mathbf{z}, \phi) p(\phi|\alpha_\phi) d\phi \cdot \int p(\mathbf{c}|\gamma) p(\gamma|\alpha_\gamma, L_d) d\gamma \\ & \cdot \int p(\mathbf{z}|\mathbf{c}, \mathbf{s}, \theta) p(\theta|\alpha_\theta) d\theta \cdot \int p(\mathbf{s}|\lambda) p(\lambda|\alpha_\lambda) d\lambda. \qquad (6) \end{aligned}$$

We use collapsed Gibbs sampler algorithm to approximate the above joint distribution, which is denoted by *c-ITM*. At each iteration of Gibbs sampling, we update the latent variables for every word position using the following processes until the latent variables converge. We use the notation in Table 1.

$$\begin{aligned} & p(c|d, z, s = 0, data) \\ & \propto \Big(inhe(d, c) + \alpha_\gamma - 1\Big) \times \Big(\tfrac{auto(c,z)+inhe(c,z)+\alpha_\theta-1}{auto(c)+inhe(c)+k\cdot\alpha_\theta-1}\Big), \end{aligned}$$

$$\begin{aligned} & p(s = 0|d, c, z, data) \\ & \propto \Big(inhe(d) + \alpha_{\lambda_0} - 1\Big) \times \Big(\tfrac{auto(c,z)+inhe(c,z)+\alpha_\theta-1}{auto(c)+inhe(c)+k\cdot\alpha_\theta-1}\Big), \end{aligned}$$

$$\begin{aligned} & p(s = 1|d, z, data) \\ & \propto \Big(auto(d) + \alpha_{\lambda_1} - 1\Big) \times \Big(\tfrac{auto(d,z)+inhe(c=d,z)+\alpha_\theta-1}{auto(d)+inhe(c=d)+k\cdot\alpha_\theta-1}\Big), \end{aligned}$$

$$\begin{aligned} & p(z'|w, d, c, s = 0, data) \\ & \propto \Big(auto(c, z') + inhe(c, z') + \alpha_\theta - 1\Big) \times \Big(\tfrac{n(w,z')+\alpha_\phi-1}{n(z')+V\cdot\alpha_\phi-1}\Big), \end{aligned}$$

$$\begin{aligned} & p(z|w, d, s = 1, data) \\ & \propto \Big(auto(d, z) + inhe(c = d, z) + \alpha_\theta - 1\Big) \times \Big(\tfrac{n(w,z)+\alpha_\phi-1}{n(z)+V\cdot\alpha_\phi-1}\Big). \end{aligned}$$

4.4 Motivation Matrix

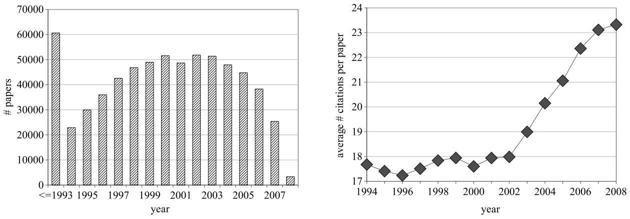
One advantage of ITM is that *c-ITM* can further refine the newly generated topic space by monitoring the inheritance relations among topics. For example, among the k topics in $D(t)$ produced by *c-ITM*, a few topics may not truly exist in $D(t)$ but are instead inherited from $D(t'), t' < t$ via citations. Since we sample words from the inherited and autonomous parts of a document separately, we can similarly separate the topic space $Z(t)$ into two parts: an *inherited* part and an *autonomous* part, to each of which a topic $\mathbf{z}_j(t)$ has a certain probability.

One simple way is to use a $k \times k$ topic motivation (correlation) matrix Q for $D(t)$. Each cell Q_{ij} represents the motivation probability of topic z_i on z_j. Each row sums to be 1. Given document d, a word w in its inherited part d^0 is assigned a topic $\mathbf{z}_i(t) \sim Multi(\psi)$. We can assume that $\mathbf{z}_i(t)$ motives another autonomous topic $\mathbf{z}_l(t) \sim Multi(\theta)$ if

$$l = \arg\max_j \{p(\mathbf{z}_i(t) \rightarrow p(\mathbf{z}_j(t)))\}.$$

The motivation probability relies on how frequently the words in d^0 and $\mathbf{z}_i(t)$ co-occur with the words in d^1 and $\mathbf{z}_l(t)$. As long as $l \neq i$ and $\mathbf{z}_i(t)$ has a same topic in $Z(t-1)$, topic $\mathbf{z}_i(t)$ can be regarded as an inherited topic that is no longer hot in the current topic space. This is reasonable because if $\mathbf{z}_i(t)$ were popular at time t, there should have been many papers in topic $\mathbf{z}_i(t)$ that cite papers from the same topic, so that the motivation probability to itself at time t is still significant. Ideally, diagonal probabilities should dominate the motivation matrix for the topic evolution category of "*same topic*".

