Understanding Evolution of Research Themes: a Probabilistic Generative Model for Citations

Xiaolong Wang, Chengxiang Zhai and Dan Roth
Department of Computer Science
University of Illinois, Urbana-Champaign
Urbana, IL
{xwang95, czhai, danr}@illinois.edu

ABSTRACT

Understanding how research themes evolve over time in a research community is useful in many ways (e.g., revealing important milestones and discovering emerging major research trends). In this paper, we propose a novel way of analyzing literature citation to explore the research topics and the theme evolution by modeling article citation relations with a probabilistic generative model. The key idea is to represent a research paper by a "bag of citations" and model such a "citation document" with a probabilistic topic model. We explore the extension of a particular topic model, i.e., Latent Dirichlet Allocation (LDA), for citation analysis, and show that such a Citation-LDA can facilitate discovering of individual research topics as well as the theme evolution from multiple related topics, both of which in turn lead to the construction of evolution graphs for characterizing research themes. We test the proposed citation-LDA on two datasets: the ACL Anthology Network (AAN) of natural language research literatures and PubMed Central (PMC) archive of biomedical and life sciences literatures, and demonstrate that Citation-LDA can effectively discover the evolution of research themes, with better formed topics than (conventional) Content-LDA.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Experimentation

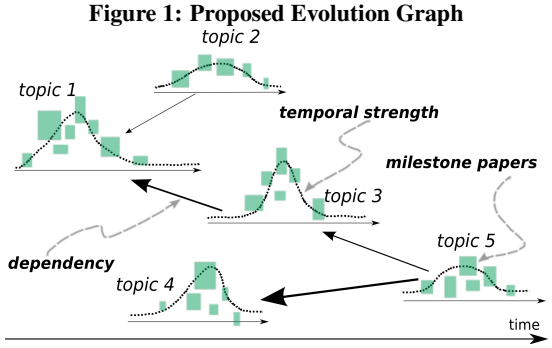
Keywords

theme evolution, citation analysis

1. INTRODUCTION

How to leverage information technologies to improve the productivity of scientific research is a highly important challenge with clearly huge impact on the society. One bottleneck in research productivity is that as a research community grows, it would be increasingly difficult for researchers to see the complete picture of how a field has been evolving, given the fact that large volume new literatures are written based on previous works. Junior researchers can often get lost in the overwhelming amount of related papers. Researchers who seek to shift to a new topic may spend lots of time preparing a reading list on his/her own. All these clearly hinder the progress of scientific research, and it would be highly beneficial to develop mining techniques to help researchers more easily and more efficiently understand research themes in scientific literature. In general, two aspects of analysis are needed for understanding research themes: First, we need to analyze *each research topic* to answer the following questions: Which papers are the milestone papers that best represent a topic and how to quantify their impact? When did the topic become popular and is it still attracting attention today? Can the topic be summarized accurately with a few keywords? Furthermore, when *investigating topics collectively*, which are the most dominant topics extensively studied? During the evolution, what are the newly generated topics initiated by the old one? Can we identify the underlying evolution patterns among topics?

Figure 1: Proposed Evolution Graph



To answer the questions raised above, ideally, we would like to automatically construct a *"research theme evolution graph"*, which we illustrate in Figure. 1. With such a graph, when zooming into the scope of individual topics, multiple types of information are provided to facilitate users to understand the research topic:

- *Topic Milestone Papers*: It is critical to recognize the papers that are best representative for a topic in the course of understanding topics. We refer to them as "topic milestone papers". Milestone papers of a topic provide a good picture how a topic is formed. In Figure. 1, milestone papers are shown in each topic as rectangles and the "size" reflects their importance with respect to topics.
- *Topic Temporal Strength*: The relative popularity of topics at different times reveals the temporal nature of topics, which can help users to identify *current* vs. *previous* research topics as well as the rough topic life spans. Intuitively, when many milestone papers occur, the topic draws more attention and becomes popular.
- *Topic Keywords*: Extracting keywords that can properly summarize a topic would enable users to obtain a brief idea about the

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.

topic even without reading its relevant papers, allowing users to fast navigate among topics in search of the most interesting ones.

While zooming out to see the big picture of all related topics in the theme, there is also meaningful information to explore:

- *Topic Importance*: Quantifying the importance of topics helps a user to discriminate the *major* vs. *minor* topics in a research theme. Topic importance also reflects how well the topic is recognized by the community.
- *Topic Dependency*: Many new topics are built on top of the old ones. Discovering the dependency relation between topics provides a good guidance for users when searching for *origin/continuing* topics. In Figure. 1, we visualize the dependency strength between topics by the "thickness" of edges.
- *Evolution Patterns*: Connecting topics by their dependency illustrates the underlying evolution patterns for research themes. Is there any trend that different topics get merged together to form a new (interdisciplinary) topic, such as Topic 3 and Topic 4 are merged into Topic 5? Or is there a general topic branched into multiple topics that address specialized problems, such as Topic 1 has led to Topic 2 and Topic 3?

To automatically construct such an evolution graph as shown in Figure. 1, the two major computational tasks are:

- ***Discovering the research topics***, which includes finding milestone papers, computing the temporal strength, and extracting keywords for each individual topic.
- ***Discovering the theme evolution***, which includes identifying the topic importance and learning the dependency relation between topics, as well as recognizing the underlying evolution patterns.

Existing approaches, notably those of topic modeling, can generate some (not all) of these components in the evolution graph, but they are far from adequate for the following reasons: First, though there are many works that aim to construct evolution map over time, they rely on pre-segmentation of text streams into fixed time windows, due to either computational issue [2, 16, 24] or modeling issue [23]. Consequently, the topic evolution result would be inevitably sensitive to the choice of temporal granularity of how time is discretized and sliced. Suboptimal granularity of time might result in missing important topics or even lead to inaccurate evolution analysis. Second, the edges in most of the existing evolution graphs, do not reflect the *dependency relation* between topics, and can only reveal the *topic similarity* and *correlation* [2, 3, 16, 23]. The fundamental limitation is that content-based topic modeling approaches are built on *word co-occurrence*, which essentially is *undirected* unlike the dependency relation. Third, it is difficult for any aforementioned models (including Pairwise Link-LDA [17]) to assess the impact of documents with respect to different topics, i.e., identifying the milestone papers. Their approaches model topics as distributions over words, and although the text similarity between document and topic can be computed, it would be a substantially different measurement from the document *impact* on a topic.

