Modeling Topic-level Academic Influence in Scientific Literatures

Jiaming Shen[1], Zhenyu Song[1], Shitao Li[2], Zhaowei Tan[1]
Yuning Mao[1], Luoyi Fu[3],Li Song[3],Xinbing Wang[3]
[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University, China
[2]Department of Mathematics, Shanghai Jiao Tong University, China
[3]Department of Electronic Engineering, Shanghai Jiao Tong University, China
{sjm940622, sunnyszy, list12356, dilevski_tan, morningmoni, yiluofu, song_li, xwang8}@sjtu.edu.cn

Abstract

Scientific articles are not born equal. Some generate an entire discipline while others make relatively fewer contributions. When reviewing scientific literatures, it would be useful to identify those important articles and understand how they influence others. In this paper, we introduce *J-Index*, a quantitative metric modeling topic-level academic influence. *J-Index* is calculated based on the novelty of each article as well as its contributions to the articles where it is cited. We devise a generative model named Reference Topic Model (RefTM) which jointly utilizes the textual content and citation information in scientific literatures. We show how to learn RefTM to discover both the novelty of each paper and the strength of each citation. Experiments on a collection of more than 420,000 research papers demonstrate that RefTM outperforms the state-of-the-art approaches in terms of topic coherence as well as prediction performance, and validate *J-Index*'s effectiveness of capturing topic-level academic influence in scientific literatures.

1 Introduction

Nowadays, there are numerous scientific articles with different qualities, which makes it unrealistic for the researcher to read all of them. Therefore, we need to select the most important papers when reviewing scientific literatures. Furthermore, understanding the scientific impact of each paper lays a foundation for intelligent academic search engine, facilitating tasks such as paper ranking and citation recommendation. If the researcher stands on the shoulders of giants, we want to find those giants (Foulds and Smyth 2013).

Modeling academic influence quantitatively is a challenging problem. In previous studies, one way of addressing this problem is using metrics related to citation counts such as impact factors. However, many citations are referenced out of "politeness, policy or piety" (Ziman 1968), and thus make literally little impact. Another attempt to solve this problem is adopting graph-based approaches. For example, (Radev et al. 2009) proposed an algorithm of PageRank (Page et al. 1999) to derive the measures of importance from the citation network. Nevertheless, this method did not utilize the textual content of articles and all citations were treated equally.

A variety of methods have been proposed in recent works for joint analysis of text and citation in scientific literatures, including classifying citations function (Teufel, Siddharthan, and Tidhar 2006), predicting citations topicality (Nallapati et al. 2008), and modeling document network (Chang and Blei 2009). However, those methods did not provide a direct measurement of academic influence of each paper. As a work more relevant to the one we present, (Dietz, Bickel, and Scheffer 2007) proposed the Citation Influence Model (CIM) to predict citation influence. CIM assumes that the citation graph is bipartite, with one set containing the citing papers while the other one containing the cited papers. To hold this assumption, CIM duplicates each paper that cites and is cited by other ones, losing the capacity of handling more complex citation networks. (Liu et al. 2010) adapted this model to heterogeneous networks for mining topical influence, but it still retained the assumption of being a bipartite graph. (He et al. 2009) relaxed this assumption by proposing Inheritance Topic Model (ITM). Nevertheless, the main goal of ITM is to detect topic evolution instead of measuring academic influence.

In this paper, we introduce *J-Index*, a quantitative metric of modeling paper's academic influence. For each paper, *J-Index* considers its citation number, the strength of each citation and the novelty of all papers where it is cited. *J-Index* resonates with the intuition that if one paper is cited by many innovative papers with high citation strength, this paper is more likely to be an influential paper itself. *J-Index* can help to rank papers, make academic recommendation, and therefore enable researchers to evaluate the scientific merit of an article.

To measure the novelty of each paper as well as the citation strength among them quantitatively, we devise an unsupervised generative model named Reference Topic Model (RefTM). Different from Latent Dirichlet Allocation (Blei, Ng, and Jordan 2003), RefTM allows each word to be drawn from either a paper's own topic or those from its references. The intuition behind is that a scholar may choose to write a word based on his/her own innovative ideas, or just "inherits" some thoughts from cited papers. RefTM posits that the paper with high novelty intends to generate a large proportion of words from its own ideas, and the citation with large strength is more likely to be selected for generating an inherited topic. We show how RefTM measures the innova-