Building the topic motivation matrix is straightforward based on LDA as below.



(a) Data volume (b) Average # citations

Figure 6: Dataset Analysis.

- Draw k topic motivation vectors $\mathbf{Q}_z \sim Dirichlet(\alpha_\beta)$, one for each topic in $Z(t)$.
- For each paper d:
  - For each word $w \in d^1$:
    - Draw a motivating topic $\mathbf{z}_i(t) \sim Multi(\psi)$, where ψ is the topic mixture of d^0 generated by *c-ITM*.
    - Given the topic (e.g., $\mathbf{z}_j(t)$) of w assigned by *c-ITM*, draw it from the topic motivation vector of $\mathbf{z}_i(t)$. $\mathbf{z}_j(t) \sim Multi(\mathbf{Q}_{z_i})$.

4.5 Complexity

The time complexity of the four algorithms fully depends on the efficiency of the k-component LDA/ITM. Let N be the dimensionality of the word sampling space $\mathbf{w}$. There are a total of kN parameters to infer during each iteration for the LDA model under Gibbs sampling. Let n be the number of iterations, the time complexity for *i-LDA*, *a-LDA* and *c-LDA* is $O(nkN)$.

For ITM, there are $kN + 2N + \overline{|L_d|} \cdot N$ parameters to be inferred in each duration, where the dimensionality of the indicator vector $\mathbf{s}$ is 2 and $\overline{|L_d|}$ is the average number of citations of each paper in the dataset (i.e., the average dimensionality of the dummy index vector $\mathbf{c}$). Since $\overline{|L_d|}$ is a constant given a document, the k-component ITM model does not grow with the size of the data. The time complexity of *c-ITM* is $O(n(k + 2 + \overline{|L_d|})N)$.

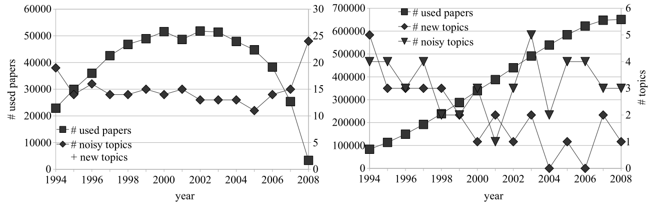
Since both LDA and ITM can be convergent after a limited number of iterations, given n, all four algorithms thus have a linear scalability with respect to N, the only factor that solely relies on the size of data.

5. EMPIRICAL EVALUATION

5.1 Dataset

We tested topic evolution models on the literature archived at CiteSeerX. The dataset contains research papers in computer and information science. We selected papers published in the last 16 years (1993-2008). After removing duplicate papers, papers without explicit publication timestamps, we obtained $650,918$ unique papers dated until early 2008. For each paper, we extracted its title and abstract as content, ignoring the rest. The distribution of number of papers over publication year is shown in Figure 6(a).

We used a year as the time unit in our analysis. The set of papers published in year t $(1993 \leq t \leq 2008)$ is fed into *i-LDA* to learn the topic space of the year. *a-LDA* uses all papers published in or before year t to learn the topic space of the year. For both *c-LDA* and *c-ITM*, we extract all cited papers prior to each year. For simplicity, only 1-hop citations are considered. Please note that only those cited papers in the dataset are used by *c-LDA* and *c-ITM*. Figure 6(b) shows the distribution of the average number of citations per paper over different years.



(a) *i-LDA* (b) *a-LDA*

Figure 8: Correlation of new/noisy topics with data.

For the LDA model, we used the free Mallet tool[2]. We implemented our ITM model in C++. For the hyper parameter settings, $\alpha_\theta = 0.1$, $\alpha_\phi = 0.01$, $\alpha_\gamma = 1.0$, $\alpha_{\lambda_0} = 3.0$, $\alpha_{\lambda_1} = 0.1$ and $\alpha_\beta = 0.1$. All these hyper parameter settings simply follow the tradition of topic modeling [3]. All experiments were conducted on a Linux server with 7 CPU processors of 2.4GHz and 16G memory.

5.2 Evaluating Our Topic Evolution Methods

We evaluated the four topic evolution algorithms for their effectiveness, efficiency, and scalability.