As hinted above, a major reason why existing topic models are insufficient is that they have not fully exploited citation relations to discover topics. In this paper, we address these limitations by doing joint analysis of citations and text. Indeed, we will rely more on citation links than on document content, which makes our work different from [17] and all others. Specifically, we leverage a similar idea to topic modeling and analyze the citation graphs in a *probabilistic* manner. We directly model the generation of citations, which are direct evidence related to *"impact"* of document as well as *"dependency"* between topics. Through citation generation, we are enabled to address the core problem of assessing milestone papers based on impact, and estimating the topic dependency. More importantly, our key insight here is that "co-cited papers" are good indicators of research topics, more effective than relying on text similarity as in most existing work. Empirical study [6] has already noticed that it is a subjective yet difficult task to annotate for each word its belonging topic even manually. However, for citations in a published paper written by experienced authors, it would be much easier to determine the topic since most authors make citations prudently and thus citation is much *less noisy* than text.

To discover topics based on citations, we propose a novel probabilistic approach to analyze citations by viewing citation graphs as a set of "citation documents" where each is a research paper represented as a *"bag of citations"*. A paper that cites k other (possibly duplicated) papers would simply be viewed as a *"document"* with k *"tokens"*, each corresponding to the ID of a cited paper. With this view, we can model all these citation documents with a generative topic model where we introduce latent topic variables over the citations. This is analogous to the application of a probabilistic topic model to model topics in text documents, but with the important difference that the discovered topics with our model would be characterized by a (multinomial) *distribution over research papers*, rather than over words as in conventional content-based topic models. In addition, when combined together with additional information, particularly the *published time* and the *title* of each paper, our model can address the computational tasks of discovering both *the research topics* and *the theme evolution*, and constructing *the evolution graph* as well.

In the rest of the paper, we first review some of the related work in Section. 2, which is followed by presenting our probabilistic model for literature citations in Section. 3. After the derivation about one specific model Citation-LDA, we focus our discussion on how to construct the theme evolution graph in Section. 4. Experiment setup and extensive evaluation results will be given in Section. 5. Finally, we conclude our work with future direction in Section. 6.

2. RELATED WORK

In recent years, many literature search engines as well as digital libraries have come into use, including Microsoft Academic Search [1], Google Scholar [2], DBLP [3] and ACM Digital Library [4]. They provide knowledge about scientific literatures through ranking and search interface, which in turn, relies on algorithms that utilize citation-related indicators such as H-index [13] and Impact Factor [9].

In the research community, one thread of study treats scientific literature as citation graphs. To assess the importance of papers, graph ranking algorithms such as PageRank and its variants have been applied [10, 20, 21, 22]. In [10], the authors further take time into consideration in order to overcome the recency bias that favors "old" papers. Apart from this, graph clustering is investigated to identify meaningful topics, such as [5, 8, 18, 19]. In [18], it is pointed out that efficient graph clustering can be combined with temporal information to identify the trends of topics in literature. Particularly, one recent paper [15] is close to our work. It leverages both citation and text (title and abstract) to generate the evolution map in computer science community. Specifically, their method relies on the temporal order of papers and the document language model to detect the formation of new topics, and then it computes the strength between two topics with the "cross citation count" (total citation numbers between the two topics), which however ignores the directed relation of topic dependency. Their method is difficult to be applied to address our problem because their method does not distinguish the difference in topic importance, nor does it recognize milestone papers through assessing the impact based on citations.

[1] http://academic.research.microsoft.com/

[2] http://scholar.google.com/

[3] http://www.informatik.uni-trier.de/ ley/db/

[4] http://dl.acm.org/

While on the other hand, existing probabilistic topic modeling over text [4, 11, 14] has been throughly studied, treating documents as mixtures of latent topics. Early attempt in modeling the topic evolution [16] investigates the Probabilistic Latent Semantic Index (PLSI) [14] to extract topics and models the evolution process as transitions between topics in Hidden Markov Model (HMM). Later, Topic Over Time (TOT) model [24] is developed based on Latent Dirichlet Allocation (LDA) [4]. The key difference between between LDA and TOT is that TOT explicitly assumes time as generated from topics, which jointly models time and word, thus enabling itself to discover time-aware topics as well as topic temporal strength. Besides, Dynamic Topic Models [2, 23] address the problem of topic evolution by modeling topics (distributions over words) changing over time. In the discrete case [2], topics at the next time-stamp deviate from the current ones by a Gaussian noise; while, in the continuous case [23], the change of topics over time is generalized as Brownian motion. One limitation of these models [2, 16, 23, 24] is that they all rely on the pre-segmentation of time: without appropriate time granularity selected, they could fall into difficulty in finding important topics. Ideally, the selection of correct time span should be made automatically. In addition to these studies, others consider the problem of modeling topic correlation [3] and document hyperlink generation [7], for which the essential difficulty is that they cannot model the *"dependency"* relation between topics. The only exception we are aware of so far is the paper [17] which jointly models text and citation generatively. One of its proposed model, named "Pairwise Link-LDA", explicitly includes the topic dependency as model parameters by extending the idea of mixed-membership block stochastic models [1]. In words, the chance of generating a particular citation is determined by the topics of citing and cited documents, which indeed addresses the topic dependency directly. Nevertheless, the Pairwise Link-LDA is not able to fulfill all the tasks we listed such as recognizing the milestone papers and so on.

To our best knowledge, there is no existing approach that can address all the questions as we raised before, i.e., the discovery of *research topics* and *theme evolution*. To this end, we directly model the generation of the citation links among literatures in this paper. In the same spirit of topic modeling, citations are generated stochastically according to a distribution with respect to the underlying topic. It is worth noting that applying the topic modeling approaches to study graphs was previously investigated for discovering communities from coauthorship networks in [12, 25] [5]. Nevertheless, our model not only discovers the topics, but also explores their dependency relationships and yields meaningful knowledge about the evolution of topics.

3. PROBABILISTIC MODELING OF LITERATURE CITATIONS

In contrast to most existing work on citation analysis, where citations are often modeled as network or graph, we propose to represent citation graph as a set of "citation documents" where each is a research paper represented as "bag of citations", and model these citation documents with a probabilistic generative model. Such a new approach has several advantages over pure graph analysis methods. First, by using a latent topic variable, we can naturally associate topics with papers and citations, enabling ranking the paper based on citation within each topic, through which milestone papers can be identified. Second, by modeling the whole set of papers in a field, we can obtain a set of topics that summarize well the major research topics in the field, with (probabilistic) weights quantifying their importance. Third, by estimating the topic level citation structure, it is possible to compute the strength of dependency relation between topics and picturing the evolution paths of research themes. Last, distribution over papers for each topic obtained by such a model can be easily used to compute a distribution over time or keywords when used together with other information such as paper published time and title, allowing modeling the topic temporal strength and summarizing topics with keywords.

