Citation Recommendation via Proximity Full-Text Citation Analysis and Supervised Topical Prior

Xiaozhong Liu[1], Jinsong Zhang[2], Chun Guo[1]
[1] Indiana University Bloomington, USA
[2] Dalian Maritime University, China

Abstract

Currently the many publications are now available electronically and online, which has had a significant effect, while brought several challenges. With the objective to enhance citation recommendation based on innovative text and graph mining algorithms along with full-text citation analysis, we utilized proximity-based citation contexts extracted from a large number of full-text publications, and then used a publication/citation topic distribution to generate a novel citation graph to calculate the publication topical importance. The importance score can be utilized as a new means to enhance the recommendation performance. Experiment with full-text citation data showed that the novel method could significantly ($p < 0.001$) enhance citation recommendation performance.

Keywords: Bibliometrics; Citation Recommendation; Supervised Topic Modeling; PageRank; Prior Knowledge
doi: 10.9776/16164

Contact: liu237@indiana.edu, jinsong_zhang@dlmu.edu.cn, chunguo@umail.iu.edu

1 Introduction and Motivation

In the last decade, scholarly publication has changed considerably: The volume of publications has increased dramatically, and many publications are now available electronically and online, rather than via traditional print media. This has had a significant effect not only on how scholars perceive, retrieve, and consume publications, but also on the types of publications that are available. The availability of many publications in online form enables a fast turnaround for researchers between the time when results are generated and when they become broadly available.

While rapid access to digital publications can accelerate research and education, several challenges must be addressed (Liu, 2013): 1) As domain knowledge in most disciplines expands at a frenetic pace, the sheer volume of scholarly publications available online makes it impossible for a researcher or student to absorb all the new information. Researchers need information retrieval (IR), information extraction (IE), and recommendation tools that can quickly filter through and locate relevant publications or scientific resources. Current scientific search tools, e.g., Google Scholar and Microsoft Academic, are limited to standardized types of queries to address users' information needs. 2) Understanding the content of scientific publications remains daunting. For instance, for a junior researcher or a researcher new to a discipline, a large number of complex publications required to read for a research topic are found too difficult, challenging, and overloading. 3) Some recent exciting developments, i.e., CiteRank (Walker, Xie, Yan, & Maslov, 2007) and Citation Influence Model (Dietz, Bickel, & Scheffer, 2007), have illustrated the possibility of using enhanced citation relationship to recommend high quality research publications to users. However, in most previous works (Liu, Zhang, & Guo, 2012; Liu, Zhang, & Guo, 2013; Liu, Yu, Guo, Sun, & Gao, 2014), while various methods were used to characterize the citation relationship, the basic assumption was easy and straightforward: all that matters is whether *Publication1 cites Publication2*, regardless of sentiment, reason, topic, or motivation. But this assumption is oversimplified and may limit the retrieval or recommendation performance or accuracy.

In this study, we propose an innovative citation recommendation method, which enhances citation recommendation performance and provides a friendly interface to locate the citations by a (user) textual working context, *T*, i.e., a publication abstract. By integrating bibliometric network analysis and supervised topic modeling techniques, for each paper, we calculate both a publication topical importance vector based on citation context proximity in the citing paper and a publication topic distribution. We then use the publication topical importance vector as the topical prior probability to enhance the classical language model for citation recommendation. In more detail, we assume that citation context in the citing paper along with citing and cited publication content can provide high-quality citation topical motivation

information, and we employ this information to infer the knowledge transitioning probability between citing and cited papers on a scholarly network. We then use the PageRank with prior algorithm (White & Smyth, 2003) to characterize publication topic importance. The importance scores were used as ***publication topical prior*** probabilities, $P_{Z_t}(pub_i)$, to enhance the language model, $P(T|pub_i)$, for citation recommendation.