5.2.1 Topic Evolution Categorization

We extracted the top 30 topics by default following the suggestion of [15]. We set the parameters $\epsilon_1 = 0.5$ and $\epsilon_2 = 0.2$. Figures 7(a)-(d) show the distribution of the different types of topic evolution found by *i-LDA*, *a-LDA*, *c-LDA*, and *c-ITM*, respectively. We can obtain some interesting observations.

i-LDA tends to produce the largest average number of *new topics* (10.53) and *noisy topics* (4.4). On average, almost half of the topics (14.93) generated by *i-LDA* are either new or noisy. Refer to Figure 8(a), the smaller amount of data, the more new or noisy topics *i-LDA* generates. For example, as the data volume in years 1998-2005 increases, *i-LDA* shares more topics from the topic spaces in the previous years. However, as the data in year 2008 is incomplete (papers crawled after early 2008 are not included) and thus much smaller (less than 1/10) compared to the other years, almost all generated topics are either new or noisy.

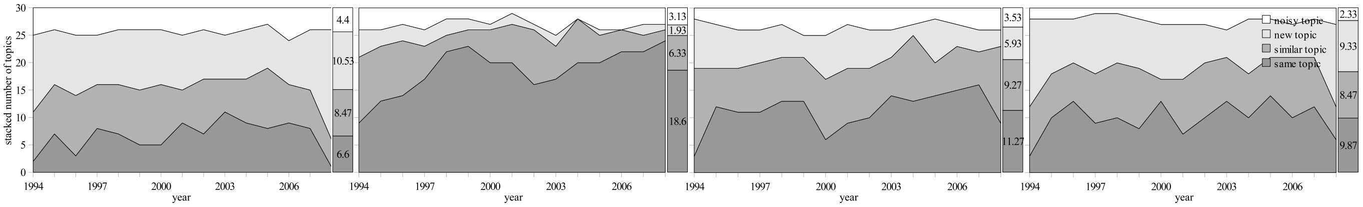
a-LDA is on the other end of the extreme: historical topics tend to dominate the topic space every year. For example, after year 1999, as the accumulation of historical data, the topic space of the current year is almost completely dominated by the previous year's topic space (on average, 2/3 generated topics are *same topics*). In contrast to *i-LDA*, a smaller data volume (of the current year) results in fewer new topics in *a-LDA*. As shown in Figure 8(b), after year 1999, the average number of new topics has a convergence range from 0 to 2. *a-LDA* generates noisy topics without a clear trend: on one hand, the dominance of historical topics can eliminate noisy topics; on another hand, the noisy words contributed to noisy topics are also accumulated along the time.

The two citation-unaware methods suffer from either heavy topic drifting or heavy topic inheritance, both are undesirable for topic evolution. Moreover, both methods are very sensitive to changes in data size.

The citation-aware methods *c-LDA* and *c-ITM* strike a good balance between *i-LDA* and *a-LDA*. Both are less sensitive to changes in data size. Specifically, *c-ITM* tends to generate more *new topics* (9.33 vs. 5.93 on average) and the fewest noisy topics (2.33 on average), while *c-LDA* tends to produce slightly more *same topics* (11.27 vs. 9.87 on average).

The cited papers may boost the importance of some old topics that are no longer hot in the current year. Therefore, the citation-

[2] http://mallet.cs.umass.edu/



(a) *i-LDA* (b) *a-LDA* (c) *c-LDA* (d) *c-ITM*

Figure 7: Categorization of 15-year topic evolution. The right bar of each sub-figure shows the average number of topics fallen into the according category.

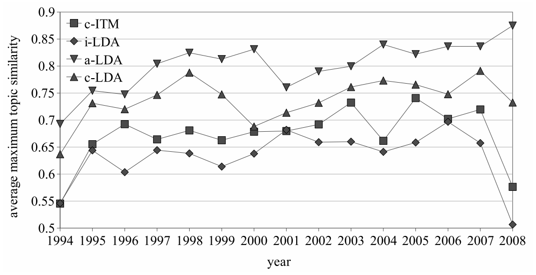


Figure 9: The 15-year topic similarity trend.

aware methods tend to produce more *same topics* and less *new topics*. The disparity can be seen by comparing *c-LDA* to *i-LDA*. This is because using citations may inherit from the historical topic space and thus affect the generation of new topics, especially when new topics appear for the first time.

Fortunately, *c-ITM* is able to reach a good balance. For example, the word space in year 2008 shrinks more than 70% relatively due to the much smaller data volume. Under the same number of partitions, most of topics in year 2008 are too specific and thus different to historical topics. That is to say, most of major historical topics have been lost in year 2008. However, *c-LDA* still generates 23 *same topics* and *similar topics* based on the small word space. Apparently, the cited papers dominate the generation of the topic space in *c-LDA* for year 2008. However, *c-ITM* does not suffer from such a problem; half of topics (15) in year 2008 are new.

