In this article, we use innovative full-text citation analysis along with supervised topic modeling and network-analysis algorithms to enhance classical bibliometric analysis and publication/author/venue ranking. By utilizing citation contexts extracted from a large number of full-text publications, each citation or publication is represented by a probability distribution over a set of predefined topics, where each topic is labeled by an author-contributed keyword. We then used publication/citation topic distribution to generate a citation graph with vertex prior and edge transitioning probability distributions. The publication importance score for each given topic is calculated by PageRank with edge and vertex prior distributions. To evaluate this work, we sampled 104 topics (labeled with keywords) in review papers. The cited publications of each review paper are assumed to be “important publications” for the target topic (keyword), and we use these cited publications to validate our topic-ranking result and to compare different publication-ranking lists. Evaluation results show that full-text citation and publication content prior topic distribution, along with the classical PageRank algorithm can significantly enhance bibliometric analysis and scientific publication ranking performance, comparing with term frequency–inverted document frequency (tf–idf), language model, BM25, PageRank, and PageRank + language model ($p < .001$), for academic information retrieval (IR) systems.

### Introduction and Motivation

Bibliometrics is a set of methods to quantitatively analyze the relatedness of scientific publications (De Bellis, 2009; Garfield, 1972), such as scholarly networks, publication or venue importance, and coauthorship analysis. Citation analysis along with graph mining is a common bibliometric method, which has been successfully used to enhance scientific information retrieval (Bernstam et al., 2006). In most previous works, while various methods were used to characterize the citation network, the basic assumption was easy and straightforward: Either Publication1 cites Publication 2 or Author1 cites Author 2, regardless of sentiment, reason, topic, or motivation.

However, this is an oversimplified assumption about the citation process (Cronin, 1984), as recent studies have shown (Shotton, 2009) that the reason or motivation for a citation matters. A number of classification schemes have been produced to capture the reasons for citing. Taking the Citation Typing Ontology as an example, Shotton (2009) theoretically captured the intent of citations and allowed authors to categorize the reasons for their citations by providing a taxonomy: confirm, correct, credit, critique, disagree with, discuss, extend, or obtain background from a study. Similarly, Liu, Qin, and Chen (2011), proposed the structural descriptive referential model to capture the domain-specific structural knowledge of each citation (i.e., research question, methodology, data set, or evaluation for information-retrieval-related papers). However, most of these studies have stayed on the conceptual level, for two reasons. First, most researchers are only able and willing to...
provide simple reference metadata due to the time and skill required to create more sophisticated metadata. Creating refined referential metadata would be beyond most authors’ capacity. Second, fully automatic citation reasoning or classification requires a large amount of training data, which is unavailable for most research disciplines. In this research, instead of classifying citations, a context window along with supervised topic modeling is utilized to statistically characterize each citation.

The combination of citation bibliometrics and text mining provides a synergy unavailable with each approach taken independently (Kostoff, del Río, Humenik, García, & Ramírez, 2001). In our research, we used a supervised topic modeling algorithm, Labeled Latent Dirichlet Allocation (LLDA) (Ramage, Hall, Nallapati, & Manning, 2009), to infer the publication and citation topic distribution, where each topic is a probability distribution of words and the label of the topic is an author-contributed publication keyword. The publication and citation topic probability distributions, then, can be converted to the vertex (publication) prior and edge (citation) transitioning probability distributions to enhance citation network PageRank (with prior distributions) for publication ranking. More specifically, we assume that words surrounding a target citation (citation context) can provide semantic evidence to infer the topical reason or motivation for the target citation, and that a citation network with prior (topic) knowledge can enhance classical bibliometric analysis (i.e., based on the citation context, if a cited paper contributes to the core topic(s) of the citing paper, this cited paper should get more credit from the citing paper (higher transitioning probability). Because each vertex or edge on the citation network is associated with a topic probability distribution, an enhanced PageRank algorithm can generate an authority vector, and each score in the vector tells the publication’s topical importance.

To evaluate these novel bibliometric analysis and publication-ranking methods, we examined 104 review papers for a number of selected topics (keywords). The cited publications for each review paper were assumed to be “important publications” for each target topic, and we used different ranking algorithms—PageRank (Page, Brin, Motwani, & Winograd, 1999), language model (Song & Croft, 1999), BM25 (Robertson, Walker, Jones, Hancock-Beaulieu, & Gatford, 1995), term frequency–inverted document frequency (tf-idf) (Jones, 1972), PageRank + language model, and our new approach—to locate these important publications in the publication repository for each scientific topic. The results based on mean average precision (MAP), to test if a paper will be cited based on a specific topic, and normalized discounted cumulative gain (nDCG) (Järvelin & Kekäläinen, 2002), to test how many times a paper will be cited based on a specific topic, show that our approach significantly ($p < .001$) outperforms other methods. For example, our approach improved on the PageRank + language model method (the strongest base method) by 34.5%.

In the remainder of this article, we (a) introduce our novel methods for constructing a bibliometric citation graph with prior distributions via full-text topic modeling; (b) review relevant literature and methodology for bibliometric analysis, topic modeling, and network mining; (c) describe the experiment setting and evaluation results; and (d) discuss the findings and limitations of the study and identify subsequent research steps.

Previous Research

In this section, we survey existing studies that have focused on three fields: citation analysis, PageRank analysis for citation network, and topic modeling for scientific publications.

Academic publications can be characterized as well-defined units of work, roughly similar in quality and number of citations as well as in their purpose. For this reason, citation analysis is a way to analyze relationships between publications and their relative influence.

Since the 1900s, scientists and librarians have been conscious of the growth of the research literature. Garfield (1972) described a method to evaluate scientific journals by the frequency and impact of citations using data from the Science Citation Index (SCI).