In order to validate this novel citation recommendation method, we extracted full text and citation context of 41,370 papers from the ACM publication corpus as candidate citations. Another 274 publications were then sampled for evaluation purposes. Two baseline algorithms were used for comparison: 1) the classical language model, and 2) the language model + the citation-based PageRank score (without topic information and citation context) as prior. The evaluation results, based on MAP and nDCG, show that the new method, which considers citation context and topic information, can significantly (*p<0.001*) enhance citation recommendation performance. Some terminology mentioned in this paper is listed below:

Term	Definition
Citation recommendation	Given a textual working context, T, i.e., a publication abstract, recommend a list of ranked publications as candidate cited papers.
Supervised topic modeling	By using Labeled LDA (LLDA), each topic, z_{key_t}, is a multinomial word probability distribution, and the topic is labeled by an author contributed keyword key_t.
Citation context	Citation's surrounding (context) words in the citing paper.
Publication (node) topic distribution	Each publication is represented as a vertex, v, and a topic distribution, $p_{z_{key_t}}(v)$, on the scholarly network.
Citation (edge) topic distribution	Each citation (between two papers v_i and v_j) is represented as an edge and a topic transitioning distribution, $p_{z_{key_t}}(v_i|v_j)$, on the scholarly network.
Publication topical prior	For a topic z_{key_t}, the prior (importance) probability of a paper, i.e., $p_{z_{key_t}}(pub_k)$.

Table 1: Terminology List

2 Research Methods

Classical content-based recommendation is performed using cosine similarity along with the TF–IDF weighting scheme for terms occurring in documents and computational user profiles. In this study, we used a textual working context, i.e., a paper abstract, to represent a user's information need, and we used this input to recommend candidate citations. A similar study launched by He, Pei, Kifer, Mitra and Giles (2010) proposed a method to recommend global and local citations based on a given piece of text.

2.1 Citation recommendation with textual working context

From a content-based retrieval or recommendation perspective, given a piece of working textual context (from user), T, and a candidate citation (cited publication), pubi, we want to estimate the probability that pubi is relevant and important to the given context T, $P(pub_i|T)$, for ranking, which can be expressed as the following formula with Bayes' rule:

$$P(pub_i|T) = \frac{P(T|pub_i) \cdot P(pub_i)}{\cancel{P(T)}}$$

The motivation for applying Bayes' rule in this formula is that the probabilities on the right-hand side, $P(T|pub_i)$, can be estimated more accurately and easily than the probabilities on the left-hand, $P(pub_i|T)$, side (Kraaij, Westerveld, & Hiemstra, 2002). For classical content-based content recommendation algorithms, we simply ignore the publication prior, $P(pub_i)$, which means, without respect to T, that all publications have an equal chance to be recommended. As P(T) is publication independent, it can be ignored in this ranking function. $P(T|pub_i)$ can be estimated by using the language model, where T has a list of words {t1, t2… tm}:

$$P(T|pub_i) = \prod_{j=1}^{m} P(t_j|pub_i)$$

From a language model perspective, $P(t_j|pub_i)$ can be calculated by different smoothing techniques (Zhai, & Lafferty, 2001). So, in this research, we focus on estimating the publication prior,$P(pub_i)$, to enhance citation recommendation performance.

2.2 Using citation relation as prior

Intuitively, highly cited publications are more important than other publications, while the more important publications should have a higher opportunity to be recommended regardless of user information need (publication prior). In addition, if a publication is cited by another important paper, this paper is important.

In terms of these hypotheses, we constructed a citation network to calculate publication importance by utilizing the PageRank (Page, Brin, Motwani, & Winograd, 1999) algorithm. Each vertex on the graph is a publication, and each edge is a citation from the citing publication to the cited publication. We then used the PageRank-based publication importance as the prior,$\mathrm{P}(\mathrm{pub_i})$, along with the language model to calculate the recommendation ranking for a piece of text. Similarly idea has been studied by Lao and Cohen (2010).

For this method, all the publications and citations were treated equally on the graph. As already mentioned, however, this hypothesis is oversimplified (Liu et al., 2012), as some citations and publications are more important than others for some scientific topics in a citing paper.

2.3 Using publication topical prior with full-text data

Statistical citation relations are important but not necessarily accurate means of telling the importance of a publication, in that they ignore the semantic information of the citing/cited publications and the citation motivation itself. In this paper, we propose three hypotheses to enhance the citation relation based publication prior for citation recommendation:

1. The (citing/cited) publication topical prior (a.k.a. publication topical importance or bias) as compensate of the publication prior should provide more accurate prior information.

2. Scientific publication (vertex) topic distribution on the citation graph is an important indicator of the publication topical prior, i.e., the publication topical importance.