Figure 9 shows the trend of average topic similarity. To draw the curves, in year t, a topic $\mathbf{z}$ is matched with a topic $\mathbf{z}'$ in year $t-1$ with the highest topic similarity. Then, the average similarity of a year is the average of similarity of the topics in the year and the matched topics in the previous year. The larger the similarity, the more similar the topic spaces in two consecutive years.

The topic similarity trends tell the differences among our four topic evolution methods: *i-LDA* always has the smallest topic similarities so that topics oscillate the most. *a-LDA* always has the highest topic similarities so that topics tend to retain. *c-LDA* and *c-ITM* stay in the middle yet sometimes *c-ITM* has a bit smaller topic similarities, so that *c-ITM* can generate a bit more *new topics*. Last, when the data volume increases in some year, the differences among the four methods become smaller.

5.2.2 Filtering Ghost Topics

Although *c-ITM* strikes a good balance between *new topics* and *same topics*, some old topics that have been declining in the current year still may be inherited along the citations. We can *optionally* build the topic motivation matrix to filter the topic space produced by *c-ITM*. We used year 2006 as an example to generate the topic motivation matrix (27×27 after removing 3 noisy topics), as shown in Figure 10.

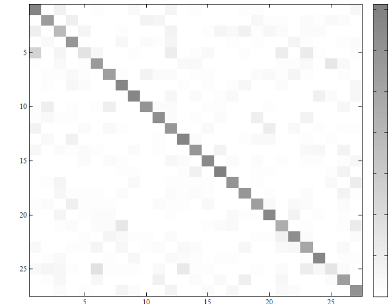


Figure 10: $Topic \times Topic$ motivation matrix in 2006.

Table 2: Top topic motivation probabilities for ghost topics in 2006.

cited topics	citing topics	probability
topic 5 mining patterns	topic 1, clustering similarity	0.2455
	self	0.1657
	topic 23, graph algorithms	0.1173
	topic 21, logic fuzzy	0.1115
	topic 12, streams information	0.1026
topic 25 coding compression	topic 6, quantum complexity	0.1685
	self	0.1520
	topic 13, memory cache	0.1319
	topic 23, graph algorithms	0.109

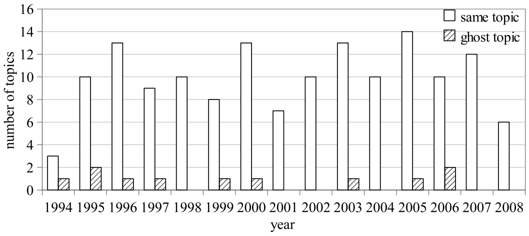


Figure 11: Distribution of detected ghost topics.

Among the 10 *same topics*, we identified two ghost topics (topics 5 and 25) that exist in the previous year topic space, and have high motivation probability to the other topics, but low motivation probabilities to themselves. These ghost topics exist only in the cited papers. Table 2 shows in detail how the two ghost topics were cited by other valid 2006 topics. Limited by space, only the top 2 words were used to represent each topic without manual labels.

Figure 11 shows the number of ghost topics yearly. Compared to the number of *same topics*, only a small portion of *same topics* are ghost topics. *c-ITM* did not generate many ghost topics.

5.2.3 Topic Evolution Case Study

To better understand the topic evolution process, here we present some real topic evolution examples related to the category of *image*