Drawing on these classic bibliometrics papers, many scholars have focused their research on citation frequency and citation impact and applied them in different domains. More recently, citation information has been successfully used to enhance information retrieval performance (e.g., Bernstam et al., 2006; Ritchie, Robertson, & Teufel, 2008). Harhoff, Narin, Scherer, and Vopel (1999) judged the value of patented inventions by citation frequency and concluded that “the higher an invention’s economic value estimate was, the more the patent was subsequently cited” (p. 511). Other authors have studied the association between the citation frequency of ecological articles and various characteristics of journals, articles, and authors (Leimu & Koricheva, 2005) and have concluded that annual citation rates of ecological papers are affected by many factors such as the hypothesis tested, article length, and authors’ information. This casts doubt on the validity of using citation counts for academic evaluation.

Increasing numbers of researchers have come to doubt the reasonableness of assuming that the raw number of citations reflects an article’s influence (MacRoberts & MacRoberts, 1996). On the other hand, full-text analysis has to some extent compensated for the weaknesses of citation counts and has offered new opportunities for citation analysis.
Traditionally, citation analysis treats all citations equally. However, in reality, not all citations are equal. Some scholars have considered location to be a factor affecting the relative importance of a citation. Herlach (1978) found that a publication cited in the introduction or literature review section and mentioned again in the methodology or discussion sections is likely to make a greater overall contribution to the citing publication than will others that have been mentioned only once. The stylistic aspects of a citation also matter. Bonzi (1982) distinguished between three broad categories of citations: Those not specifically mentioned in the text (e.g., “Several studies have dealt with...”), those barely mentioned (e.g., “Smith has studied the impact of...”), direct quotation or discussion point (e.g., “Smith found that...”), and two or more quotations or discussion points in the text. Small (1978) utilized citation context to characterize each citation as concept symbol for a set of highly cited articles in chemistry, and he found that many highly cited documents in chemistry have uniform or standardized usage and meaning. More recently, Ritchie et al. (2008) found that the citation context, a text window containing the target citation, can be employed to identify the semantics of the cited paper. Indexing those terms can effectively help a system improve retrieval and ranking effectiveness.

PageRank has become a significant method for evaluating the most important nodes in complex graph analysis. Examples include social networks, web graphs, telecommunication networks, and biological networks. From the point of citation analysis in bibliometrics, PageRank is also an efficient way to evaluate a paper’s ranking score in a specific domain and decide “which entities are most important in the network relative to a particular individual or set of individuals” (White & Smyth, 2003, p. 266). The PageRank algorithm, first proposed by Page et al. (1999) and used in Google Search, is a method for computing a ranking score for every web page based on a graph of the web to measure the relative importance of web pages. Different than the traditional method of simple backlink counts, PageRank utilizes the graph to recalculate the ranking of each web page based on backlinks. This means that a page has a high rank when it has many backlinks or has a few highly ranked backlinks. PageRank is the most widely used method for citation analysis of web pages and has become a popular research area.

White and Smyth (2003) first proposed the priors idea in their formalization of a relative-rank extension to both PageRank and Hyperlink-Induced Topic Search (HITS). They experimentally evaluated different properties of some algorithms (social networks, graph theory, Markov models, and web graph analysis) on toy graphs and demonstrated how the approach could be used to study relative importance in real-world networks. Rodriguez and Bollen (2006) described implementation of a particle-swarm that can simulate the performance of the popular PageRank algorithm in both its global-rank and relative-rank incarnations. PageRank with priors is used in this article to compute the publication topic importance score with the node prior and edge transitioning probability vectors.

Although more and more publications have focused on PageRank, most previous research for improving the ranking of search-query results has computed a single vector using a link structure of the network, which is independent of particular search queries. Chakrabarti, Joshi, Punera, and Pennock (2002) showed that pages tend to point to other pages on the same “broad” topic, suggesting that it may be possible to improve the performance of link-based computations by taking page topics into account. Based on this theory, Haveliwala (2003) proposed computing a set of PageRank vectors biased using a set of representative topics to capture more accurately the notion of importance with respect to a particular topic. By computing the topic-sensitive PageRank scores using the topic of the context in which the query appeared and then generating context-specific importance scores for pages using linear combinations of biased PageRank vectors, the proposed algorithm can generate more accurate rankings compared with a single, generic PageRank vector. Topics were represented by term vectors extracted from web documents under 16 top categories of the Open Directory.

Latent Dirichlet Allocation (LDA; Blei, Ng, & Jordan, 2003) is another widely used method for topic modeling. It is a generative probabilistic model for collections of discrete data such as text corpora. Each topic is modeled as an infinite mixture over an underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide, in theory, an explicit representation of a document. Recent work has shown that LDA-based topic modeling can be integrated to scholarly network-based citation analysis. For instance, Nallapati Ahmed, Xing, and Cohen (2008) used the Pairwise-Link-LDA and the Link-PLSALDA models for citation prediction, and the Relational topic model (Chang & Blei, 2009) was used to summarize a network of documents, predict links between them, and predict words within them. Meanwhile, the topic model can be used to identify most influential documents in a corpus without using citation linkage (Gerrish & Blei, 2010). By using changes in the thematic content of documents over time, a dynamic topic model based on LDA is employed for quantifying and qualifying the impact of these documents. However, optimizing topic number for LDA and interpreting each statistical topic are challenging for citation analysis.

LLDA (Ramage et al., 2009) is a supervised topic model that constrains LDA by defining a one-to-one correspondence between LDA’s latent topics and user tags. LLDA can directly learn word–tag correspondences, which have been demonstrated to improve expressiveness over traditional LDA with visualizations of a corpus of tagged web pages. It is another promising method to model topics from full-text documents, and could be used to optimize the PageRank algorithm, especially for bibliometric analysis evaluation and interpretation.
Research Methods

Most previous citation analysis studies share a common assumption: if paper1 cites paper2, then paper1 and paper2 are connected. Most of the time, the reasons or motivations for this putative connection are ignored. Here, we characterize citation relations in terms of two kinds of prior knowledge: publication (citing or cited paper) topic probability distribution and citation topic probability distribution; these are illustrated in Figure 1.