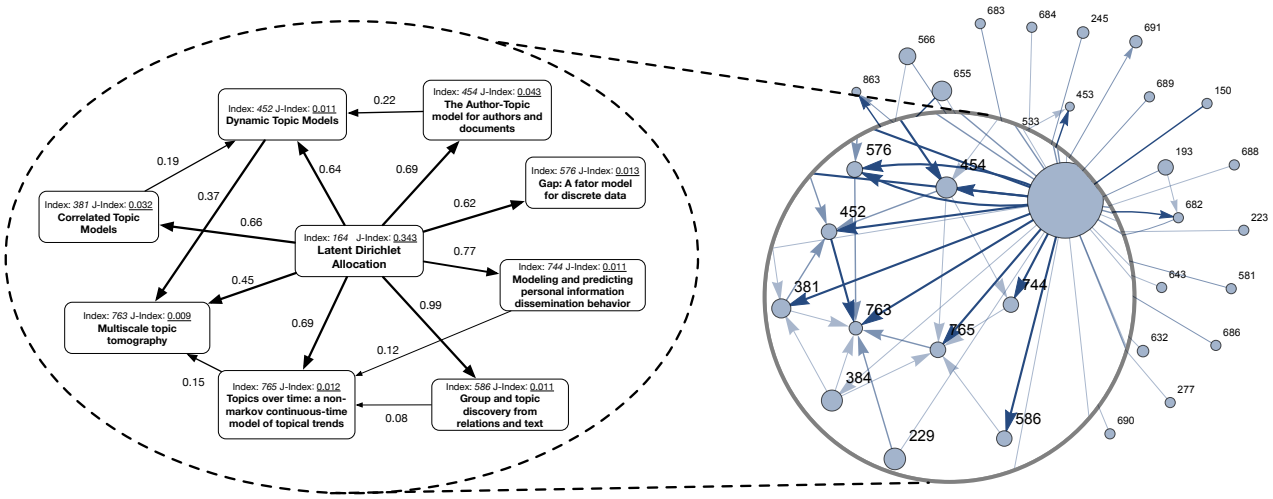


Figure 1: Right hand side is an illustrative citation graph in which the thickness of edge represents the citation strength and the vertex size indicates one paper's academic influence. Left hand side presents each paper's *J-Index* and quantitative measurement of the citation strength.

tiveness of article as well as the citation strength by jointly utilizing the textual content and citation network in scientific literatures.

We conduct extensive experiments on a collection of more than 420,000 research papers with over two million citations. Our results show that RefTM can effectively discover topics of high quality, model paper novelty and predict citation strength. We also calculate the *J-Index* of all research papers and the results validate its effectiveness of capturing topic-level influence in scientific literatures.

2 Academic Influence Metric

We model a collection of scientific literatures as a directed graph $G = (N, E)$ in which each node $e \in N$ represents an articles and each edge $(u, v) \in E$ indicates a citation from paper u to paper v. Our goal is to find a metric $F(\cdot)$ such that $F(e)$ represents the academic influence of paper e. An illustrative example is shown in Figure 1.

Naturally, we want the metric value correlated with the ground truth. However, this ground truth is unobservable, making it not obvious how one may quantify such a notion of "influence". Consequently, we need some commonsense knowledge when designing this influence metric. Here we have three general assumptions.

Assumption 1. *A paper's academic influence increases as it gains more citations.*

Citation is the most direct indicator of scientific merit, reflecting the academic influence of a paper. This assumption resonates with the intuition that a paper will increase its influence when there are more papers citing it. Put this mathematically, suppose we denote the set of paper m's citations as $C(m)$, then $F(\cdot)$ should be a monotonically increasing function in terms of $|C(m)|$, the citation number of paper m. Notice that $F(\cdot)$ is generally not a monotonic function over the whole corpus. A paper with 800 citations may be less influential than another paper with 650 citations due to many other factors like the function of each citation, which leads to our second assumption.

Assumption 2. *A paper with stronger citations intends to be more influential.*

Many citations are referenced out of "politeness, policy or piety" and have little impact on another work. We need to consider the strength of each citation when measuring an article's academic influence. Therefore, $F(\cdot)$ should include a component function $\delta(\cdot)$, defined on edge set E, to assess the citation strength. Moveover, $F(u)$ should increase more if one citation (u, v) has a larger value of $\delta(u, v)$. The conception of citation strength enables us to filter those citations made in passing by adding a relatively small influence score.

Assumption 3. *A paper cited by more innovative papers is more influential.*

In many cases, simply relying on the citation strength falls short of considering the difficulty of obtaining that citation. An innovative paper intends to generate most words from its own ideas, leading to small strengths of all citations associated with it. For this reason, $F(\cdot)$ should contain another node-weight function $\lambda(\cdot)$ to take into account the innovativeness of each paper.