3. Citation (edge) topic distribution, a.k.a. transition topic probability distribution, extracted from the citation context in the citing paper is an important indicator of both the citation topic importance and the publication topic importance, i.e., the publication topical prior.

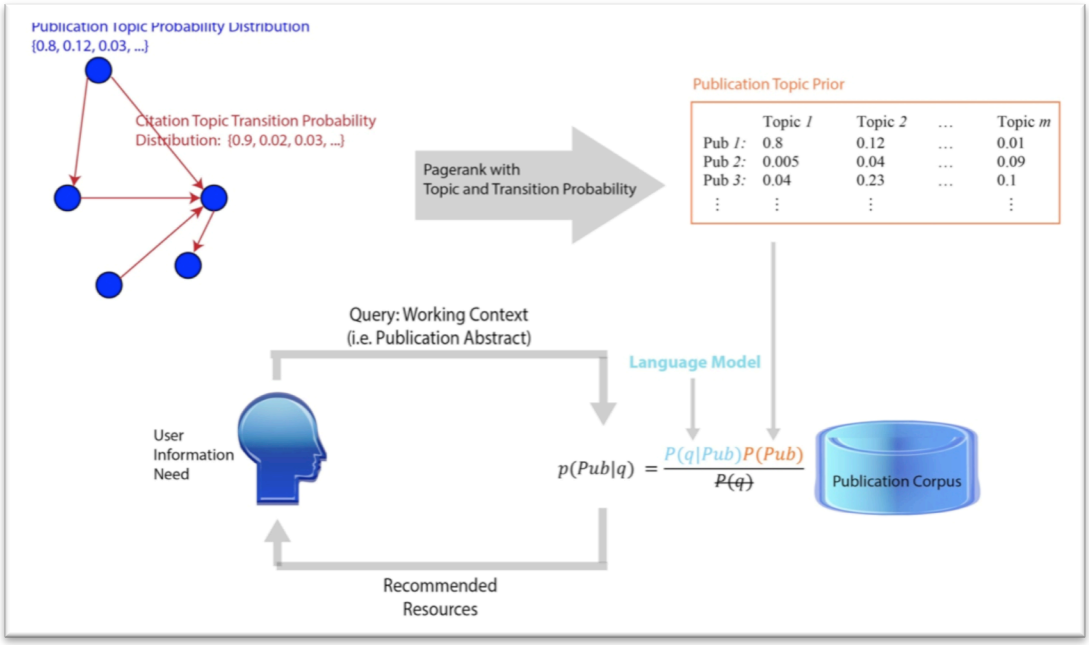

Figure 1. Citation recommendation with topic prior workflow

As Figure 1 shows, we first constructed a new citation graph associated with publication (vertex) and citation (edge) topic distributions, which were inferred from publication content and citation context. And then, we got the publication topic prior by PageRank with Topic and Transition Probability, as $\mathrm{P}(\mathrm{pub_i})$. When the user has information need,(query q), we recommended citations by calculating $\mathrm{P}(\mathrm{pub}|\mathrm{q})$, in this paper, query could be a working context, i.e., publication abstract. As shown in the Figure 1, $\mathrm{P}(\mathrm{pub}|\mathrm{q})$ is Language Model method, and $\mathrm{P}(\mathrm{pub_i})$ is the prior.

Please note, as we employ publication full-text data, multiple edges can exist between each citing and cited papers, where each edge is represented by a topic distribution (inferred from a citing context).

We then used the enhanced PageRank algorithm based on White and Smyth's work (2003) to calculate the publication topic prior. For different topics the publication importance (prior) can be different.

Finally, based on user textual input *T*, i.e., a publication abstract, we infer the user information need topic distribution, $\{T_{Z_1}, T_{Z_2} \ldots T_{Z_k}\}$ to represent the user's topical information need. The publication prior can then be calculated by:

$$P(pub_i) = P_{\theta_T}(pub_i) = \sum_{t=1}^{k} P(Z_t|T) \cdot P_{Z_t}(pub_i)$$

where $P_{Z_t}(pub_i)$ is the publication topic prior of $\mathrm{pub_i}$ for topic $\mathrm{Z_t}$. Unlike the classical (query-independent) publication prior probability, in this study, the publication prior depends on the user textual input T based topic distribution, θ_T. As a result, for each publication, as Figure 1 shows, a topic prior vector is used to characterize the importance of this publication for different scientific topics. When a user inputs a textual working context T, a topic modeling algorithm will infer the topic distribution θ_T, and calculate the publication topic prior $P_{\theta_T}(pub_i)$ by using topic probability $P(Z_t|T)$ and publication topic prior $P_{Z_t}(pub_i)$.