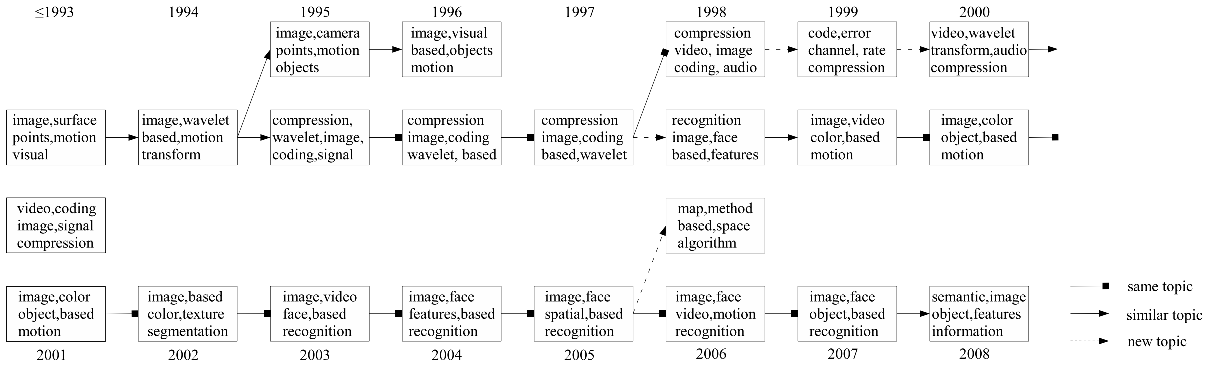

≤1993
1994
1995
1996
1997
1998
1999
2000
image,camera points,motion objects
image,visual based,objects motion
compression video, image coding, audio
code,error channel, rate compression
video,wavelet transform,audio compression
image,surface points,motion visual
image,wavelet based,motion transform
compression, wavelet,image, coding,signal
compression image,coding wavelet, based
compression image,coding based,wavelet
recognition image,face based,features
image,video color,based motion
image,color object,based motion
video,coding image,signal compression
map,method based,space algorithm
image,color object,based motion
image,based color,texture segmentation
image,video face,based recognition
image,face features,based recognition
image,face spatial,based recognition
image,face video,motion recognition
image,face object,based recognition
semantic,image object,features information
2001
2002
2003
2004
2005
2006
2007
2008
same topic
similar topic
new topic

Figure 12: Topic evolution examples related to image processing.

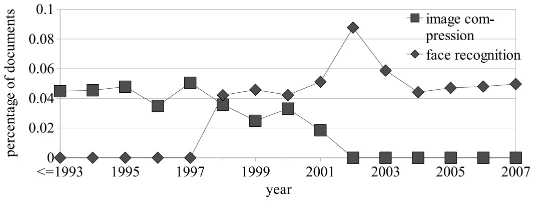


Figure 13: Topic strength trend for image-related topics.

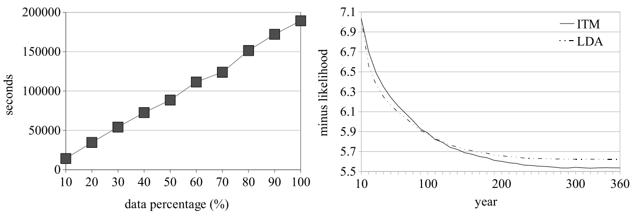
processing, as shown in Figure 12. Each topic is described using the top 5 words without any human labels.

There are two main topics related to image processing: *image compression* which mainly evolved from 1994 to 2001, and *face recognition* which mainly evolved from 1998 to 2007. Specifically, *image compression* evolved from the topic *image surface* in 1994 and subsequently evolved into the new topic *face recognition* in 1998. In 1998, *image compression* further evolved from *static image compression* to *video compression*. Except for the year 1999 in which *video compression* was suddenly interrupted by *channel coding* which was still hot for image compression, we believe the other evolutions are consistent and reasonable. We can also conclude that *wavelet coding* is a very important tool for both *static image compression* and *video compression*, rather than others like *channel coding*.

Figure 13 further depicts the topic strength trends for two main topics: *image compression* and *face recognition*. In Figure 13, we simply group *static image compression*, *video compression* and *channel coding* into the same topic named *image compression*, discard the other unimportant branches like *camera objects* etc., and assume that year 2008 does not have enough data to support the topic evolution in image processing. The topic strength trend clearly tells how these two topics evolve over the time. Interestingly, when *image compression* reaches the bottom in 2002, the topic *face recognition* also reaches its peak at the same time.

5.2.4 Scalability and Time Efficiency

We analyze the scalability and time efficiency of our topic evolution methods. The scalability of LDA has been tested in many previous work. Here, we only test the *ITM* model. The total number of word occurrences N across 16 years in our dataset is $42,389,066$. Accordingly, we sampled 10%, 20%, $\ldots$, 100% of the word occurrences to test the scalability. Note that except for *a-LDA*, we will never have a chance to use 100% of all word occurrences. We showed that the time complexity of *ITM* is linear with respect to the number of word occurrences, the number of iterations, and the size of topic space. Figure 14(a) further verifies our claim with $k = 30$ and $1,000$ Gibbs sampling iterations.



(a) Scalability of *ITM* (b) Convergence rates

Figure 14: Scalability and convergence rate of models.

Previous work (e.g., [14]) reported that LDA under Gibbs sampling normally requires around 500-$1,000$ iterations to reach convergence. Here, we also compared the convergence rate for two topic models: *ITM* and *LDA* under Gibbs sampling. We used the minus likelihood of the data to measure how our model fits the data (the whole word space $\mathbf{w}$), which is defined as

$$p(\mathbf{w}) = -\frac{1}{N} \cdot \sum_{w_i \in \mathbf{w}} logp(w_i). \qquad (7)$$

Suppose that d is the document from which the word w_i originates,

$$p(w_i) = \sum_{z \in Z} \phi_z(w_i)[\theta_d(z)p(s=1) + \sum_{d' \in L_d} \theta_{d'}(z)\gamma_{d'}p(s=0)].$$

Figure 14(b) shows the convergence rate of models. After about 100 iterations, the likelihood of the data stabilizes and does not change significantly for both models. Overall, *LDA* converges a bit faster and stabilizes after 200 iterations. Instead, *ITM* cannot improve the likelihood further after 300 iterations. But after convergence, *ITM* has a higher likelihood. The result indicates that the convergence speed of our model is comparable to *LDA* under the Gibbs sampling; and our model fits the citation graph data better.