J-Index

Based on three above-mentioned assumptions, we introduce *J-Index*, a quantitative metric modeling topic-level academic influence. *J-Index* is actually a metric framework, including two key components $\delta(\cdot)$ and $\lambda(\cdot)$, obtained from subsequent model. We define the *J-Index* of paper u as follows:

$$\textit{J-Index-Score}(u) = \sum_{c \in C(u)} \lambda(c) \times \delta(c, u) \qquad (1)$$

J-Index is calculated as a sum of all positive numbers, and thus the *J-Index* score of one paper will never decrease as

more citations are added, which enables *J-Index* to satisfy the first assumption. Besides, *J-Index* encodes the strength of each citation as well as the novelty of each citing paper in $\delta(c, u)$ and $\lambda(c)$. Consequently, a paper's *J-Index* will increase more if one of its citations is more influential, or one of its citing papers is more innovative. This correlates with the second and third assumptions. The exact choice of $\delta(c, u)$ and $\lambda(c)$ will be discussed in following section.

3 Reference Topic Model

We propose the Reference Topic Model (RefTM) to measure the novelty of each paper as well as the citation strength among them. Different from basic LDA model, RefTM is able to utilize both the textual content and citation information in scientific literatures.

The intuition of RefTM is that a scholar may choose to write a word based on his/her own innovative ideas, or just "inherits" some thoughts from references. In RefTM, each paper can generate the topic of a word from either its own topic distribution or from one of its references' topic mixtures. RefTM uses an unfair coin s to model this choice. We draw s from a Bernoulli distribution with parameter λ. Each paper d has its own parameter λ, which to some extent reflects its novelty. For each word in paper d, if its corresponding coin s equals 0, then we draw that word's topic by paper d's own topic mixture, otherwise we first choose one paper from d's references and draw the word's topic from that reference's topic mixture. The selection of that referred paper is modeled by RefTM as a multinomial distribution with parameter $\boldsymbol{\delta}$, which represents the strength of each citation. Both $\boldsymbol{\delta}$ and λ can be learned by RefTM.

Given a collection of M scientific articles and K topics expressed over V unique words, we can extract the citation network from this collection and thus the number of each paper's reference L_m is known. The whole generative process described by RefTM is as follows:

1. For each topic index $k \in \{1, \dots, K\}$
   (a) Draw a word distribution $\boldsymbol{\varphi_k} \sim \text{Dir}(\boldsymbol{\beta})$
2. For each document index $m \in \{1, \dots, M\}$
   (a) Draw a topic distribution $\boldsymbol{\theta_m} \sim \text{Dir}(\boldsymbol{\alpha})$
   (b) Draw a reference distribution $\boldsymbol{\delta_m} \sim \text{Dir}(\boldsymbol{\eta}|L_m)$
   (c) Draw an inheritance index $\lambda_m \sim \text{Beta}(\alpha_{\lambda_n}, \alpha_{\lambda_c})$
   (d) For each word $n \in \{1, \dots, N_m\}$ in document m:
      (i) Flip a coin $s_{m,n} \sim \text{Bern}(\lambda_m)$
      (ii) if $s_{m,n} = 0$:
         Draw a topic $z_{m,n} \sim \text{Multi}(\boldsymbol{\theta}_m)$
         Draw a word $w_{m,n} \sim \text{Multi}(\boldsymbol{\varphi}_{z_{m,n}})$
      (iii) else ($s_{m,n} = 1$):
         Draw a reference $c_{m,n} \sim \text{Multi}(\boldsymbol{\delta}_m)$
         Draw a topic $z_{m,n} \sim \text{Multi}(\boldsymbol{\theta}_{c_{m,n}})$
         Draw a word $w_{m,n} \sim \text{Multi}(\boldsymbol{\varphi}_{z_{m,n}})$

where $\boldsymbol{\theta}_m$ denotes a K-dimension multinomial distribution over topics, $\boldsymbol{\varphi_k}$ defines a V-dimension multinomial distribution over words, and $\boldsymbol{\delta_m}$ represents a L_m-dimension multinomial distribution over references. The graphical representation of RefTM is shown in Figure 2.