2.4 Topic modeling and citation topic inference

Latent Dirichlet Allocation (LDA) is a significant topic modeling proposed by Blei et al. (2003), allowing sets of observations to be explained by unobserved groups that explain why some parts of the data are similar. However, one limitation of LDA is the challenge of interpreting and evaluating topic statistics. In addition, arbitrary numbers of topic 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 semi-supervised topic modeling algorithm, labeled LDA (LLDA) (Ramage, Hall, Nallapati, & Manning, 2009). 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 $\beta_{key_{i,}}$ over all vocabulary words.

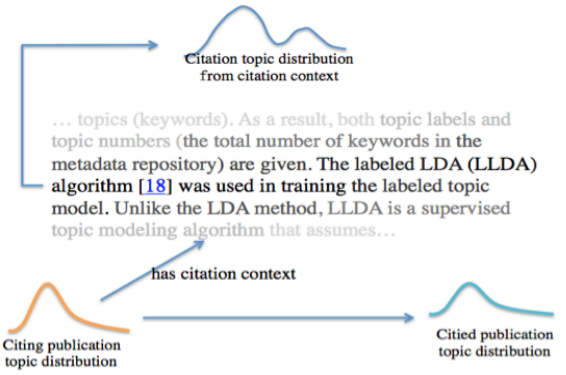

Figure 2. Publication and citation topic distributions

On the other hand, there is a common assumption in bibliometrics studies and citation recommendation approaches, that If $paper_1$ cites $paper_2$, then $paper_1$ and $paper_2$ are connected. Therefore, we created a large citation-directed network, $G = (V, E)$. While, classical citation networks tend to ignore citation and publication content. Here, we characterize citation relations in terms of two kinds of knowledge: publication (citing or cited paper) topic probability distribution, and citation topic probability distribution. These are illustrated in Figure 2.

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. Therefore, the citation-directed network in this paper with two kinds of prior knowledge: publication topic priors and a citation topic transitioning probability distribution.

Each vertex, $v \in V$, on the citation graph represents a publication, with the publication topic prior probability vector $\{p_{z_{key_1}}(v), p_{z_{key_2}}(v), \ldots p_{z_{key_n}}(v)\}$, where $p_{z_{key_t}}(v)$ is the prior probability of vertex v for topic z_{key_t} and $\sum_{i=1}^{n} p_{z_{key_i}}(v) = 1$.

$$p_{z_{key_t}}(v) = \frac{P(z_{key_t}|paper_v)}{\sum_{x=1}^{|V|} P(z_{key_t}|paper_x)}$$

Each edge, $e \in E$, on the graph represents a citation connecting v_i and v_j (v_i cites v_j). The topic transitioning vector for each edge is$\{ p_{z_{key_1}}(v_i|v_j), p_{z_{key_2}}(v_i|v_j), \ldots p_{z_{key_n}}(v_i|v_j) \}$, where $p_{z_{key_t}}(v_i|v_j)$ is the probability of transitioning from vertex v_i to v_j for topic z_{key_t}.

$$p_{z_{key_t}}(v_i|v_j) = \frac{P(z_{key_t}|citation_{j,i})}{\sum_{x=1}^{d_{out}(v_j)} P(z_{key_t}|citation_{j,x})}$$

where $P(z_{key_t}|paper_v)$ is the publication topic inference score, and $P(z_{key_t}|citation_{j,i})$ is the citation topic inference score.