Within this framework, each publication makes different degrees of contribution for different scientific topics, and each citation is characterized by a topic probability distribution inferred by the citation’s surrounding (context) words.

There are three major contributions of this research. First, even with the same citation network topology, different publications can make different contributions (i.e., have different authority) to different scientific topics. In addition, topic authorities cannot be uniformly distributed to other cited publications in terms of the citation topic distributions’ inferred transitioning probabilities. Second, unlike classical, unsupervised topic modeling algorithms, the topics in this research are associated with scientific keywords (supervised learning), which can help to interpret and evaluate the results. Last, because we utilize full-text citation analysis, one paper can have more than one citation edge with the other paper. For instance, if paper1 cites paper2 three times, there will be three distinct edges on the citation network between these two papers. Hence, the accumulated transitioning probabilities between paper1 and paper2 can be higher than others, resulting in more accurate PageRank random walk modeling.

Topic Modeling with Labels

Blei et al. (2003) proposed LDA as a promising unsupervised topic modeling algorithm. LDA employs a generative probabilistic model in the hierarchical Bayesian framework, and extends Probabilistic Latent Semantic Indexing (PLSI) by introducing a Dirichlet prior on θ. As a conjugate prior for the multinomial topic distribution, the Dirichlet distribution assumption has some advantages, which can simplify the problem. The probability density of a T-dimensional Dirichlet distribution over the multinomial distribution

\[ p = (p_1, p_2, \ldots, p_T), \]  

where \( \sum p_j = 1 \), is defined by:

\[
\text{Dir}(\alpha_1, \alpha_2, \ldots, \alpha_T) = \frac{\Gamma \left( \sum_j \alpha_j \right)}{\prod_j \Gamma(\alpha_j)} \prod_j p_j^{\alpha_j-1} 
\]

where \( \alpha_1, \alpha_2, \ldots, \alpha_T \) are parameters of the Dirichlet distribution. These parameters can be simplified to a single value \( \alpha_{LDA} \), the value of which is dependent on the number of topics.

However, one limitation of LDA is the challenge of interpreting and evaluating the statistical topics. For example, it is difficult to automatically assign a label to (i.e., provide a semantics for) each statistical topic. In addition, arbitrary numbers of topics may not be appropriate for bibliometric studies because while some topics may be very sparse, others may only focus on quite detailed knowledge of the same scientific topic. These limitations motivated us to utilize a supervised or semisupervised topic modeling algorithm, one stemming from LDA, which employs existing topics from scientific metadata.

Here, we assume that each (author-assigned) scientific keyword is a topic label and that each scientific publication is a mixture of its author-assigned topics (keywords). As a result, both topic labels and topic numbers (the total number of keywords in the metadata repository) are given. The LLDA algorithm (Ramage et al., 2009) was used in training the labeled topic model. Unlike the LDA method, LLDA is a supervised topic modeling algorithm that assumes the availability of topic labels (keywords) and the characterization of each topic by a multinomial distribution \( \beta_{key} \) over all vocabulary words. For example, Table 1 is an example of the keyword–topic probability.

During the Bayesian generative topic modeling process, each word \( w \) in a publication is chosen from a word distribution associated with one of that paper’s labels (keywords). The word is picked in proportion to the publication’s preference for the associated label \( \theta_{paper, key} \) and the label’s preference for the word \( \beta_{key,w} \). Figure 2 visualizes the LLDA generative process. For each topic (keyword) \( key \), one draws a multinomial distribution \( \beta_{key} \) from the symmetric Dirichlet prior \( \phi \). Then, for each publication, one builds a label set \( \Lambda \) paper for the deterministic prior \( \phi \). Finally, one selects a multinomial distribution \( \theta_{paper} \) over the labels \( \Lambda \) paper from Dirichlet prior \( \alpha \).

Publication Topic Inference

Paper (author provided) keywords can provide high-quality topic labels for each scientific publication; however, this is not an ideal solution in that a large number of publications in the metadata repository have very few keywords,
and often not enough to cover all potential topics of the target paper. For example, after examining 200,000 publications from the ACM digital library, we found that 41.49% had no keyword information (either keyword metadata were missing or authors did not provide any), and 6.13% had only one or two keywords, which is probably not enough to cover the whole paper.

To cope with this problem, we used several different approaches to infer the topic distribution for each publication:

- **All topics (ALL):** As the easiest approach, we assumed that all publications in the repository may be related to all possible topics extracted by LLDA. So we used publication title + abstract + full text to infer the topic distribution on any \( \alpha \) in the topic space. For this approach, author keyword metadata were not used.
- **Greedy match (GM):** For this approach, we assumed that author-assigned keywords were not enough to cover the semantics of the paper, and used greedy matching to expand the paper topic space. First, we loaded all possible keyword (topic label) strings into memory and then searched each keyword from the paper title and abstract by using greedy matching. For example, if **music information retrieval** existed in the title, we did not use the keyword **information retrieval**. Matched keywords were used as "pseudo-keyword" metadata for the target publication. For the \{"Author keywords" + "Pseudo-keywords"\} collection, we used LLDA inference to assume topic probability scores. All topics not in this collection were ignored.
- **Mutual information (MI):** For this approach, we assumed that all keywords from the greedy match approach were related and that mutual information could be used to further expand the publication keywords. Thus, if a target paper had keyword collection \( \text{Paper}_{\text{key}} = \{\text{key}_1, \text{key}_2, \ldots, \text{key}_m\} \), then further keywords \( \text{key}_i \), where \( \text{key} \notin \text{Paper}_{\text{key}} \), were scored based on mutual information (MI):

\[
\text{Score}(\text{key}_i) = \frac{\sum_{i=1}^{n} P(\text{Paper}_{\text{key}})\cdot \text{MI}(\text{key}_i, \text{key}_x)}{\sum_{i=1}^{m} P(\text{Paper}_{\text{key}})}.
\]

As a ranking function, any keyword \( \text{key}_i \) can be related to the paper if \( \text{Score}(\text{key}_i) \) is large and \( \text{key}_i \) is highly ranked.