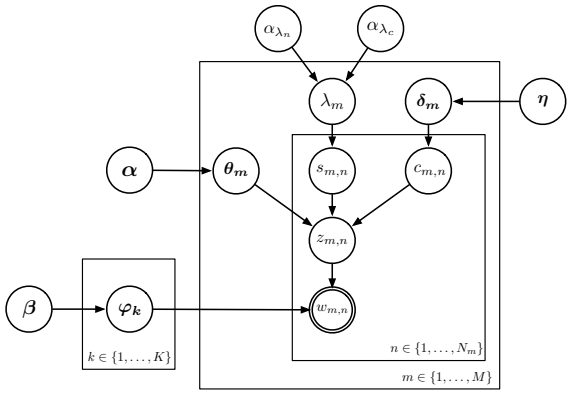


Figure 2: Graphical Representation of RefTM

4 Parameter Estimation

In RefTM, we need to estimate the parameter set $\{\boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\delta}, \boldsymbol{\lambda}\}$. We adopt collapsed Gibbs sampling (Griffiths and Steyvers 2004) for approximate estimation. Gibbs sampling allows to learn the parameters by alternatively updating each latent variable conditioned on the current assignments of all remaining variables. In order to derive these update equations, we first write out the joint distribution of all variables in generative process:

$$p(\boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c}, \boldsymbol{s}|\boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{\eta}, \alpha_{\lambda_n}, \alpha_{\lambda_c}) = \int p(\boldsymbol{w}|\boldsymbol{z}, \boldsymbol{\phi}) p(\boldsymbol{\phi}|\boldsymbol{\beta}) d\boldsymbol{\phi} \cdot \int p(\boldsymbol{z}|\boldsymbol{s}, \boldsymbol{c}, \boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\alpha}) d\boldsymbol{\theta} \cdot \int p(\boldsymbol{s}|\boldsymbol{\lambda}) p(\boldsymbol{\lambda}|\alpha_{\lambda_n}, \alpha_{\lambda_c}) d\boldsymbol{\lambda} \cdot \int p(\boldsymbol{c}|\boldsymbol{\delta}, \boldsymbol{L}) p(\boldsymbol{\delta}|\boldsymbol{\eta}, \boldsymbol{L}) d\boldsymbol{\delta}$$

Based on this joint distribution, we can derive the Gibbs sampling update equations for three different types of latent variables s_i, c_i and z_i associated with each word w_i, where subscript $i = (m, n)$ means it is the n-th word in m-th document.

During each iteration, we first sample a new flip coin s_i following two equations below:

$$p(s_i = 0|\boldsymbol{s}_{\neg \boldsymbol{i}}, \boldsymbol{w}, \boldsymbol{z}, \cdot) \propto \frac{(n_m^{z_i(0)} - 1) + n_m^{z_i(1)} + \alpha}{n_m^{(\cdot)(0)} + n_m^{(\cdot)(1)} + K\alpha - 1} \cdot \frac{N_m^{(0)} - 1 + \alpha_{\lambda_n}}{N_m^{(1)} + (N_m^{(0)} - 1) + \alpha_{\lambda_n} + \alpha_{\lambda_c}} \quad (2)$$

$$p(s_i = 1|\boldsymbol{s}_{\neg \boldsymbol{i}}, \boldsymbol{w}, \boldsymbol{z}, \boldsymbol{c_i}, \cdot) \propto \frac{n_{c_i}^{z_i(0)} + (n_{c_i}^{z_i(1)} - 1) + \alpha}{n_{c_i}^{(\cdot)(0)} + n_{c_i}^{(\cdot)(1)} + K\alpha - 1} \cdot \frac{N_m^{(1)} - 1 + \alpha_{\lambda_c}}{(N_m^{(1)} - 1) + N_m^{(0)} + \alpha_{\lambda_n} + \alpha_{\lambda_c}} \quad (3)$$

where $n_m^{z_i(0)}$ is the number of tokens in document m assigned to topic z_i through topic innovation and $n_m^{(\cdot)(0)} = \sum_{z_i} n_m^{z_i(0)}$. Furthermore, RefTM allows a paper to spread its influence through citations. This influence is partially reflected by $n_m^{z_i(1)}$, the number of topic z_i generated in the whole corpus by document m through topic inheritance and $n_m^{(\cdot)(1)} = \sum_{z_i} n_m^{z_i(1)}$.

Suppose the new coin s_i is equal to 1, we should next sample a new reference c_i. This update equation is shown below:

$$p(c_i|\boldsymbol{c}_{\neg \boldsymbol{i}}, \boldsymbol{w}, \boldsymbol{z}, s_i = 1, \cdot) \propto \frac{n_{c_i}^{z_i(0)} + (n_{c_i}^{z_i(1)} - 1) + \alpha}{n_{c_i}^{(\cdot)(0)} + n_{c_i}^{(\cdot)(1)} + K\alpha - 1} \cdot \frac{R_m^{c_i} - 1 + \eta}{R_m^{(\cdot)} + L_m\eta - 1} \quad (4)$$

Table 1: Notations

Variable	Meaning
M	number of documents
K	number of topics
V	vocabulary size
N_m	number of words in document m
L_m	number of references in document m
$\boldsymbol{\alpha}$	hyper-parameter of topic distribution
$\boldsymbol{\beta}$	hyper-parameter of word distribution
$\boldsymbol{\eta}$	hyper-parameter of reference distribution
$\alpha_{\lambda_n}, \alpha_{\lambda_c}$	hyper-parameters of inheritance index
$\boldsymbol{\theta}_m$	topic distribution of document m
$\boldsymbol{\varphi}_k$	word distribution of topic k
$\boldsymbol{\delta}_m$	reference distribution of document m
λ_m	inheritance index of document m

where $R_m^{c_i}$ represents the number of reference c_i selected by document m and $R_m^{(\cdot)}$ is the summation over all references.