Lastly, we tested the running time of all topic evolution methods with $k = 30$, as shown in Figure 15. For a fair comparison, we ran $1,000$ iterations for each method. The running time of all methods grows/declines linearly as the data volume (refer to Figure 6) and the word sampling space increase/decrease. Under the same data distribution, *c-ITM* is slower than *c-LDA* as the former needs to infer 2 additional latent variables.

5.3 Comparison with Previous Topic Evolution Methods

In this section, we compare our topic evolution results with the previous topic evolution work on the CiteSeerX data, which is narrowed to Jo *et al.* [16] (denoted by *term-graph*), Zhou *et al.* [33] (denoted by *author-interaction*) and Bolelli *et al.* [6, 7] (denoted by

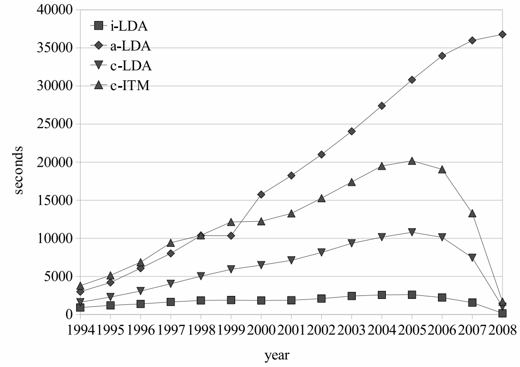


Figure 15: Comparisons of running time.

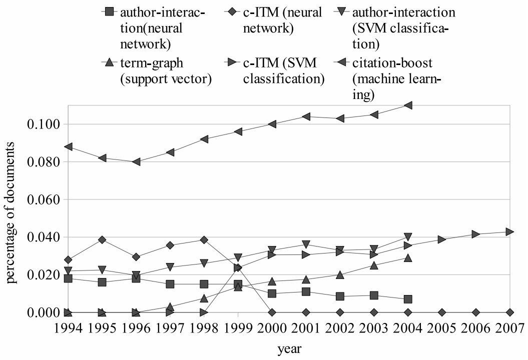


Figure 16: Comparing the topic evolution for topics related to *machine learning*.

citation-boost). Topics were examined from 1991 to 2004 in [33], from 1994 to 2004 in [16], and from 1990 to 2004 in [6, 7].

5.3.1 Comparison Case Study

We use the topic category "machine learning" as the example to compare with previous work within the common time periods 1994-2004. The *term-graph* generated thousands of topics based on terms, yet the other three only only produced k topics each time; so that much fewer documents were grouped to topics by *term-graph* then others. We multiply the percentage of documents by five for the topics generated by *term-graph*, *only* for a comparable visualization.

Figure 16 visualizes the evolution of topics related to *machine learning* for all methods. The topics *neural network* and *SVM classification* are related. In our method, we detected *neural network* at the very beginning. It started to decline in 1998 and evolved to *SVM classification* around 1999-2000. In 1999, the top 5 words of the according topic in our model are *learning, neural, models, network and classification*, which can be seen as a mixture of *neural network* and *classification*. We can thus treat the year 1999 as a transitional period between these two topics; and the mixed topic *learning, neural, models, network and classification* as the according transitional topic. This transitional topic was announced as *similar topic* by our model. Before year 1999, *neural network* was hot; after year 1999, *SVM classification* was uplifted to the hot topic list. We simply assumed that these two topics co-existed with the same weight in 1999 to highlight such a transitional period in Figure 16.

In *author-interaction*, *neural network* has a decreasing trend and *SVM classification* has an increasing trend, both from the beginning to the end. After considering the impact of authors, *author-interaction* found the reason for the declining of *neural network* and increasing of *SVM for classification*: some authors worked on *neural network* might move to the area of *SVM classification*. This finding is consistent with our results.

In *term-graph*, the topic *neural network* was not found and a topic denoted by the term *support vector* was found with an increasing trend as the term *support vector* at the 5th position in the list of top topics since 2000. In *citation-boost*, a more general topic labeled by *machine learning* was found in the very beginning. It first has a decreasing trend until year 1996; after that, its topic strength increases.