Compared with pure content-based topic models, our use of topic model is entirely on capturing topics through citation structures, roughly corresponding to discovering topics based on *co-citation relation*, which is intuitively more accurate in finding research topics: if there is a *"stable"* set of *"core papers"* that are often cited together, then it generally indicates the existence of a major research topic and the core papers are actually *milestone papers* in that topic. Specifically, we use a probabilistic model to explain how an author generates the references (citations) for a paper (which we may also refer to as a document for convenience sometimes). More specifically, given a paper, he/she would "generate" all the references cited in the paper independently. When generating each citation, the author would first sample a topic according to a document-specific topic distribution (doc_topic distribution), and then draw a reference document to cite from the citation distribution of the sampled topic ($topic\_doc$ distribution). One may easily notice that such a generation process is essentially similar to the one over words for documents assumed in probabilistic topic models for text data. Indeed, our work is a novel way of using topic models for citation analysis, and just as topic models are very effective for discovering and analyzing topics in *text documents*, our model can also be very useful for discovering and analyzing topics in *scientific literatures* where the citation graph is available. Another advantage over content-based topic models we may anticipate is that the computational complexity is greatly reduced because the number of citations is much less than the number of words in the corpora.

3.1 The General Model

Formally, suppose each document d cites a subset of other documents $\{c_t\}$ $(t = 1, 2, \ldots)$, where c_t is a cited reference. We assume the following generation process for a citation that links to document c_t in document d (i.e., document d cites document c_t):

- draw topic sample: $z_t \sim D_{doc\_topic}(z; d)$
- draw citation sample: $c_t \sim D_{topic\_doc}(c; z_t)$

The doc-topic distribution $D_{doc\_topic}(\cdot; d)$ and topic-doc distribution $D_{topic\_doc}(\cdot; z)$ are parameterized by the citing document d and the topic z respectively, and are the two key components in the model that would enable many interesting ways to analyze topics and evolution relations among topics. Indeed, $D_{doc\_topic}(\cdot; d)$ gives us a probability distribution over (latent) topics conditioned on document d, and can be interpreted as the *topic coverage* in document d when generating citations, whereas $D_{topic\_doc}(\cdot; z)$ gives a *"reverse"* conditional distribution of documents given a topic, and can be interpreted as how a topic is characterized by a set of papers (documents) that are cited. Thus if a document c_i has a higher probability than c_j according to $D_{topic\_doc}(\cdot; z)$, it would suggests that c_i better characterizes topic z than c_j, or c_i represents topic z better as being a more important paper with higher impact upon z than c_j. With such a distribution over papers, we can easily compute the *expected time* for a topic based on the time when the paper was published as well as the *topic keywords* based on the paper titles (or abstracts if available). Note that a substantial advantage of such a probabilistic model is that it can *"decode"* why document d cites document c_t by inferring the latent topic associated with this citation relation and quantifying with uncertainty, which enables "disambiguation" of citation relations to some extent. As will be further discussed, we can use such a model to perform the computational analysis for discovering research topics and theme evolution, which finally lead to the construction of evolution graph as proposed in Figure. 1.

[5] We thank the anonymous reviewer for pointers to these works

3.2 Citation-LDA

Though we may have different ways to refine the general probabilistic model defined above, in this paper as a first step, we focus on exploring the use of the basic Latent Dirichlet Allocation (LDA) [4] model, which we call "Citation-LDA" and show that even with this simple model setting, we can already discover a lot of interesting knowledge that is useful for understanding research theme evolution.

Specifically, Citation-LDA assumes that D_{doc_topic} and D_{topic_doc} are multinomial distributions with parameters drawn from conjugated Dirichlet prior α and β respectively [6]. We follow the convention to denote $D_{doc\_topic}(\cdot; d)$ and $D_{topic\_doc}(\cdot; z)$ by θ_d and ϕ_z respectively, and we have: $\theta_d \sim \mathrm{Dir}(\alpha)$ and $\varphi_z \sim \mathrm{Dir}(\beta)$. The citation generation process for document d_{i^*} is:

- sample a topic $z = k^* \sim \mathrm{Multi}(\theta_{i^*})$
- sample a document to cite $c = d_{j^*} \sim \mathrm{Multi}(\varphi_z)$

We use the collapsed Gibbs sampling [11] to make inferences with the model. The sampling is initialized by assigning random topic labels $\{z\}$ and updates each of them iteratively. In particular, for the t-th citation that links to d_{j^*} in document d_{i^*}, the topic assignment is updated according to the probability [7]:

$$\begin{aligned} & \Pr(z = k^* | c_{i^*,t} = d_{j^*}, Z_{\neg(i^*,t)}, C^{\neg(i^*,t)}) \\ \propto \quad & \left(\alpha_{k^*} + \#^{\neg(i^*,t)}(z = k^*, d = i^*)\right) \\ & \times \frac{\beta_{j^*} + \#^{\neg(i^*,t)}(z = k^*, c = d_{j^*})}{\sum_j \beta_j + \#^{\neg(i^*,t)}(z = k^*, c = d_j)} \end{aligned} \tag{1}$$

The sampling converges to the true posterior distribution after the burn-in stage [8]. Posterior expectation of θ_{i^*,k^*} and φ_{k^*,j^*} is given by [9]:

$$\hat{\theta}_{i^*,k^*} = \left\langle \frac{\#(d = i^*, z = k^*) + \alpha_{k^*}}{\sum_k \#(d = i^*, z = k) + \alpha_k} \right\rangle \tag{2}$$

$$\hat{\varphi}_{k^*,j^*} = \left\langle \frac{\#(z = k^*, c = j^*) + \beta_{j^*}}{\sum_j \#(z = k^*, c = j) + \beta_j} \right\rangle \tag{3}$$

In addition, the empirical posterior distribution over topics can be computed as:

$$\hat{\Pr}(z = k^* | C) = \left\langle \frac{\#(z = k^*)}{\sum_k \#(z = k)} \right\rangle \tag{4}$$

4. CONSTRUCTION OF THEME EVOLUTION GRAPH

The results obtained from Equation. 2 - 4 form the basis for exploring the knowledge that leads to the construction of the evolution graph, which includes the discovery of not only individual research topics but also theme evolution. We investigate them in details in following discussion.

4.1 Discovery of Research Topics

Zooming into individual topics identified by Citation-LDA, we are interested in finding *milestone papers*, generating *keywords*, and computing the *temporal strength* for each topic.

[6] In experiments, α and β are symmetric prior with weight 1×10^{-3} to encourage sparse topic distributions

[7] We use $\#(\cdot)$ as the *count* function that computes the number of instances satisfy the conditions specified in ( ), and $\neg(i^*, t)$ denotes all the citations except the t-th citation in document d_{i^*}

[8] In experiments, this is empirically measured by parallel gibbs sampling

[9] We use $\langle\cdot\rangle$ to denote averaging the statistics specified over the iterations in sampling

4.1.1 Topic Milestone Papers

The *topic-doc* distribution $\{\hat{\varphi}_{k,j}\}$, as computed in Equation. 3 indicates how well a single paper d_j represents the topic z_k. The ranking of papers based on $\{\hat{\varphi}_{k,j}\}$ in essence provides the topic-aware impact assessment for papers with the milestone papers for topic z_k ranked at the top.