Finally, we draw the latent topic z_i for each word w_i based on following two equations:

$$p(z_i|\boldsymbol{z_{\neg i}}, \boldsymbol{w}, s_i = 0, \cdot) \propto \frac{n_{z_i}^{w_i}+\beta-1}{n_{z_i}^{(\cdot)}+V\beta-1} \cdot \frac{(n_m^{z_i(0)}-1)+n_m^{z_i(1)}+\alpha}{n_m^{(\cdot)(0)}+n_m^{(\cdot)(1)}+K\alpha-1} \quad (5)$$

$$p(z_i|\boldsymbol{z_{\neg i}}, \boldsymbol{w}, s_i = 1, c_i, \cdot) \propto \frac{n_{z_i}^{w_i}+\beta-1}{n_{z_i}^{(\cdot)}+V\beta-1} \cdot \frac{n_{c_i}^{z_i(0)}+(n_{c_i}^{z_i(1)}-1)+\alpha}{n_{c_i}^{(\cdot)(0)}+n_{c_i}^{(\cdot)(1)}+K\alpha-1} \quad (6)$$

where $n_{z_i}^{w_i}$ is the number of tokens of word w_i assigned to topic z_i, and the same as $n_{z_i}^{(\cdot)}$ above represents the summation of $n_{z_i}^{w_i}$ over all topics.

The Gibbs sampling for RefTM is outlined in Algorithm 1 with notations defined in Table 1. After the sampling process converges (i.e., after a sufficient number of iterations), we can obtain the multinomial parameter sets corresponding to the state of Markov chain, using following equations:

$$\theta_{m,k} = \frac{n_m^{(k)}+\alpha}{n_m^{(\cdot)(0)}+n_m^{(\cdot)(1)}+K\alpha} \quad (7)$$

$$\varphi_{k,t} = \frac{n_k^{(t)}+\beta}{n_k^{(\cdot)}+V\beta} \quad (8)$$

$$\lambda_m = \frac{N_m^{(1)}+\alpha_{\lambda_c}}{N_m^{(0)}+\alpha_{\lambda_n}+N_m^{(1)}+\alpha_{\lambda_c}} \quad (9)$$

$$\delta_{m,c} = \frac{R_m^{(c)}+\eta}{R_m^{(\cdot)}+L_m\eta} \quad (10)$$

The equations resonate with our intuition statistically. Take λ_m for instance, if more words are generated from paper m's references instead of its own idea (i.e., $N_m^{(1)} \gg N_m^{(0)}$), then we may conclude that this paper is less innovative, reflected by a large inheritance index. Notice that following equation (7), the update of one term in paper m will have influence on all papers where it is cited. Therefore, RefTM is able to directly model influence propagation in the citation network.

Complexity Analysis

The algorithm above needs to sample three latent variables associated with each word. Each sampling operation requires a time of $O(K)$, where K is the number of topics given as the input of algorithm. Suppose there are W words in the corpus, and Gibbs sampling runs N_{iter} iterations, then the time complexity of RefTM equals $O(N_{iter}WK)$. With extra latent variables sampled, the time complexity of RefTM remains the same as that of LDA's.

As for the space complexity, the procedure itself uses four large data structures, the count caches $n_m^{(k)(0)}, n_m^{(k)(1)}$ of dimension $M \times K$, n_k^w of dimension $K \times V$, and R_m^c of dimension E, where E denotes the number of citations. In addition, the row sums of these four data structures, namely $n_m^{(0)}, n_m^{(1)}, n_k, R_m$, take space of dimension $3M + K$. In summary, the space complexity of RefTM is $O(MK + KV + E + W)$. Compared with the space complexity of LDA, RefTM requires another $O(E)$ space, in order to measure the influence of citation network.