For either the GM or MI approaches, a subset of keywords (topics) from the training LLDA model were used to infer the paper topic distribution. The topic scores for \( \text{key}_i \) (i.e., \( P(z_{\text{key}}, \text{paper}) \)), were normalized for future experiments, where \( \sum_{i=1}^{m} P(z_{\text{key}}, \text{paper}) = 1 \).

### Citation Topic Inference

As Figure 1 shows, each citation context in the citing paper is located for this research. One reference could be cited more than once in a paper, and the citation distributions could be different.

The text window surrounding the target citation, \([-n \text{ words}, +n \text{ words}]\), is used to infer the citation topic distribution via LLDA. Intuitively, \( n \) should be a small number, as nearby words should provide more accurate citation information. However, \( n \) should not be too small to minimize randomness. In this experiment, we used an arbitrary parameter setting, where \( n = 150 \). However, the ideal parameter setting should be further trained; that is a task for future work.

We considered two different hypotheses for citation topic distributions:

**H1: All topics (ALL).** As with publication topic inference, we assumed that all citations in the repository were related to all possible topics extracted by LLDA.

**H2: Citing + cited topics only (CC).** For this approach, we assumed that citations may not relate to all topics in the LLDA model. Instead, citations may relate only to topics provided by citing or cited topics. For any topic, \( z_{key} \), not in a citing or cited paper, we gave the citation a lower score, \( P'(z_{key}, \text{citation}) = \psi \cdot P(z_{key}, \text{citation}) \). We set \( \psi = 0.1 \) for this research, as we did not want to totally remove these citations in the graph or make the citation transitioning probability equal zero in the citation network. As with publication topic inference, citation distributions for this method were normalized.

![LLDA allocation algorithm](image)
Citation Networks with Publication and Citation Priors

Classical citation networks tend to ignore citation and publication content. In this study, we created a large citation-directed network, \((G = V, E)\), with two kinds of prior knowledge: publication topic prior and citation topic transitioning probability distribution.

Each vertex, \(v \in V\), on the citation graph represents a publication, with the publication topic prior probability vector \(\{p_{z_{key}^v}(v), p_{z_{key}^v}(v), \ldots p_{z_{key}^v}(v)\}\), where \(p_{z_{key}^v}(v)\) is the prior probability of vertex \(v\) for topic \(z_{key}\) and \(\sum_{v=1}^{n} p_{z_{key}^v}(v) = 1\).

Each edge, \(e \in E\), on the graph represents a citation connecting \(u_i\) and \(u_j\) (\(u_i\) cites \(u_j\)). The topic transitioning vector for each edge is \(\{p_{z_{key}^v}(v_i|v_j), p_{z_{key}^v}(v_i|v_j), \ldots p_{z_{key}^v}(v_i|v_j)\}\), where \(p_{z_{key}^v}(v_i|v_j)\) is the probability of transit from vertex \(v_i\) to \(v_j\) for topic \(z_{key}\).

For a given publication \(u\), we used \(S_u(u)\) and \(S_u^i(u)\) to represent a set of incoming and outgoing edges (citations) to node \(u\), with “in” degree \(d_s(u) = \sum_{u=1}^{n} S_u(u)\) and “out” degree \(d_o(u) = \sum_{u=1}^{n} S_u(u)\). Thus, \(\sum_{u=1}^{n} p_{z_{key}^v}(v_i|v_j) = 1\). For example, if a publication cites only three papers, and for a specific topic, the transitioning probabilities to these three papers are \(.1, .1, \) and \(.8\), then most of the paper’s credit on this topic (topic authority) goes to the third paper.

Based on these definitions, we can calculate each vertex’s (i.e., each publication’s) prior probability:

\[
p_{z_{key}^v}(v) = \frac{P(z_{key}^v, paper_v)}{\sum_{v=1}^{n} P(z_{key}^v, paper_v)}
\]

We can also calculate each edge’s (i.e., each citation’s) transitioning probability:

\[
p_{z_{key}^v}(v_i|v_j) = \frac{P(z_{key}^v, citation_{i,j})}{\sum_{v=1}^{n} P(z_{key}^v, citation_{i,j})},
\]

where \(P(z_{key}^v, paper_v)\) is the publication topic inference score and \(P(z_{key}^v, citation_{i,j})\) is the citation topic transitioning score discussed earlier.

Unlike classical PageRank, a citation graph with vertex and edge priors permits nonuniformly distributed random jumps. Based on earlier discussion, topic distributions for each publication could be sparse for the GM and MI assumptions, and for a given topic, \(z_{key}\), the vertex prior probability, \(p_{z_{key}^v}\), for many publications could be zero. Thus, for each topic, the updated PageRank algorithm can tell the “relative importance” of vertices in \(G\) with respect to a set of “root vertices” \(R \subseteq V\), where for each \(r \in R\), \(p_{z_{key}^v} \neq 0\). Those root vertices can be thought of as the important publications given a topic (prior knowledge). A special case is the “all topics” approach, where all the topics are considered, and root vertices \(R = V\).