There are advantages over naive ranking of papers based on the citation counts, which can be inaccurate since there are cases that in one area people tend to include more references than people from another area. Even sophisticated citation-based measurement, e.g., [10, 20, 21, 22], without taking into account of topics, can lead to bad judgement: a well recognized theoretic paper about graphic model in "Bayes learning" might receive less credit in "data engineering" and "very large database" due to the computational difficulty that limits its application.

4.1.2 Topic Temporal Strength

For topic z_k, there is a time point when it began attracting attention, a time point when it enjoyed its glory days with most important milestone papers emerged, and possibly a time point when interest decreased and the topic faded out. If it is a long lasting topic, it might span over decades while if not, the active period can be as short as only a few years.

Topic temporal distribution sufficiently maintains the information. Viewing topic z_k as a distribution over papers, the proportion of accumulated probability for published papers until time t forms the cumulative distribution function (CDF):

$$\begin{aligned} \Pr(time \le t | z = k) &= \sum_{j, time(d_j) \le t} \Pr(c = j | z = k) \\ &= \sum_{j, time(d_j) \le t} \hat{\varphi}_{k,j} \end{aligned} \tag{5}$$

For the discrete time case, which is also our case, the probability mass function (PMF) for temporal distribution of z_k is:

$$\Pr(time = t | z = k) = \sum_{j, time(d_j) = t} \hat{\varphi}_{k,j} \tag{6}$$

In addition, the expectation can be computed as:

$$\mathbf{E}_{c|z=k}[\mathrm{time}(c)] = \sum_j \mathrm{time}(d_j) \hat{\varphi}_{k,j} \tag{7}$$

The standard deviation can also be easily computed, which, together with *topic expected time*, concisely show the major occurring time and provide a rough estimation about the life span for a topic.

4.1.3 Topic Keywords

In general it would be desirable to summarize the topic with only a few words [6]. With Citation-LDA, we accomplish this by leveraging words in title (or abstract if available) as tags for each paper and summarize the topic by those words with high *expected occurrences*. Specifically, to compute the word occurrence expectation over $\{\hat{\varphi}_{k,j}\}$ for word w in topic z_k:

$$\mathbf{E}_{c|z=k}[\#(w, c)] = \sum_j \hat{\varphi}_{k,j} \cdot \#(w, d_j) \tag{8}$$

As shown later in experiments, the topic keywords generated from titles are surprisingly indicative yet discriminative for especially seemingly similar topics.

4.2 Discovery of Theme Evolution

In order to help a researcher see the big picture of all research topics, we can also easily use Citation-LDA to discover the theme evolution, which would involve the exploration of assessing the *topic importance* as well as the *topic dependency relation*, and recognizing the underlying *evolution patterns*.

4.2.1 Topic Importance

By Equation. 4, the distribution of $\{\hat{\Pr}(z = k)\}$ represents the chance of documents from one topic getting cited. Consequently, it can be associated as the topic importance in the research community since topics with higher importance are those who receive more citations and vice versa. The top important topics reflect the major research progress and reveal the dominant research interest in one area.

4.2.2 Topic Dependency

In Citation-LDA, topics are represented as multinomial distributions over papers $\{\hat{\varphi}_{k,j}\}$ while the *doc-topic* distribution $\{\hat{\theta}_{i,k}\}$ implies the topic mixture of document d_i. More precisely, $\hat{\theta}_{i,k^{(2)}}$ is the probability of topic $k^{(2)}$ occurring in document d_i with an (outlink) citation. Consequently, when marginalizing over papers d_j discounted by $\{\hat{\varphi}_{k^{(1)},j}\}$, the probability of citing topic $k^{(2)}$ (by topic $k^{(1)}$) conditioned on topic $k^{(1)}$ is:

$$\begin{aligned} & \Pr(k^{(1)} \to k^{(2)} | k^{(1)}) \\ = \; & \mathbf{E}_{c|z=k^{(1)}}[\Pr(z = k^{(2)} | d = c)] \\ = \; & \sum_j \Pr(c = j | z = k^{(1)}) \Pr(z = k^{(2)} | d = j) \\ = \; & \sum_j \hat{\varphi}_{k^{(1)},j} \hat{\theta}_{j,k^{(2)}} \end{aligned} \tag{9}$$

An intuitive explanation of Equation. 9 is: whenever randomly drawing a document d_j from topic $k^{(1)}$, and then emitting a citation from that document, $\Pr(k^{(1)} \to k^{(2)} | k^{(1)})$ is the chance of that citation being associated with *latent* topic $k^{(2)}$.

More importantly, Equation. 9 explains the *topic level citation structure*, as well as quantifies the *topic dependency* between any two topics precisely — the amount of influence of topic $k^{(2)}$ upon topic $k^{(1)}$, from which we can tell if a topic is developed on top of another.

4.2.3 Evolution Patterns

Topic level citation structure $\{\Pr(k^{(1)} \to k^{(2)} | k^{(1)})\}_{K \times K}$ reveals the topic dependency. Nevertheless, it is indeed a $K \times K$ matrix with most entries being sparse. In our work, we propose two pruning criteria:

- *Threshold cutting-off*: By setting a threshold ξ [10] empirically, all citation dependencies between topics with strength less than ξ would be removed.
- *Temporal regularization*: As previously investigated in [15, 16], the citation dependencies of the "old" topics upon the "new" topics can be roughly regarded as noise and safely discarded.

After applying pruning to the *topic level citation structure*, significant yet meaningful influences between topics are kept. Closely dependent topics form the themes, in which different *evolution patterns* can be found: some topics may get merged into a new topic which is highly dependent on them (*merging*). Alternatively, one topic might have multiple subsequent topics that are developed on top of it (*branching*). In other cases, topics stop evolution and gradually *fade out*. We will discuss evolution patterns with concrete examples in the following experiment section.

5. EXPERIMENTS & RESULTS

In this section, we first formally describe the two datasets AAN and PMC on which we demonstrate our Citation-LDA. Further, extensive evaluation results of discovery of research topics and theme evolutions are discussed. Last, we show that our Citation-LDA over-performs conventional Content-LDA baseline with two evaluation metrics: *forward-citation* and *journal conditioned entropy*.

Due to space limit, here we only show some representative results in our paper. The complete results as well as the source code for Citation-LDA can be found at:
`http://sifaka.cs.uiuc.edu/~xwang95/citation_lda/`

5.1 Dataset

In our experiments, two public scientific literature datasets are investigated: AAN from natural language processing domain and PMC from biomedical and life sciences.