Based on the above results, we conclude a few findings to differentiate our method from the previous work.

- All methods except for *citation-boost* have consistent topic strength trend: *neural network* declines before 1999 and *SVM classification* boosts after 1999. It is not clear why *citation-boost* has an inconsistent concave point in 1996 (partially because *machine learning* covers other unknown topics).
- Only our method and *term-graph* are able to tell cause and effect for two related topics, where one topic evolves from another one. Our method finds out such hot topic transition through a transitional topic (it is very likely in the topic evolution category *similar topic* or *new topic*), yet *term-graph* bridges two related topics by counting the common authors.
- The topics found by *term-graph* are more fine-grained (with fewer in-topic documents). In the contrary, *citation-boost* generated rather general topics (with more in-topic documents). Both methods cannot find out pairwise relations for topics.

5.3.2 Comparison on Top 20 topics

Since no benchmark topics exist for topic evolution, we evaluate the quality of the automatically detected topics by comparing them to the top manually-confirmed topics found by *term-graph*. The *term-graph* offered two ranking lists for topics before and since 2000. Those top topics of each list are the most frequent terms happened in each time period. Thus, *term-graph* provided a benchmark for evaluating the topic evolutions. For example, if an important topic has evolved from the past after 2000, it might appear in the top list of topics since 2000 only (not in the list before 2000). We picked the top 25 topics from the list since 2000, and removed those topics that also appeared in the top 100 of the list before 2000. In the end, we got 20 topics in total that did not appear in top 100 list before 2000 but appear in the top 25 list since 2000, as shown in Table 3. These 20 topics are then treated as the benchmark that evolved from the part and newly became hot after 2000.

In Table 3, we checked the meaningfulness of these benchmark topics manually, as well as how these topics evolved over our method and the other two: *author-interaction* and *citation-boost*. We found that only 12 benchmark topics are proper hot topics since 2000. Among the rest 8 topics, 4 topics were detected repeatedly and another 4 are too specific to be proper topics (they might be specific methods rather than topics). This result indicates that the enormous topic space (thousands of topics) produced by *term-graph* is not clean.

Based on Table 3, we made the following conclusions.

- Assuming that *term-graph* has the highest recall (100%), our method successfully detected most of important topic evolutions with a recall of 91.67% (11 over 12), only missing 1 topic *image retrieval* which was covered by the topic *information retrieval*. The *citation-boost* has a recall of 58.33% by missing 5 topics. The *author-interaction* has a recall of 25% by detecting 3 topics.
- Among the hit topics, our method and *author-interaction* have a finer grain. The generated topics match the most frequent terms well. However, *citation-boost* only produced coarse topics. Based on *citation-boost*, it is not clear how topics evolved from one topic to another in detail.

Table 3: Evolutions of top 20 topics ranked by [16] since 2000.

top topics [16]	human label	author-interaction [33]	c-ITM	citation-boost [6, 7]
sensor networks	√	N/A	evolved from system model since 2002	network, increasing trend
hoc networks	√	evolved from network community	evolved from network communication since 1994	network, increasing trend
image retrieval	√	N/A	hidden in information retrieval from 1993	N/A
support vector	√	evolved from neural network	evolved from neural network since 2000	machine learning, increasing trend
decision diagrams	too specific	N/A	N/A	N/A
wireless sensor	√	evolved from network community	evolved from sensor networks since 2003	network, increasing trend
ad hoc	√	N/A	evolved from network routing since 2003	network, increasing trend
intrusion detection	√	N/A	evolved from protocol security since 2000	N/A
vector machines	duplicated	-	-	-
mobile ad	duplicated	-	-	-
binary decision	too specific	N/A	N/A	N/A
sensor network	duplicated	-	-	-
energy consumption	too specific	N/A	N/A	N/A
content-based image	√	N/A	evolved from video compression since 2001	N/A
semantic web	√	N/A	evolved from knowledge ontology since 2002	web data analysis, increasing trend
fading channels	√	N/A	evolved from channel coding since 2004	N/A
xml data	√	N/A	evolved from database since 2003	database, increasing trend
source separation	too specific	N/A	N/A	N/A
signature scheme	√	N/A	evolved from protocol security since 2004	N/A
xml documents	duplicated	-	-	-

√: proper hot topic, too specific: too specific to be a topic, duplicated: the topic has been listed before, N/A: not found, -: no need to compare again

- In *term-graph*, only the top topics are meaningful as *term-graph* generated many specific topics. Heavy duplications also exist in *term-graph*. Instead, the other three work only produced general topics without ranking. Only *author-interaction* and our method produced pairwise topic relations. Compared to *author-interaction*, our method can tell the exact boundary of topic evolution, but in *author-interaction*, topics were spanned over the whole time range.