We used the PageRank with prior algorithm (White & Smyth, 2003) to calculate each vertex’s (topic relative) importance, \(I_{z_{key}}(v|R) = \pi_{z_{key}}(v)\) and:

\[
\pi_{z_{key}}(v)^{t+1} = (1 - \beta_k) \sum_{u=1}^{n} p_{z_{key}^v}(v|u) \pi_{z_{key}}(u)^{t} + \beta_k p_{z_{key}^v}(v).
\]

This equation represents a Markov chain for a random surfer who transits “back” to the root vertexes \(R\) with probability \(\beta\) at each time step. For each incoming link (citation) from \(v\) the PageRank score is updated with respect to edge (citation) transitioning probability \(p_{z_{key}^v}(v|u)\).

The output, for each vertex (publication), \(u\), is an authority vector \(\{A_{z_{key}^v}(v), A_{z_{key}^v}(v), \ldots A_{z_{key}^v}(v)\}\). Each authority score in the vector indicates the publication topic importance with respect to both paper topic and full-text citation priors. We can get \(n\) ranking lists as a result.

Example

To help readers better understand full-text citation analysis with citation and publication priors, we provide an example.

Assume that we have only three keywords (topics) and four publications in the repository, as depicted in the following diagram:

In this illustrative citation network, each publication or citation has a topic distribution of three keyword-based topics (upper graph). One reference may correspond to more than one citation (e.g., Publication 1 here cites Publication 2 twice).

If we only focus on Topic 1 (Keyword 1), Publication 1’s prior probability, \(p_{z_{key}^v}(paper_1)\), is 0.26. This is calculated by \(\frac{P(z_{key}^v, paper_1)}{\sum_{v=1}^{n} P(z_{key}^v, paper_1)} = 0.4/(0.4 + 0.4 + 0.33 + 0) = 0.3\).

Similarly, we can calculate all edge (citation) transitioning probabilities by using the outgoing links’ topic probability, as Figure 3 (lower graph) shows for Citation 3, by using \(\frac{P(z_{key}^v, citation_{i,j})}{\sum_{v=1}^{n} P(z_{key}^v, citation_{i,j})}\).

We can then use PageRank with these priors to calculate each vertex’s authority vector. Please note that for Topic 1, Publication 4’s prior probability is zero, which means that the root vertices for Topic 1 (highlighted) are Publications 1, 2, and 3 (i.e., these publications are more important for this topic). However, \(\pi_{z_{key}}(paper_1)\) will be greater than zero due to Publication 3’s citation with transitioning probability = 1. For example, if Keyword 1 is information retrieval, Publications 1, 2, and 3 are all focusing on this topic (author-provided prior knowledge), but Publication 4 concentrates on machine learning. Although Publication 4 is not directly related to the topic, this paper could also be important for this topic (topical authority could be large.), as it may always provide methodological support for information retrieval research (based on citation context). Also note that Publication 2 will get more credits from Publication 1, as compared with Publication 3, because there are two citation links (Citations 1 and 2) and because the associated transitioning probability scores are higher than those of Citation 3.
Evaluation Methods

Unlike unsupervised topic modeling approaches, in this study we projected full-text scientific publications and citations onto labeled topic spaces, where each topic’s label is a scientific keyword. As a result, we are able to assess and interpret the topic publication authority vector and topic publication ranking by using keyword information. However, as this research focuses on the method of calculating publication topic importance, we can hardly compare the authority vector with other classical bibliometric indicators such as the h-index or impact factor, which are topic independent.

For evaluation, we tried to find the “ground truth” of the most important publications for a specific scientific topic (keyword). We cannot use citation count for this evaluation for two reasons. First, citation topical motivation is not available on a large scale, and general citation count ranking cannot be used to validate topic-based publication ranking. Second, in this research, we assume that citation count is not an ideal indicator to demonstrate publication importance; for instance, some newly published papers or papers from lesser-known authors may not get a large number of citations even though they are important for the target topic.

To achieve this goal, a list of review or survey papers along with their cited papers was collected. Collected review papers were screened so that they focused on only one topic (keyword) (see Table 2).

All candidate review papers were manually judged “accept” or “reject” by coders from an automatically preselected pool.

We assume that if a publication is cited by a review paper, and if this review paper concentrates on keyword keyi, then this publication is important for keyi. Since the degree of importance of cited papers may be different, we used the number of citations (by a review paper) to characterize the importance. Thus, if a review paper for keyword keyi cited paperi twice and paperj once, then \( \text{Importance}_{keyi}(\text{paperi}) = 2 \) and \( \text{Importance}_{keyj}(\text{paperj}) = 1 \). We also assume that if a paper is not cited by the target review paper, then the importance of this paper for the target topic is zero. We also assume that if a paper is cited four or more times by the review paper, then its importance is equal to 4.

We clearly understand that review papers do not cover all important publications for the target topic, even though the quality of the review papers could be very high. The goal of this evaluation, however, is to compare the performance of PageRank with paper and citation priors against a list of baseline algorithms:

Citation PageRank: For this baseline, we built the citation network without publication and citation prior knowledge. We then calculated the PageRank authority score for each publication without considering keyword or topic information or citation context.

Tf–idf: For this baseline, we used the keyword (topic label) as the query (e.g., multimedia information retrieval) to search all the paper content (abstract and full text). Ranking lists based on tf–idf + vector space model, for a list of keywords, were used as the baseline. For a specific keyword, if a publication received a high ranking score, this publication was assumed to be important for the target topic.

BM25: Same as for tf–idf, except we replaced vector space with BM25.

Language model: Same as for tf–idf, except we replaced vector space with language model and used Dirichlet smoothing (Zhai & Lafferty, 2001).