For any text window surrounding the target citation, as Figure 3 shows, the proximity of the word (to the citation) is intuitively important for inferring the citation topic distribution. More specifically, as Figure 3 shows, words closer (to the citation) can provide more significant information as compared with words further away from the target citation. In this research, we used a decay function to characterize the proximity. For text before or after the target citation, we used to following formula to infer the topic distribution by using the words before $\{w_{-1}, w_{-2} \dots w_{-m}\}$ and after the citation $\{w_1, w_2 \dots w_m\}$, where the subscript is the distance from the segment to the target citation.

$$P\left(z_{key_i}\middle|citation\right) = \frac{\sum_{j=1}^{m} P\left(z_{key_i}\middle|w_{-j}, w_j\right) \cdot (log(1+j))^{-1}}{\sum_{j=1}^{m} (log(1+j))^{-1}}$$

The citation topic distribution is decided by (normalized) word distance (to the citation) before or after the citation, and the contribution of word wj decays when the distance of this word, j, gets larger, i.e., $(log\ (1+j))^{-1}$. In this research, based on findings of Ritchie, Robertson, & Teufel (2008), we used preliminary parameter setting with context length m = 150 words, but more parameter training should be validated in the future. Compared with Ritchie et al. (2008), context size is less important for this study as the closer words always made more contribution to the citation topic distribution.

2.5 Publications & Citation Topic Inference

In this study, we used the following approaches to infer the publication and citation topic probability distributions:

Publication topic inference with all topics (ALL): As the easiest approach, we assumed that all publications in the repository were related to all possible topics extracted by LLDA. So we used publication {title + abstract + full} text to infer the topic distribution on any α_i in the topic space. For this approach, author keyword metadata was not used for topic inference.

Publication topic inference with keyword greedy matching (KGM): As we used supervised topic modeling algorithm, LLDA, for this study, each author-contributed keyword is used as the topic label, and we don't need to set up the total topic number. One limitation of this approach, however, is that a large number of publications in the corpus don't have keyword metadata. In order to solve this problem, we used greedy matching to generate pseudo-keywords for each paper, which has been used in Guo, Zhang, and Liu's work (2013).

First, we loaded all possible keyword (topic label) strings into memory. We then searched each keyword from the paper title and abstract by using greedy match. For example, if "music information retrieval" existed in the title, we didn't 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.

Publication topic inference with keyword greedy matching + smoothing (KGMS): One limitation of keyword-based topic modeling is that each publication consists of only a few topics, and all other topic probabilities are 0. This may limit citation recommendation performance because, for instance, authors may not assign enough keywords (topics) for the target publication. Consequently, we applied smoothing techniques, where:

$$p_{z_{key_t}}(paper_v) = \sigma \cdot P\left(z_{key_t}\middle|paper_v\right) + (1-\sigma) \cdot P\left(z_{key_t}\middle|Corpus\right)$$

where the paper topic probability is calculated by a linear smoothing function. As $P\left(z_{key_t}\middle|Corpus\right)$ is always larger than 0, the paper topic score is also always positive. The parameter, σ, controls the amount of smoothing. In this research, we used as a tentative value, $\sigma = 0.8$.

Publication topic inference with all topics + smoothing (ALLS): Similar as KGMS, we used publication {title + abstract + full} text to infer the topic distribution, and then, used $P\left(z_{key_t}\middle|Corpus\right)$ for smoothing. The linear smoothing function parameter $\sigma = 0.8$.

Citation topic inference without keyword information (ALL): As with publication topic inference, we assumed that all citations in the repository are related to all possible topics extracted by LLDA. For this approach, we didn't use keyword information from citing and cited publications.

Citation topic inference with citing and cited publication keywords (CC): For this approach, we assumed that citations may not relate to all topics in the LLDA model. Instead, citations may only relate to topics provided by citing or cited topics. For any topic, z_{key_x}, not in a citing or cited paper, we gave the citation a lower score, $P'(z_{key_x}|citation) = \psi \cdot P(z_{key_x}|citation)$. We set $\psi = 0.1$ for this research, as we didn't want to totally remove these citations in the graph or make the citation transitioning probability = 0. As with publication topic inference, citation distributions for this method were normalized.

Based on the above methods, topic distributions for each publication could be sparse for the KGM assumption, and for a given topic, z_{key_t}, the vertex prior probability, $p_{z_{key_t}}(v)$, 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 $r \in R$, $p_{r,z_{key_t}} \neq 0$. Those root vertices can be thought of as the important publications given a topic (prior knowledge). A special case is the "All topics" ALL approach or KGMS approach, where all the topics are considered, and all are root vertices, $R = V$.