Algorithm 1 Gibbs Sampling Algorithm for RefTM

Input: $K, \boldsymbol{w}, \alpha, \beta, \eta, \lambda_c, \lambda_n$
Output: Parameter sets $\{\boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\delta}, \boldsymbol{\lambda}\}$
Read in data and zero out all count caches
Randomly initialize $\boldsymbol{z_i}, \boldsymbol{c_i}, \boldsymbol{s_i}$
for $iter = 1$ to N_{iter} **do**
  for all documents $m \in [1, M]$ **do**
    for all words $n \in [1, N_m]$ in document m **do**
      if $s_{m,n}$ = 0 **then**
        Update the counts $n_m^{(k)(0)}, n_m^{(0)}$
      else
        Update the counts $n_c^{(k)(1)}, n_c^{(1)}, R_m^c, R_m$
      Draw a new $\tilde{s}$ from Eqs.(2-3)
      if $\tilde{s} = 0$ **then**
        Update the counts $n_k^{w_{m,n}}, n_k$
        Draw a new topic $\tilde{k}$ from Eq.(5)
        Update the counts $n_m^{(\tilde{k})(0)}, n_m^{(0)}, n_{\tilde{k}}^{w_{m,n}}, n_{\tilde{k}}$
      else
        Draw a new reference $\tilde{c}$ from Eq.(4)
        Update the counts $R_m^{\tilde{c}}, R_m, n_k^{w_{m,n}}, n_k$
        Draw a new topic $\tilde{k}$ from Eq.(6)
        Update the counts $n_{\tilde{c}}^{(\tilde{k})(1)}, n_{\tilde{c}}, n_{\tilde{k}}^{w_{m,n}}, n_{\tilde{k}}$
Read out parameters set $\boldsymbol{\theta}, \boldsymbol{\varphi}, \boldsymbol{\lambda}, \boldsymbol{\delta}$ by Eqs.(7-10)

5 Experiments

In this section, we analyze the performance of RefTM and validate the effectiveness of *J-Index* of modeling academic influence in scientific literatures both quantitatively and qualitatively.

Datasets

We consider two scientific corpora in the experiment part, i.e., a large unsupervised collection of 426728 articles with over 209 million citations gathered from the Internet and another small supervised collection of 799 papers obtained from (Liu et al. 2010). Papers in the first dataset are related to the "network" field, while those in the second corpus are

of more specific topics like "sentiment analysis" and "privacy security". We extract the title and abstract of each each paper as its textual contexts. After stop words removal, the average paper length of two corpora are 83 and 98 words, respectively. The distribution of paper's citation number is shown in Figure 3.

Evaluation Aspects

We conduct the experiments to analyze the performance of RefTM and effectiveness of *J-Index* from three aspects.

First, we evaluate the coherence of topics learned from RefTM since good topic cluster performance is the foundation of a good understanding of topic-level academic influence. Two metrics are used to assess the topic quality, including PMI-Score (Newman et al. 2010) and *topic coherence-Score* (Mimno et al. 2011).

Second, we adopt a prediction task to explore whether RefTM can effectively learn the citation strength, which is another key component of *J-Index*. We conduct experiments on the relatively small dataset in which the strength of each citation is manually labelled. We compare the results of RefTM with previous approaches and prove that our model has better performance concerning the prediction of citation strength.

Finally, we validate the effectiveness of *J-Index* in terms of capturing topic-level academic influence through one case study. We rank all 426728 papers in the first large dataset based on their *J-Index* scores and compare the results with each paper's corresponding assessment of research scientists.

The specific settings of hyper-parameters in RefTM and comparative methods are discussed in following subsections.

In all our experiments, we set $\alpha = 50/K$, $\beta = 0.01$, following the convention of (Griffiths and Steyvers 2004). As for three newly-added hyper-parameters in RefTM, we give the recommending values as $\eta = \bar{L}$, $\alpha_{\lambda_n} = 0.01 \cdot \bar{N}$, and $\alpha_{\lambda_c} = 0.04 \cdot \bar{N}$, where $\bar{L}$ is the average reference number of each paper, and $\bar{N}$ represents the average length of papers.

Topic Coherence Analysis

We evaluate the coherence of each learned topic based on two metrics. The first one is PMI-Score, which represents the average Pointwise Mutual Information between the most probable words in each topic. A larger PMI-Score indicates that the topic is more coherent. The calculation of PMI-Score requires external dataset such as Wikipedia Data. We construct our own reference collection based on 3.34 million scientific articles with 395.3 million word tokens to better reflect the language usage in academic domain.

We compare the PMI-Score of topics generated by LDA and RefTM at the left hand side of Figure 4. As we can see, the PMI-Score increases as the number of topics ranges from 10 to 50. Besides, we can discover that RefTM outperforms the LDA by 12% when topic number K equals 50, while the performance of these two models is fairly close with a small number of topics.