5.1.1 ACL Anthology Network (AAN)

The ACL Anthology Network (AAN) [20] is a public dataset which includes all papers published by Association for Computational Linguistics (ACL) and related organizations over the period from 1965 till now. Major conference and journal papers in the area of natural language processing (NLP) can be found in the dataset. In our experiments, there are are in total $18,041$ papers (including citing and cited papers) from 13 venues with $82,944$ citations.

5.1.2 PubMed Central (PMC)

The PubMed Central (PMC) [11] is a free archive of biomedical and life sciences journal literature. Compared with AAN, it is a much larger yet sparser dataset, with a coverage of much wider areas than NLP. In our experiments, we includes the papers published after year 1960 and there are $145,317$ article papers with $274,133$ citations from $1,726$ journals.

Unlike AAN, the large number of journals in PMC provide a *"coarse topical annotation"* for papers, as in life sciences journals are commonly specialized in only a few research topics. For example, the journal *"Nucleic Acids Research"* covers research on nucleic acids such as DNA and RNA, but the journal *"Environmental Health Perspectives"* mainly publishes research on environmental health such as toxicology, exposure science and public health, etc. Later, we would utilize the journal information to evaluate the modeling performance of Citation-LDA and Content-LDA.

5.2 Results of Research Topics Discovery

Before the discussion of the results, however, a nontrivial question is how to determine the *number of topics* to be modeled? In following experiments, we perform the Citation-LDA with 100 topics in AAN and 500 topics in PMC, leaving the discussion of selecting the topic number in Section. 5.4.

5.2.1 Finding Milestone Papers

Milestone papers for two topics: "sentiment analysis" from AAN and "air pollution" from PMC, both of which are of great importance, are presented in Table. 1 - 2 respectively (10 milestone papers for each topic). Together, the *topic-doc* probability $\hat{\varphi}_{k,j}$ and the venue/journal sources are included. Clearly, the milestone papers listed are truly representative and recognized by the community based on the impact with respect to the topic.

One might notice that the top milestone papers in Table. 2, unlike those of topic "sentiment analysis" from AAN, are actually all from one journal "Environmental Health Perspectives", which is generally regarded as among the most top tier journals in the area of "environment health" with especially established reputation in the topic "air pollution". In fact, the top milestone papers for topics in PMC being from the same (or only a few) journal(s) are actually quite common. Given that the journals in PMC are closely related to a variety of specialized topics, it can be taken as "noisy" topic labels of fair quality for evaluation purpose.

[10] $\xi = 0.1$ in experiments

[11] `http://www.ncbi.nlm.nih.gov/pmc/`

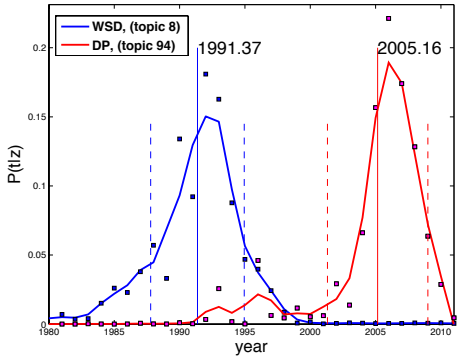
Table 1: Top 10 High Impact Papers in Topic "Sentiment Analysis" (Topic 89, AAN)

$\hat{\varphi}$	Venue	Paper Title
0.078533	EMNLP'02	Thumbs Up? **Sentiment Classification** Using Machine Learning Techniques
0.067202	ACL'02	Thumbs Up Or Thumbs Down? **Semantic Orientation** Applied To Unsupervised Classification Of Reviews
0.048269	HLT'05	Recognizing Contextual Polarity In Phrase-Level **Sentiment Analysis**
0.043634	ACL'04	A Sentimental Education: **Sentiment Analysis** Using Subjectivity Summarization Based On Minimum Cuts
0.036498	ACL'97	Predicting The **Semantic Orientation** Of Adjectives
0.031173	COLING'04	Determining The **Sentiment Of Opinions**
0.030686	HLT'05	Extracting Product Features And **Opinions** From Reviews
0.028673	EMNLP'03	Towards Answering **Opinion** Questions: Separating Facts From **Opinions** And Identifying The **Polarity Of Opinion Sentences**
0.027851	EMNLP'03	Learning Extraction Patterns For **Subjective Expressions**
0.016856	ACL'05	Seeing Stars: Exploiting Class Relationships For **Sentiment Categorization**ñ With Respect To Rating Scales

Table 2: Top 10 High Impact Papers in Topic "Air Pollution" (Topic 175, PMC)

$\hat{\varphi}$	Venue	Paper Title
0.035435	Environ_Health_Perspect	Ultrafine Particulate **Pollutants** Induce Oxidative Stress and Mitochondrial Damage
0.018051	Environ_Health_Perspect	Ambient **Air Pollution** and Atherosclerosis in Los Angeles
0.017836	Environ_Health_Perspect	Effects of **Air Pollution** on Heart Rate Variability: the VA Normative Aging Study
0.014414	Environ_Health_Perspect	Acute Blood Pressure Responses in Healthy Adults during Controlled **Air Pollution** Exposures
0.014233	Environ_Health_Perspect	The Effect of Particulate **Air Pollution** on Emergency Admissions for Myocardial Infarction
0.013984	Environ_Health_Perspect	Diabetes, Obesity, and Hypertension May Enhance Associations Between **Air Pollution** and Markers of Systemic Inflammation
0.013690	Environ_Health_Perspect	Nanotoxicology: an Emerging Discipline Evolving from Studies of **Ultrafine Particles**
0.013266	Environ_Health_Perspect	Association of Fine **Particulate** Matter From Different Sources With Daily Mortality in Six U.S. Cities
0.013090	Environ_Health_Perspect	**Ultrafine Particles** Cross Cellular Membranes by Nonphagocytic Mechanisms in Lungs and in Cultured Cells
0.012830	Environ_Health_Perspect	**Ambient Particulate Air Pollution**, Heart Rate Variability, and Blood Markers of Inflammation in a Panel of Elderly Subjects

Figure 2: Topic Temporal Strength for "WSD" and "DP"



5.2.2 *Discovering Temporal Strength*

To demonstrate that our model discovers the topic over time correctly, we show the topic temporal strength of two topics, namely "word sense disambiguation" (WSD) and "dependency parsing" (DP) from AAN in Figure. 2, and the computational details can be found in Equation. 5-7.

In fact, the topic "WSD" was once a popular topic around early 90s while "DP" was newly popularized around year 2005. Based on our model, "WSD" has the expected time 1991.37, with a standard deviation 3.58. For "DP", the expectation is 2005.16 and standard deviation is 3.84. These estimations are all consistent with the expert knowledge.