We used the PageRank with priors algorithm (Shi, Leskovec, & McFarland, 2010; Haveliwala, 2003) to calculate each vertex's (topic relative) importance, $I_{key_t}(v|R) = \pi_{key_t}(v)$, and:

$$\pi_{key_t}(v)^{i+1} = (1-\beta_b)\left(\sum_{u=1}^{d_{in}(v)} p_{z_{key_t}}(v|u)\,\pi_{key_t}{}^{i+1}(u)\right) + \beta_b p_{z_{key_t}}(v)$$

This equation represents a Markov chain for a random surfer who transitions "back" to the root vertexes *R* with probability β_b at each time-step. For each incoming link (citation) from v the PageRank score is updated with respect to edge (citation) transitioning probability $p_{z_{key_t}}(v|u)$.

The output, for each vertex (publication), v, is an authority vector $\{A_{z_{key_1}}(v), A_{z_{key_2}}(v), \ldots A_{z_{key_n}}(v)\}$ (as Figure 1 shows). 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.

2.6 Evaluation Methods

In this study, we recommend citations based on a textual working context. In order to evaluate this work, we randomly sampled a number of publications with full text. We used the publication abstract as the working context input (to represent the user's information need), and then extracted all the citations and citation frequency from the publication text by using regular expressions.

We used two indicators to measure recommendation ranking algorithm performance: *mean average precision* (MAP), and *normalized discounted cumulative gain* (nDCG) (Järvelin & Kekäläinen, 2002). nDCG estimates the cumulative relevance gain a user receives by examining recommendation results up to a given rank on the list. In this research, we used an importance score, 0–4, as the citation importance to calculate nDCG scores. For instance, if a citation is not cited by the target publication, the importance score is 0, and if a citation is cited 4 or more times in the citing paper, then it is probably very important for the target citing publication, and its importance score is 4.

We clearly understand that author-provided citations don't cover all important publications that should be cited for the working context (publication abstract). The goal of this evaluation, however, is to compare citation recommendation performance with different publication priors.

3 Previous Research

3.1 Citation relationship and text

As aforementioned, most previous studies (Liu et al., 2012) in text mining, bibliometrics, and scholar information retrieval/recommendation used citation as a statistical relation between citing and cited papers, while the topic and motivation is ignored.

With further study of citation analysis, increasing numbers of researchers have come to doubt and challenge the reasonableness of assuming that the raw citations reflects an article's influence. For instance, CiteRank (Walker et al., 2007) is an enhanced ranking algorithm over PageRank, which enables ranking method to estimate the traffic $\mathrm{T_i}(\tau_{\mathrm{dir}}, \alpha)$ to a given paper i. For this method, a recent paper is more likely to be selected with a probability that is exponentially discounted according to the age of the paper, τ_{dir}. At every step of the path, with probability α the researcher is satisfied/saturated and halts his/her line of inquiry. Citation Influence Model (Dietz, Bickel, & Scheffer, 2007) is another effective method to weight the importance of citation relation, which employed citing and cited paper topic

distribution and the compatibility-based citation weighting of two topic mixtures is measured by the Jensen-Shannon Divergence. Similarly research is implemented by Erosheva, Fienberg, and Lafferty (2004), which capturing the notion of topical similarity between the contents of the cited and citing documents. Based on these work, Nallapati, Ahmed, Xing, & Cohen (2008) proposed Pairwise-Link-LDA and Link-PLSA-LDA, which goal to predict important unseen citation between papers by using topic based graph models. Vice versa, citation relation can also be used to characterize the topic models. For instance, He et al. (2009) used citation relation to detect the topic evolution by using Inheritance Topic Model. Similar studies, i.e., Topic-Link LDA (Liu, Niculescu-Mizil, & Gryc, 2009) and Topic-level Influence (Liu, Tang, Han, Jiang, & Yang, 2010), investigated topic level propagation and aggregation.

Unlike those studies, we employed citation context along with citation topology to estimate topic based citation motivation, while we assume full-text analysis has to some extent compensated for the weaknesses of citation counts and has offered new opportunities for citation analysis. Moreover, the citation graph with supervised topic analysis is converted to publication topical prior for language model, which is used to address user textual information need. Ritchie et al. (2008) and Bernstam et al. (2006), found citation context can provide important information for retrieval task. They also found that the closeness of a word in the citation context provides stronger semantic information about the cited paper. Meanwhile, Gerrish, and Blei (2010) used dynamic influence model to characterize scholar impact without using citation information. These studies motivate us to use the proximity for citation topic inference at the topic level for recommendation task.