Topic Coherence-Score is another metric to assess the topic quality. *Topic Coherence-Score* depends only on internal training data, specifically the word co-occurrence statistics gathered from the corpus being modeled, and thus it does not rely on the external reference corpus like PMI-Score does. The comparative results of LDA and RefTM are shown at the right hand side of Figure 4, with number of topic K fixed as 30. We can see RefTM outperforms LDA in terms of median, lower quartile and upper quartile.

Citation Strength Prediction

We conduct this experiment using the small supervised datasets in which the strength of each citation is manually classified into three levels, i.e., strong, middle and weak, labeled as 1, 2 and 3, respectively. Similar to (Liu et al. 2010), we use the averaged AUC value for decision boundaries "1 vs. 2, 3" and "1, 2 vs. 3" as the quality measure for prediction performance. A larger AUC value indicates the prediction is more accurate. We compare the result with another two baseline methods – LDA-JS and LDA-post in (Dietz, Bickel, and Scheffer 2007). We set the hyper-parameters $\beta = 0.01$, $\alpha = 50/K$ where the number of topics K ranges from 10 to 50 in all three methods. After reducing the normalization constraint of RefTM in equation (9), we train each model 20 times and present the result in Figure 5. Clearly, we can see RefTM outperforms another two methods in all five scenarios.

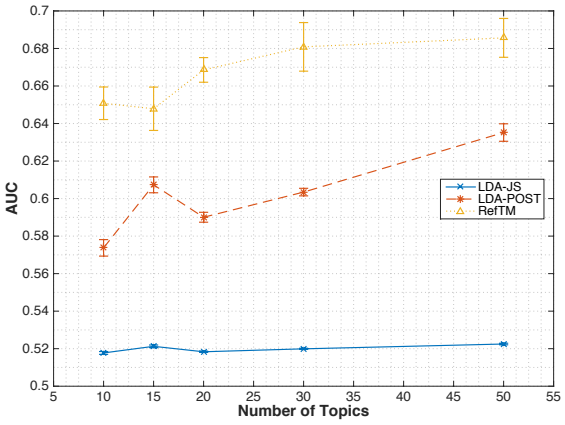


Figure 5: Citation Strength Prediction by AUC

Academic Influence Exploration

J-Index is a metric modeling one paper's influence rather its quality. These two notions differ in that a paper's influence may change over time while its quality is fixed. To reduce the bias from different publication dates, we select a subset of 224 papers published on INFOCOM in the same year 2003, and further adopt *J-Index* to measure a paper's own quality. We set the number of topic in RefTM to be 20, and rank each paper by its *J-Index* as well as citation numbers. Notice that here "citation number" actually means the number of citation within the corpora, which is only a fraction of a paper's overall citations. Due to space limitations, we only

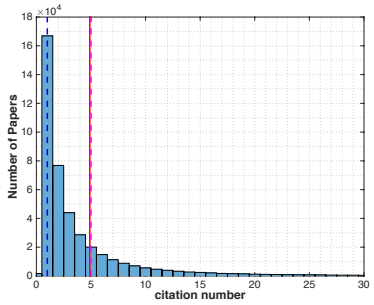
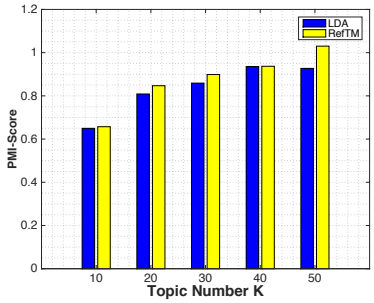


Figure 3: Histogram of Citation Number



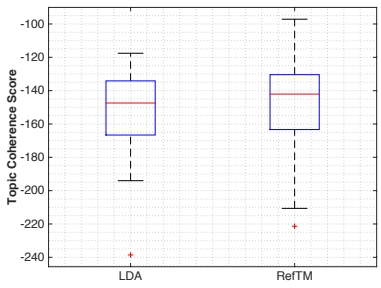


Figure 4: Topic Coherence Evaluation

Table 2: Top 5 Articles in INFOCOM 2003 randked by *J-Index* & citaions

Title	*J-Index*	citation counts
Top 5 Articles in INFOCOM 2003 ranked by *J-Index*		
Ad hoc positioning system (APS) using AOA	6.75	115
Performance anomaly of 802.11b	5.17	127
Packet leashes: a defense against wormhole attacks in wireless networks	4.13	74
Unreliable sensor grids: coverage, connectivity and diameter	4.00	82
Sensor deployment and target localization based on virtual forces	3.61	60
Top 5 Articles in INFOCOM 2003 ranked by citation number		
Performance anomaly of 802.11b	5.17	127
Ad hoc positioning system (APS) using AOA	6.75	115
Optimal routing, link scheduling and power control in multihop wireless networks	2.26	109
Sprite: a simple, cheat-proof, credit-based system for mobile ad-hoc networks	2.43	88
Unreliable sensor grids: coverage, connectivity and diameter	4.00	82

show the Top-5 papers in terms of two different metrics in Table 2.