5.2.3 *Extracting Topic Keywords*

We list the extracted keywords (phrases) [12] in Table. 3-4. As will be explained in details later, the topics are the dominant 10 topics in AAN and PMC datasets. The extracted keywords are mainly about the *problem, task, model* and *methodology* of the topics. For Topic 73 in AAN, it shows that the topic investigates the problem of "part-of-speech tagging", models the problem as "sequential labeling", and approaches it with "discriminative parsing" methods. For Topic 61 in PMC, the nature of the topic can be recovered as research on the risks of "children exposure" against "agricultural spraying" such as "pesticides" and "organophosphorus". In general, it is easy to conclude the research problems or detailed methodology for each topic through the extracted keywords along. Besides, based on the spotted keywords, Topic 92, Topic 96, Topic 80, and Topic 50 in AAN are all about the research theme "statistical machine translation". But keywords reveal that topics differ from each other as concerning about *distinct* methods/models (phrase-based models (92) *v.s.* discriminative learning (96)) or problems (reordering, alignment (80) *v.s.* evaluation (50)), which evidently substantiates that the keywords are adequately discriminative even for quite related topics, serving as accurate yet succinct summary for topics.

5.3 Results of Theme Evolution Discovery

5.3.1 *Identifying Important Topics*

As earlier implied, Table. 3-4 show the dominant 10 topics for AAN and PMC, which are selected based on the topic weight $\{\hat{\Pr}(z = k)\}$ as computed in Equation. 4. Identified dominant topics cover major research progress and interest in NLP and life sciences. In AAN, it is obvious that the research theme "statistical machine translation" plays the most important role in the community, thriving and diverse with multiple different topics such as Topic 92, 96, 80, and 50. In PMC, many topics related to "public health" are dominant such as Topic 175, 61, and 86, though the detailed research topics are distinguishable from the keywords.

Taking the topic temporal strength into account,

$$\Pr(z = k, time = t) = \Pr(time = k|z = k) \cdot \hat{\Pr}(z = k)$$

is the joint probability of topic strength and time, allowing us to compare the topic strength in different time periods *with each other topics*. We visualize this for AAN and PMC in Figure. 3-4, and it shows that the major research development occurred after year 2000 for both two dataset [13], except that Topic 8 ("word sense disambiguation") of AAN was dominant compared with others in early 90s while Topic 2 of "yeast", "saccharomyces cerevisiae" in PMC was a extensively studied around entire 90s.

5.3.2 *Topic Dependency & Evolution Patterns*

After applying the pruning to the *topic level citation structure* the evolution graph for research themes can be plotted. We show the

[12] Top word phrases are generated from top 20 keywords and then matched with n-grams in titles of the milestone papers

[13] However, there is possibility that our datasets are biased as being rich in citations after year 2000

Table 3: Dominant 10 Topics in AAN (100 topics)

Topic	Weight	$\mathbf{E}(t)$	$\mathrm{stdev}(t)$	Top Keyword Phrases
94	0.02806	2005.16	3.84	dependency parsing, non-projective, shared tasks, multilingual
89	0.02761	2004.64	3.25	sentiment classification, opinion analysis, orientation, learning
8	0.02509	1991.37	3.58	word sense disambiguation, lexical semantics
92	0.02428	2004.98	3.26	machine translation, phrase-based models, alignment
96	0.02277	2005.45	3.59	machine translation, online, margin, discriminative learning
84	0.02093	2003.94	3.36	semantic role labeling, shared tasks
80	0.02069	2003.44	3.83	machine translation, reordering, alignment
73	0.01965	2002.76	4.09	discriminative parsing, sequential labeling, part-of-speech
50	0.01908	2000.87	4.13	machine translation, minimum error rate training, BLEU evaluation
72	0.01804	2002.74	4.45	coreference resolution, machine learning, anaphora, pronoun

Table 4: Dominant 10 Topics in PMC (500 topics)

Topic	Weight	$\mathbf{E}(t)$	$\mathrm{stdev}(t)$	Top Keyword Phrases
484	0.00624	2006.45	8.95	protein, molecular interaction, biomolecular, database
499	0.00504	2007.36	9.89	ensemble, gene, genome, resources
488	0.00478	2006.48	19.37	gnome-scale metabolic reconstruction, escherichia coli, malaria
175	0.00450	2004.48	10.67	air pollution, ambient particulates, heart rates, exposure
373	0.00388	2005.35	11.77	non-coding RNA, sequence alignment, structure prediction, genome
492	0.00382	2006.56	11.39	sorcerer II, global ocean sampling, metagenomics, atlantic
61	0.00351	2003.22	12.12	children exposure, agricultural spraying, pesticides, organophosphorus
2	0.00350	1998.00	13.85	yeast, actin,saccharomyces cerevisiae, protein, myosin, cell
38	0.00338	2002.67	12.78	cell, regulatory T cell, CD4, CD25, human, Foxp3, expression, induction
86	0.00320	2003.64	14.12	phthalate exposure, human, urine, infants, metabolites, prenatal, health

Figure 3: Topic-Temporal Joint Strength In AAN

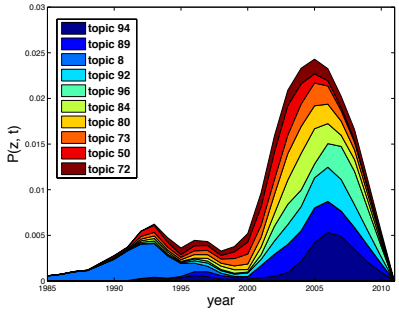


Figure 4: Topic-Temporal Joint Strength In PMC

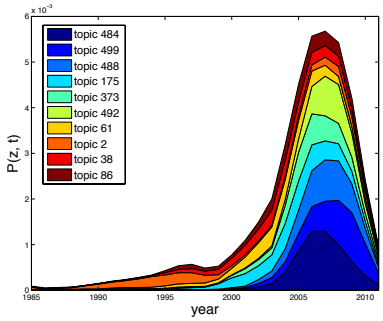
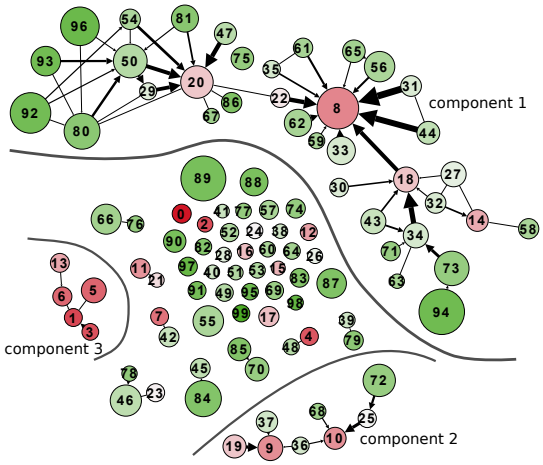


Figure 5: Theme Evolution Graph of AAN



evolution graph of AAN with 100 topics in Figure. 5: each node represents a topic and the importance of topics are discriminated by the size of nodes. The green nodes are new topics while the red ones are *relatively* old. In addition, the dependency between topics are reflected by the thickness of edges .