Language model + PageRank: For this baseline, we combined the language model with PageRank (without topic priors), using random walk probability as the model prior:

\[
P(\text{paper}, \text{key}) = P(\text{key}, \text{paper}) \cdot P(\text{paper})
\]

where \( P(\text{key}, \text{paper}) \) is the language model score, the likelihood of keyword key given paper-content paper; with Dirichlet smoothing, and where \( P(\text{paper}) \) is the model prior, the PageRank (without
topic priors) random walk probability. \(P(\text{key})\) was ignored (paper independent) for this ranking function. For this approach, a publication was deemed important if its content matched the keyword and the paper was popular on the citation graph.

We used two indicators to measure and compare ranking algorithm performance: MAP and nDCG (Järvelin & Kekäläinen, 2002). nDCG estimates the cumulative relevance gain a user receives by examining retrieval results up to a given rank on the list. In this research, we used the importance score, 0–4, as the relevance label to calculate nDCG scores.

**Experiments**

In this section, we describe the experimental setting and results. Analysis and conclusions are presented in the next section.

**Data**

We used 41,370 publications from 111 journals and 1,442 conference proceedings or workshops on computer science for the experiment (mainly from the ACM Digital Library), where full text and citations were extracted from the PDF files. The selected papers were published between 1951 and 2011. From these, we extracted 28,013 publications’ text (accounting for 67.7% of all the sampled publications), including titles, abstracts, and full text. For the other publications, we used the title, the abstract, and information from a metadata repository to represent the content of the paper.

We then wrote a list of regular expression rules to extract all the possible citations from paper’s full text. For instance, the rules could extract “... [number] ...” and “... [number, number ... , number] ...” as citations from the content of publication. Each citation extracted from the publication text was associated with a reference (cited paper ID). Of course, citation extraction based on regular expressions is not a perfect solution because differences in encoding, format, or citation style may threaten citation-extraction performance. In a total of 223,810 references (paper1 cites paper2 relations), we successfully identified 94,051 references, which accounted for 42.0% of all references. Of course, references may have been cited more than once in a citing paper and located in multiple sections.

**LLDA Model Training**

We sampled 10,000 publications (with full text) to train the LLDA topic model. Author-provided keywords were used as topic labels. For instance, the current article has six author-provided keywords. Thus, our LLDA training would have assumed that this article is a multinomial distribution over these six topics. During preprocessing, we also clustered similar keywords if the edit distance between them was very small (e.g., \(k\text{-means}\) and \(k\ mean\)) or if two keywords shared the same stemmed root (e.g., \(web\ searches\) and \(web\ search\)).

If a keyword appeared less than 10 times in the selected publications, we removed it from the training topic space. For publication content, we first used tokenization to extract words from the title, abstract, and publication full text. If the character length of the word was greater than three, this word was removed. Snowball stemming was then employed to extract the root of the target word. We also removed the most frequent 100 stemmed words and words appearing less than three times in the training collection.

Finally, we trained an LLDA model with 3,911 topics (keywords). These topics were used to infer the publication and citation topic distribution.

**Results**

By using the method proposed earlier, we constructed a directed citation graph with each vertex as a publication, with its associated publication topic distribution, and each edge as a citation, with its citation topic distribution. For each topic, we then calculated each publication’s root prior probabilities and each citation’s transitioning probabilities. Note that from each node there are three ways to compute a publication’s topic distribution: (a) \(ALL\) (all 3,911 topics), (b) \(GM\) (author keyword + greedy match from title and abstract), and (c) \(MI\) (GM + keywords with high mutual information score). Further, from each citation there are two ways to calculate a topic distribution: (a) \(ALL\) (all 3,911 topics) and (b) \(CC\) (topics from citing or cited papers).

Hence, there were five groups of results (i.e., only five because if we used all topics for the publication inference, then the two citation inference methods would be the same).

As mentioned earlier, MAP scores for this evaluation are fairly small (as compared with information retrieval systems) because we used review papers’ references as the ground truth data, which only cover a small proportion of important publications on the target topics. However, the evaluation data set quality is very high because those references come from domain (topic) experts.

In terms of result comparison, we found that topic publication ranking performance was sensitive to the publication topic inference methods. We found that GM and MI outperformed ALL. On the other hand, various citation distribution inference methods achieved similar results. In particular, the GM method was slightly better than was MI for publication topic inference.

We then used GM + CC to compare with other baseline methods, including PageRank, tf–idf, BM25, language model, and PageRank + language model. The results are presented in the following tables. The top performing algorithm is highlighted for each row, and these results are visualized in Figures 4 and 5.

For baseline ranking methods, the PageRank + language model clearly achieved the top performance, and PageRank alone (topic independent) performed the worst. For \(MAP@n\), PageRank + language model outperformed our method (PageRank with prior, highlighted) when \(n \leq 50\). But for \(n \geq 100\), PageRank with priors surpassed all other
methods. We also used significance testing to compare PageRank with prior and PageRank + language model (*$p < 0.01$, **$p < 0.005$, ***$p < 0.001$). After $n \geq 1,000$, PageRank with priors was significantly better than was PageRank + language model.

nDCG@$n$ is a more important indicator in this research because it tells the degree of (publication topic) importance. If the nDCG score is large, the target algorithm can prioritize the most important on the ranking list. In Table 6 and Figure 5, it is clear that PageRank with priors is always