6. CONCLUSIONS

In this paper we studied the topic evolution problem for scaled scientific literature. We first investigated the citation-unaware approaches based on the LDA model, along with their limitations on topic evolution, i.e., the correlated topics were generated independently. We then proposed the citation-aware approaches for topic evolution. Moreover, an iterative topic learning framework based on citation network was presented to fully utilize the impact of citations. A novel Inheritance Topic Model was then naturally proposed for this learning process. Our algorithm can be quickly convergent under the Gibbs sampling and has a linear scalability with respect to the size of dataset. The experimental results show that our approach can track the topic evolution in a large dataset containing more than 650,000 papers over 16 years. The experimental results clearly indicate that citations are able to portray the inherent dependence among correlated topics, and citation-aware approaches are thus good choices for tackling the sequential topic evolution problem.

7. ACKNOWLEDGMENTS

This material is based upon work supported by the National Science Foundation under Grant Nos. 0535656 and 0845487.

8. REFERENCES

[1] J. Allan. Topic detection and tracking. event-based information organization. *Kluwer Academic Publishers*, 2002.

[2] L. AlSumait *et al.* On-line LDA: Adaptive Topic Models for Mining Text Streams with Applications to Topic Detection and Tracking. In *ICDM'08*.

[3] D. Blei *et al.* Latent dirichlet allocation. *Journal of Machine Learning Research 2003*.

[4] D. Blei and M. Jordan. Modeling annotated data. In *SIGIR'03*.

[5] D. Blei and J. Lafferty. Dynamic topic models. In *ICML'06*.

[6] L. Bolelli *et al.* Finding Topic Trends in Digital Libraries. In *JCDL'09*.

[7] L. Bolelli *et al.* Topic and Trend Detection in Text Collections using Latent Dirichlet Allocation. In *ECIR'09*.

[8] R. Brown *et al.* Link detection results and analysis. In *1999 TDT-3 Evaluation Project Workshop 1999*.

[9] F. Chen *et al.* Story link detection and new event detection are asymmetric. *HLT-NAACL 2003*.

[10] D. Cohn and T. Hofmann. The missing link - a probabilitstic model of document content and hypertext connectivity. In *NIPS'01*.

[11] L. Dietz *et al.* Unsupervised prediction of citation influences. In *ICML'07*.

[12] E. Erosheva *et al.* Mixed-membership models of scientific publications. *PNAS 2004*.

[13] A. Gohr and A. Hinneburg. Topic Evolution in a Stream of Documents. In *SDM'09*.

[14] T. Griffiths and M. Steyvers. Finding scientific topics. *PNAS 2004*.

[15] G. Heinrich. Parameter estimation for text analysis. *Technical Note, University of Leipzig, 2008*.

[16] Y. Jo *et al.* Detecting research topics via the correlation between graphs and texts. In *SIGKDD'07*.

[17] G. Mann *et al.* Bibliometric impact measures leveraging topic analysis. In *JCDL'06*.

[18] A. McCallum *et al.* Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research 2007*.

[19] Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *SIGKDD'05*.

[20] Q. Mei *et al.* A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *WWW'06*.

[21] Q. Mei *et al.* Topic modeling with network regularization. In *WWW'08*.

[22] F. Morchen *et al.* Anticipating annotations and emerging trends in biomedical literature. In *SIGKDD'08*.

[23] S. Morinaga and K. Yamanishi. Tracking dynamics of topic trends using a finite mixture model. In *SIGKDD'04*.

[24] R. Nallapati *et al.* Joint latent topic models for text and citations. In *SIGKDD'08*.

[25] D. Newman *et al.* Statistical entity-topic models. In *SIGKDD'06*.

[26] G. Salton and C. Buckley. Term-weighting approaches in automatic text retrieval. *Information Processing and Management 1988*.

[27] R. Schult and M. Spiliopoulou. Discovering emerging topics in unlabelled text collections. In *Proc. East European ADBIS Conference 2006*.

[28] M. Spiliopoulou *et al.* Monic: modeling and monitoring cluster transitions. In *SIGKDD'06*.

[29] M. Steyvers *et al.* Probabilistic author-topic models for information discovery. In *SIGKDD'04*.

[30] Y. Teh *et al.* Hierarchical dirichlet processes. In *Technical report, UC Berkeley Statistics TR-653, 2004*.

[31] C. Wang *et al.* Continuous time dynamic topic models. In *UAI'08*.

[32] C. Zhai and J. Lafferty. A study of smoothing methods for language models applied to information retrieval. *TOIS 2004*.

[33] D. Zhou *et al.* Topic evolution and social interactions: How authors effect research. In *CIKM'06*.