There are three major connected component, each of which contains themes developing over time: Component 3 is about the theme "grammar", and corresponding topics entirely *faded out* during early 90s. Nevertheless, Component 2 has the theme of "discourse/dialogue" and "summarization", showing mildly progress recently (e.g., Topic 72 (2003) of *"machine learning" based "coreference resolution"*). Observing the Component 1, which is the largest, is interesting with discovery of various theme evolution patterns: Topic 8 (1991) about "word sense disambiguation" was *branched* into many topics, with one of them (Topic 18) being about "prepositional phrase attachment" (1994). Soon, Topic 18 further enabled Topic 34 (1999) of "statistical parsing", and again Topic 73 of "discriminative parsing" was established by 2003 on top of Topic 34. Later, Topic 94 of "dependency parsing" raised and has grown as one dominant topic since 2005.

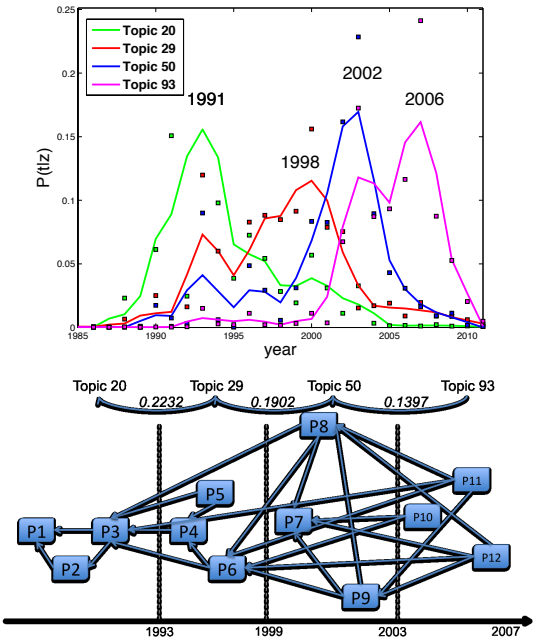
Another key thread of theme in Component 1 was initiated by Topic 20, which was the very beginning topic of the theme "statistical machine translation" (SMT). The topics along the theme evolution path are presented in Table. 5, including 4 topics (Topic 20, 29, 50, and 93), together with the milestone papers (top 3 for each). In addition, the temporal distribution over time is given in Figure. 6, where the citations among the milestone papers, and the dependency strength between consecutive topics are also depicted.

Specifically, Topic 20 began increasing its impact around early 90s, introducing basic statistical methods to machine translation; Later, around 1998, its popularity was shifted to Topic 29 which was specialized in subproblems such as "decoding", "alignment" and "reordering" in SMT; By 2002, however, Topic 50 emerged, and soon grew as the new dominant topic by proposing "BLEU" as the standard evaluation metric and investigating "discriminative methods" such as "minimum error rate training"; The current state of the art approach in SMT, "phrase-based model", accompanied

Table 5: SMT Example for Theme Evolution

Topic	Year	Paper ID	Paper Title	$\hat{\varphi}$
Topic 20	1990	P1	A Statistical Approach To Machine Translation	0.036542
	1991	P2	A Program For Aligning Sentences In Bilingual Corpora	0.047619
	1993	P3	The Mathematics Of Statistical Machine Translation: Parameter Estimation	0.060931
Topic 29	1996	P4	HMM-Based Word Alignment In Statistical Translation	0.097162
	1997	P5	Decoding Algorithm In Statistical Machine Translation	0.030390
	1999	P6	Improved alignment models for statistical machine translation	0.036367
Topic 50	2002	P7	BLEU: A Method For Automatic Evaluation Of Machine Translation	0.087902
	2002	P8	Discriminative Training & Maximum Entropy Models For Statistical Machine Translation	0.027799
	2003	P9	Minimum Error Rate Training In Statistical Machine Translation	0.027027
Topic 93	2003	P10	Statistical Phrase-Based Translation	0.036239
	2005	P11	A Hierarchical Phrase-Based Model For Statistical Machine Translation	0.022442
	2007	P12	Hierarchical Phrase-Based Translation	0.043163

Figure 6: Temporal Evolution in Topics of Theme SMT



by the raise of Topic 93, was actually built on top of previous work, especially milestone papers of $P7$-$P9$ of Topic 50.

In Figure. 6, citation links among milestone papers across topics are illustrated, which clearly show the formation of topics through the "stable core set" of milestone papers that *get cited together* (co-cited). More importantly, it is evident that the "co-citation" of "core" papers is the direct contributing factor in the dependency relation between two consecutive topics.

5.4 Model Selection & Comparison Results

We now discuss how to select the topic numbers for Citation-LDA and compare the performance with Content-LDA on two metrics, namely, *Forward Citation* and *Jounral Conditional Entropy*.

We investigate the conventional Content-LDA [4] as our baseline, using the title and abstract to represent the papers in both datasets. In order to make the output of Content-LDA aligned with that of Citation-LDA, we need to derive the missing *topic-doc* distribution: the distribution over papers (instead of tokens) for each topic. As in our experiments, we assume $\Pr(d|k) \propto \Pr(k|d) \cdot \Pr(d)$ whereas $\Pr(d) \propto |d|$ with $|d|$ being the document length for d.

5.4.1 Evluation on Forward Citation for AAN

We compute the *topic forward citation* probability based on the topic dependency (Equation. 9) and expected topic time (Equation. 7). In words, the forward citation probability reflects the chance a topic *cites* future topics that arise after itself (though it is impossible for a paper to cite a future paper). We compute the model's loss on topic k by the topic *future citation probability* , which is given by: $l(k) = \sum_{\tilde{k}, t(\tilde{k})>t(k)} \Pr(k \to \tilde{k}|k)$ for topic k. To assess the total *loss for Forward Citation* of a model, we define it as follows:

$$\text{Loss}_{FC} = \sum_{k} \Pr(k) \cdot l(k)$$

Table 6: Loss on Forward Citation (AAN)

#topic	20	100	200
Citation-LDA	0.3148	**0.1917**	0.2488
Content-LDA	0.3745	0.3816	0.3924

We show the evaluation based on *Forward Citation* for AAN in Table. 6, from which we see: 1) Citation-LDA has better performance on Forward Citation compared with Content-LDA and 2) 100 topics are a good choice for AAN dataset.