<table>
<thead>
<tr>
<th></th>
<th>ALL + ALL</th>
<th>GM + ALL</th>
<th>GM + CC</th>
<th>MI + ALL</th>
<th>MI + CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP@10</td>
<td>0.1467</td>
<td>0.2058</td>
<td>0.1955</td>
<td>0.1853</td>
<td>0.1876</td>
</tr>
<tr>
<td>MAP@30</td>
<td>0.1134</td>
<td>0.1797</td>
<td>0.1728</td>
<td>0.1653</td>
<td>0.1668</td>
</tr>
<tr>
<td>MAP@50</td>
<td>0.1047</td>
<td>0.1598</td>
<td>0.1581</td>
<td>0.1465</td>
<td>0.1510</td>
</tr>
<tr>
<td>MAP@100</td>
<td>0.0993</td>
<td>0.1437</td>
<td>0.1440</td>
<td>0.1346</td>
<td>0.1358</td>
</tr>
<tr>
<td>MAP@300</td>
<td>0.0766</td>
<td>0.1206</td>
<td>0.1207</td>
<td>0.1109</td>
<td>0.1129</td>
</tr>
<tr>
<td>MAP@500</td>
<td>0.0718</td>
<td>0.1145</td>
<td>0.1144</td>
<td>0.1042</td>
<td>0.1059</td>
</tr>
<tr>
<td>MAP@1000</td>
<td>0.0653</td>
<td>0.1069</td>
<td>0.1064</td>
<td>0.0972</td>
<td>0.0974</td>
</tr>
<tr>
<td>MAP@3000</td>
<td>0.0595</td>
<td>0.1016</td>
<td>0.1011</td>
<td>0.0913</td>
<td>0.0918</td>
</tr>
<tr>
<td>MAP@5000</td>
<td>0.0584</td>
<td>0.1010</td>
<td>0.1004</td>
<td>0.0906</td>
<td>0.0910</td>
</tr>
<tr>
<td>MAP@ALL</td>
<td>0.0517</td>
<td>0.0814</td>
<td>0.0816</td>
<td>0.0741</td>
<td>0.0751</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>ALL + ALL</th>
<th>GM + ALL</th>
<th>GM + CC</th>
<th>MI + ALL</th>
<th>MI + CC</th>
</tr>
</thead>
<tbody>
<tr>
<td>nDCG@10</td>
<td>0.0626</td>
<td>0.1027</td>
<td>0.0980</td>
<td>0.0923</td>
<td>0.0916</td>
</tr>
<tr>
<td>nDCG@30</td>
<td>0.0791</td>
<td>0.1187</td>
<td>0.1187</td>
<td>0.1140</td>
<td>0.1150</td>
</tr>
<tr>
<td>nDCG@50</td>
<td>0.0903</td>
<td>0.1370</td>
<td>0.1367</td>
<td>0.1305</td>
<td>0.1320</td>
</tr>
<tr>
<td>nDCG@100</td>
<td>0.106</td>
<td>0.1559</td>
<td>0.1526</td>
<td>0.1504</td>
<td>0.1517</td>
</tr>
<tr>
<td>nDCG@300</td>
<td>0.1333</td>
<td>0.1988</td>
<td>0.1988</td>
<td>0.1938</td>
<td>0.1967</td>
</tr>
<tr>
<td>nDCG@500</td>
<td>0.1471</td>
<td>0.2199</td>
<td>0.2179</td>
<td>0.2214</td>
<td>0.2207</td>
</tr>
<tr>
<td>nDCG@1000</td>
<td>0.1675</td>
<td>0.2424</td>
<td>0.2423</td>
<td>0.2452</td>
<td>0.2488</td>
</tr>
<tr>
<td>nDCG@3000</td>
<td>0.2049</td>
<td>0.2747</td>
<td>0.2737</td>
<td>0.2794</td>
<td>0.2808</td>
</tr>
<tr>
<td>nDCG@5000</td>
<td>0.2176</td>
<td>0.2830</td>
<td>0.2825</td>
<td>0.2867</td>
<td>0.2877</td>
</tr>
<tr>
<td>nDCG@ALL</td>
<td>0.2743</td>
<td>0.3202</td>
<td>0.3189</td>
<td>0.3193</td>
<td>0.3204</td>
</tr>
</tbody>
</table>

Note. MAP = mean average precision; ALL = all topics; GM = greedy match; CC = cited and citing topics only; MI = mutual information.
FIG. 5. Different publication and citation inference methods (nDCG). [Color figure can be viewed in the online issue, which is available at wileyonlinelibrary.com.]

TABLE 5. Different publication and citation inference methods (MAP).

<table>
<thead>
<tr>
<th></th>
<th>PageRank</th>
<th>tf–idf</th>
<th>BM25</th>
<th>Language model</th>
<th>Language model + PageRank</th>
<th>PageRank with prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAP@10</td>
<td>0.0168</td>
<td>0.1551</td>
<td>0.1637</td>
<td>0.163</td>
<td><strong>0.2039</strong></td>
<td>0.1955</td>
</tr>
<tr>
<td>MAP@30</td>
<td>0.0192</td>
<td>0.1387</td>
<td>0.1397</td>
<td>0.1498</td>
<td><strong>0.1872</strong></td>
<td>0.1728</td>
</tr>
<tr>
<td>MAP@50</td>
<td>0.0186</td>
<td>0.1295</td>
<td>0.1254</td>
<td>0.138</td>
<td><strong>0.1702</strong></td>
<td>0.1581</td>
</tr>
<tr>
<td>MAP@100</td>
<td>0.0182</td>
<td>0.1171</td>
<td>0.1151</td>
<td>0.1198</td>
<td>0.1424</td>
<td>0.144</td>
</tr>
<tr>
<td>MAP@300</td>
<td>0.0162</td>
<td>0.0918</td>
<td>0.0904</td>
<td>0.0935</td>
<td>0.1106</td>
<td>0.1207</td>
</tr>
<tr>
<td>MAP@500</td>
<td>0.0145</td>
<td>0.0858</td>
<td>0.0851</td>
<td>0.0864</td>
<td>0.1001</td>
<td>0.1144</td>
</tr>
<tr>
<td>MAP@1000</td>
<td>0.011</td>
<td>0.0754</td>
<td>0.0756</td>
<td>0.0759</td>
<td>0.0918</td>
<td><strong>0.1064</strong>*</td>
</tr>
<tr>
<td>MAP@3000</td>
<td>0.0072</td>
<td>0.064</td>
<td>0.0652</td>
<td>0.0672</td>
<td>0.081</td>
<td><strong>0.1011</strong>*</td>
</tr>
<tr>
<td>MAP@5000</td>
<td>0.006</td>
<td>0.0614</td>
<td>0.0626</td>
<td>0.0646</td>
<td>0.078</td>
<td><strong>0.1064</strong>*</td>
</tr>
<tr>
<td>MAP@ALL</td>
<td>0.0037</td>
<td>0.0415</td>
<td>0.0418</td>
<td>0.0438</td>
<td>0.0542</td>
<td><strong>0.0816</strong>**</td>
</tr>
</tbody>
</table>

Note. *\(t < 0.01\). **\(t < 0.005\). ***\(t < 0.001\).