Although the citation number and *J-Index* have positive correlation in general, they tend to rank some specific papers differently. For example, the most cited paper, "Performance anomaly of 802.11b", by Heusse, M. et al., is ranked second place according to *J-Index*. Another example is "Packet leashes: a defense against wormhole attacks in wireless networks", in which a novel mechanism is presented for defending against a severe attack in ad hoc networks called wormhole attack. *J-Index* ranks this paper at 3rd place, up from 11th place by citation count. Suppose we consider the citation number on Google Scholar, which is based on enormous data volume, as a partial ground truth, we find "Packet leashes" is actually ranked 2nd place among all papers in INFOCOM 2003 with over 1840 citations. After detailed observation, we discover that "Packet leashes" possesses a dominant position in the references of those papers where it is cited. This explains the behavior of *J-Index* and further validates its effectiveness in capturing paper's novelty.

6 Conclusions & Future Work

This paper introduces *J-Index*, a quantitative metric modeling topic-level academic influence. *J-Index* encodes each paper's novelty and its contribution to the articles where it is cited. A generative model named Reference Topic Model (RefTM) is further proposed to recover the innovativeness of each paper and the strength of each citation. RefTM is able to jointly utilize the textual content and citation relationship in scientific literatures during its training process, and thus plays a key role in the calculation of *J-Index*. Experiments on two real-world datasets demonstrate RefTM's ability to discover high-quality topics, predict citation strength and validate the effectiveness of *J-Index* for modeling topic-level academic influence.

There are several interesting future directions. For example, RefTM can be extended to model more inherent relationship in scientific literatures such as co-authorship, co-reference and co-citation, enabling *J-Index* to cover more information beyond word level. Another possible direction is to model the dynamics of citation network as well as *J-Index*. Currently, *J-Index* is only applicable to a static network and it has to be recalculated when new papers are added or time passes by. Therefore, an online version of RefTM as well as an explicit time component in *J-Index* is able to capture influence changes in scientific literatures. Finally, we intend to develop a system such as CiteSeer in which *J-Index* can facilitate a large pool of applications like paper ranking and academic recommendation.

References

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *the Journal of machine Learning research* 3:993–1022.

Chang, J., and Blei, D. M. 2009. Relational topic models for document networks. In *International Conference on Artificial Intelligence and Statistics*, 81–88.

Dietz, L.; Bickel, S.; and Scheffer, T. 2007. Unsupervised prediction of citation influences. In *Proceedings of the 24th international conference on Machine learning*, 233–240. ACM.

Foulds, J. R., and Smyth, P. 2013. Modeling scientific impact with topical influence regression. In *EMNLP*, 113–123.

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

He, Q.; Chen, B.; Pei, J.; Qiu, B.; Mitra, P.; and Giles, L. 2009. Detecting topic evolution in scientific literature: how can citations help? In *Proceedings of the 18th ACM conference on Information and knowledge management*, 957–966. ACM.

Liu, L.; Tang, J.; Han, J.; Jiang, M.; and Yang, S. 2010. Mining topic-level influence in heterogeneous networks. In *Proceedings of the 19th ACM international conference on Information and knowledge management*, 199–208. ACM.

Mimno, D.; Wallach, H. M.; Talley, E.; Leenders, M.; and McCallum, A. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 262–272. Association for Computational Linguistics.

Nallapati, R. M.; Ahmed, A.; Xing, E. P.; and Cohen, W. W. 2008. Joint latent topic models for text and citations. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 542–550. ACM.

Newman, D.; Lau, J. H.; Grieser, K.; and Baldwin, T. 2010. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 100–108. Association for Computational Linguistics.

Page, L.; Brin, S.; Motwani, R.; and Winograd, T. 1999. The pagerank citation ranking: bringing order to the web.

Radev, D. R.; Joseph, M. T.; Gibson, B.; and Muthukrishnan, P. 2009. A bibliometric and network analysis of the field of computational linguistics.

Teufel, S.; Siddharthan, A.; and Tidhar, D. 2006. Automatic classification of citation function. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 103–110. Association for Computational Linguistics.

Ziman, J. M. 1968. *Public knowledge: An essay concerning the social dimension of science*, volume 519. CUP Archive.