5.4.2 Evaluation on Journal Conditional Entropy for PMC

As discussed before, the journal sources are fairly good "coarse" annotation for topics in PMC. For topic k, we can derive the *journal conditional distribution on topic* k, yielding the conditional entropy [14]:

$$H(J|z) = \sum_{z=k} \Pr(z=k) \cdot H(J|z=k)$$

The $H(J|z)$ would have low value if the journal labels and topic labels are *consistent*, by which we mean that for papers with the *same topic label* (in a probabilistic sense), there is *one journal label* being as dominant as possible, ideally being purely the only journal label. Hence, we can compute the *loss for Journal Conditional Entropy* of a model as:

$$\text{Loss}_{CE} = H(J|z)$$

Table 7: Loss on Journal Conditional Entropy (PMC)

#topic	100	300	500	1000
Citation-LDA	3.5047	3.2144	**3.18729**	3.4118
Content-LDA	4.2048	4.2805	4.06496	4.4725

Based on the journal conditional entropy on topics (Table. 7), we again demonstrate the advantage of Citation-LDA over Content-LDA: the topic formed in Citation-LDA is more consistent with the "journal labels" than Content-LDA. In addition, we verify that for PMC dataset, 500 topics might be a reasonable choice.

6. CONCLUSION AND FUTURE WORK

In this paper, we proposed a novel approach for analyzing research theme evolution of scientific literature data where citation

[14] Entropy $H(X) = -\sum_{x} \Pr(x) \log \Pr(x)$

links are available. Our tasks have two folds: 1) to discover research topics, which includes finding milestone papers, computing topic temporal strength, and extracting keywords for topics; 2) to discover theme evolution, which includes identifying topic importance, learning topic dependency relation, and recognizing the evolution patterns. These computational components together enable us to understand evolution of research themes by constructing the evolution graph. In experiments, we investigated two datasets, namely AAN and PMC from two domains, with extensive results showing that our proposed model, Citation-LDA, which represents article paper as "bag of citations" and model the generation of citation links within a probabilistic framework, can effectively accomplish the tasks defined above, with the performance better than Content-LDA. Our proposed Citation-LDA, together with the developed mining techniques, can be very useful to help researchers digest literature quickly, thus speeding up scientific research discovery and delivering very broad positive impact on the society.

In general, our model can also be applied to any graph data for tasks such as network clustering and ranking, as well as modeling the evolution of network generation, which we leave as future work directions.

7. ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their constructive comments and suggestions, which significantly contribute to improving the manuscript and help us to notice the related works [12, 25].

Our work is supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20155. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

8. REFERENCES

[1] E. M. Airoldi, D. M. Blei, S. E. Fienberg, E. P. Xing, and T. Jaakkola. Mixed membership stochastic block models for relational data with application to protein-protein interactions. In *Proceedings of the international biometrics society annual meeting*, 2006.

[2] D. Blei and J. Lafferty. Dynamic topic models. In *Proceedings of the 23rd international conference on Machine learning*, pages 113–120. ACM, 2006.

[3] D. M. Blei and J. D. Lafferty. A correlated topic model of science. *The Annals of Applied Statistics*, pages 17–35, 2007.

[4] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, Mar. 2003.

[5] L. Bolelli, S. Ertekin, and C. Giles. Clustering scientific literature using sparse citation graph analysis. *Knowledge Discovery in Databases: PKDD 2006*, pages 30–41, 2006.

[6] J. Boyd-Graber, J. Chang, S. Gerrish, C. Wang, and D. Blei. Reading tea leaves: How humans interpret topic models. In *Proceedings of the 23rd Annual Conference on Neural Information Processing Systems*, 2009.

[7] J. Chang and D. Blei. Relational topic models for document networks. In *Artificial Intelligence and Statistics*, pages 81–88, 2009.

[8] G. W. Flake, R. E. Tarjan, and K. Tsioutsiouliklis. Graph clustering and minimum cut trees. *Internet Mathematics*, 1(4):385–408, 2004.

[9] E. Garfield. The history and meaning of the journal impact factor. *JAMA: the journal of the American Medical Association*, 295(1):90–93, 2006.

[10] R. Ghosh, T.-T. Kuo, C.-N. Hsu, S.-D. Lin, and K. Lerman. Time-aware ranking in dynamic citation networks. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 373–380. IEEE, 2011.

[11] T. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

[12] K. Henderson and T. Eliassi-Rad. Applying latent dirichlet allocation to group discovery in large graphs. In *Proceedings of the 2009 ACM symposium on Applied Computing*, pages 1456–1461. ACM, 2009.

[13] J. E. Hirsch. An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the United states of America*, 102(46):16569, 2005.

[14] T. Hofmann. Unsupervised learning by probabilistic latent semantic analysis. *Machine Learning*, 42(1):177–196, 2001.

[15] Y. Jo, J. E. Hopcroft, and C. Lagoze. The web of topics: discovering the topology of topic evolution in a corpus. In *Proceedings of the 20th international conference on World wide web*, WWW '11, pages 257–266, New York, NY, USA, 2011. ACM.

[16] Q. Mei and C. Zhai. Discovering evolutionary theme patterns from text: an exploration of temporal text mining. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, KDD '05, pages 198–207, New York, NY, USA, 2005. ACM.

[17] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen. Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '08, pages 542–550, New York, NY, USA, 2008. ACM.

[18] A. Popescul, G. W. Flake, S. Lawrence, L. H. Ungar, and C. L. Giles. Clustering and identifying temporal trends in document databases. In *Advances in Digital Libraries, 2000. ADL 2000. Proceedings. IEEE*, pages 173–182. IEEE, 2000.

[19] V. Qazvinian and D. Radev. Scientific paper summarization using citation summary networks. In *Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1*, pages 689–696. Association for Computational Linguistics, 2008.

[20] D. Radev, P. Muthukrishnan, and V. Qazvinian. The acl anthology network corpus. In *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*, pages 54–61. Association for Computational Linguistics, 2009.

[21] H. Sayyadi and L. Getoor. Futurerank: Ranking scientific articles by predicting their future pagerank. In *Proc. of the 9th SIAM International Conference on Data Mining*, pages 533–544, 2009.

[22] D. Walker, H. Xie, K.-K. Yan, and S. Maslov. Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, 2007(06):P06010, 2007.

[23] C. Wang, D. Blei, and D. Heckerman. Continuous time dynamic topic models. *arXiv preprint arXiv:1206.3298*, 2012.

[24] X. Wang and A. McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 424–433. ACM, 2006.

[25] H. Zhang, B. Qiu, C. L. Giles, H. C. Foley, and J. Yen. An lda-based community structure discovery approach for large-scale social networks. In *Intelligence and Security Informatics, 2007 IEEE*, pages 200–207. IEEE, 2007.