TABLE 6. Different publication and citation inference methods (normalized discounted cumulative gain; nDCG).

<table>
<thead>
<tr>
<th></th>
<th>PageRank</th>
<th>tf–idf</th>
<th>BM25</th>
<th>Language model</th>
<th>Language model + PageRank</th>
<th>PageRank with prior</th>
</tr>
</thead>
<tbody>
<tr>
<td>nDCG@10</td>
<td>0.0093</td>
<td>0.0674</td>
<td>0.0689</td>
<td>0.0713</td>
<td>0.0901</td>
<td><strong>0.098</strong></td>
</tr>
<tr>
<td>nDCG@30</td>
<td>0.0076</td>
<td>0.0741</td>
<td>0.0738</td>
<td>0.0757</td>
<td>0.0945</td>
<td><strong>0.1187</strong>*</td>
</tr>
<tr>
<td>nDCG@50</td>
<td>0.0084</td>
<td>0.0833</td>
<td>0.0832</td>
<td>0.0861</td>
<td>0.1071</td>
<td><strong>0.1367</strong>*</td>
</tr>
<tr>
<td>nDCG@100</td>
<td>0.0107</td>
<td>0.0975</td>
<td>0.0957</td>
<td>0.1006</td>
<td>0.1251</td>
<td><strong>0.1526</strong>*</td>
</tr>
<tr>
<td>nDCG@300</td>
<td>0.0198</td>
<td>0.1266</td>
<td>0.126</td>
<td>0.1329</td>
<td>0.1552</td>
<td><strong>0.1988</strong>**</td>
</tr>
<tr>
<td>nDCG@500</td>
<td>0.0261</td>
<td>0.1391</td>
<td>0.138</td>
<td>0.1446</td>
<td>0.1685</td>
<td><strong>0.2179</strong>***</td>
</tr>
<tr>
<td>nDCG@1000</td>
<td>0.0392</td>
<td>0.1541</td>
<td>0.1525</td>
<td>0.1616</td>
<td>0.1859</td>
<td><strong>0.2425</strong>***</td>
</tr>
<tr>
<td>nDCG@3000</td>
<td>0.0719</td>
<td>0.1827</td>
<td>0.1808</td>
<td>0.1872</td>
<td>0.2128</td>
<td><strong>0.2737</strong>***</td>
</tr>
<tr>
<td>nDCG@5000</td>
<td>0.0917</td>
<td>0.1932</td>
<td>0.1895</td>
<td>0.1987</td>
<td>0.2227</td>
<td><strong>0.2825</strong>***</td>
</tr>
<tr>
<td>nDCG@ALL</td>
<td>0.1904</td>
<td>0.2141</td>
<td>0.213</td>
<td>0.2174</td>
<td>0.2371</td>
<td><strong>0.3189</strong>***</td>
</tr>
</tbody>
</table>

Note. *\(t < 0.01\). **\(t < 0.005\). ***\(t < 0.001\).
then obviously cited 1 in this subgraph makes a significant contribution to the citing paper because (a) this paper is cited three times by the citing paper and because (b) the sum of the Topic 1 transitioning probability between the citing paper and cited 1 = 0.85. Note that (a) and (b) are not independent; “multiple cited” could increase the accumulated transitioning probability from the citing paper to the cited paper while this score also is determined by the citation context. This shows that cited 1 made a major contribution to the citing paper on Topic 1. So, most of the citing paper’s credit on the first topic will go to cited 1, but not cited 2. For these reasons, our method can make a new paper or an unknown paper become popular in a shorter time as compared with other algorithms such as PageRank. This is very important for academic information retrieval and recommendation systems.

Limitations and Future Work

Limitations of this work are twofold. With respect to data, our test corpus came mostly from the ACM Digital Library, from which we cannot access full-text data for all papers. In our experiment, we extracted only 67.7% of the papers’ full text, and most of those papers were published after 1995 (because old paper PDF files are scanned, we cannot extract text directly from them). As mentioned earlier, when full text was unavailable, we used the title and abstract as a compromise, but this can be biased. In the future, this problem could be fixed by using image-based text recognition.

Another problem is that we identified only 42.0% of the references in the paper text. The main reason is, again, lack of full-text data. But we also faced additional challenges having to do with different citation styles, formatting errors, and encoding problems. These problems need to be addressed in future work.

With respect to evaluation, because we proposed a topic-based ranking method, some well-established, but not topic-based, bibliometric algorithms such as the h-index and impact factor cannot be used directly as the baseline. In future work, we will tailor our method to facilitate comparison with other bibliometric methods. In addition, we will couple our method with other bibliometric methods for better scientific publication, author, and venue characterization by, for example, introducing topical h-index or topical impact factors based on full-text citation.

References


De Bellis, N. (2009). Bibliometrics and citation analysis: From the Science Citation Index to cybermetrics. Lanham, MD: Scarecrow Press.


Shotton, D. (2009). CiTO, the Citation Typing Ontology, and its use for annotation of reference lists and visualization of citation networks. Paper presented at the Bio-Ontologies 2009 Special Interest Group meeting at ISMB.