3.2 Citation recommendation

Scientific recommendation is an important research area, where a scientific publication, venue, or author is recommended to users based on the similarity between the recommended resource and user profiles or samples of text they are working on. Chandrasekaran et al. (2008), for example, present a method of recommending scientific papers of potential interest to users by using the ACM Computing Classification System along with hierarchical concept information from both author profiles and paper content. Based on this work, He et al. (2010) proposed a method to recommend global and local citations based on a piece of given text under both context-oblivious and context-aware conditions. In this article, the authors recommend citations to users based on the similarity between a candidate publication's in-link citation contexts and a user's input text. More recently, He et al. (2011) have used more comprehensive methods, i.e., the language model, contextual similarity, and the dependency feature model, to enhance citation recommendation performance. Unsupervised topic modeling is also used for citation analysis (Xia, Tang, & Moens, 2012), where visible candidate citations, hidden scientific topics, and visible words are represented in different layers. A restricted Boltzmann machine model was used for building the relationship between user input and recommended citation ranking. Similarly, recommendation methods can also be used for domain expert recommendation, for instance bag-of-words (Basu, Hirsh, & Cohen, 2002), LSI concept (Dumais & Nielsen, 1992) and subject category (Conry, Koren, & Ramakrishnan, 2009) similarity is used for expert recommendation based on text input.

All of the above-mentioned recommendation studies involve item–item or item–user-based content-sensitive collaborative filtering algorithms. In these cases, potential scientific resources for recommendation should be similar to the target working context, and words or unsupervised latent topics were used to build the relationship between user information need and candidate resources.

As another important approach, scholarly or bibliographic networks—i.e., networks based on citation or co-authorship—have also been used to recommend scientific resources. For instance, Shi, Leskovec, and McFarland (2010) developed citation projection graphs by investigating citations among publications that a given paper cites. In this study, authors investigated high-impact and low-impact citation behavior, where citation impact is defined as the number of citations a publication receives normalized by the average number of citations of all other publications published in the same year and same area. More recently, Lao and Cohen (2010) used both supervised and unsupervised methods with the Random Walk with Restart (RWR) algorithm for citation, author, and venue recommendation. In this study, a large heterogeneous network (with venue, author, and publication as vertices, and co-author and citation as edges) was constructed for the recommendation task. Evaluation results show that supervised RWR can significantly enhance recommendation performance.

The proposed work differs from previous research in that we used similarity-based (i.e., the language model) and network-based (i.e., prior probability) methods for citation analysis. Moreover, in the citation network, we used supervised topic models to characterize each vertex and edge. The citation

topical motivation probability distribution is extracted by using the proximity-based citation context, where each topic is a keyword-labeled unigram word probability distribution.

4 Experiments

4.1 Data

We used 41,370 publications (as **candidate citation collection**) 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. From these we extracted 28,013 publication texts (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 keyword information from a metadata repository to represent the content of the paper. In order to get the citation context, we then wrote a list of regular expression rules to extract all the possible citations from paper's full text. For example, the rules could extract *"… [number]…"* and *"… [number, number…, number]…*" as citations from the content of a publication. Each citation extracted from the publication text was associated with a reference (cited paper ID). In a total of 223,810 references ($paper_1$ *cites* $paper_2$ 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.

For later citation recommendation evaluation, we also used a **test collection** with 274 papers. The selected papers met the following conditions. First, the selected papers were exclusive from the 41,370 publication candidate citation collection. Second, each selected paper had more than 15 citations from the candidate citation collection. Thirdly, each paper's abstract had at least 150 words. The paper's abstract was used as a working context to represent a user's information need, and we recommended citations from the candidate citation collection.

Besides that, we sampled 10,000 publications (with full text) to train the LLDA topic model. Author-provided keywords were used as topic labels. 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 less than 3, 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 3 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.

4.2 Experimental 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 vertex topic probabilities and each citation's transitioning probabilities. Please note that from each node there are different ways to compute publication and citation topic distributions. In this evaluation we investigated the following groups (defined in section 2.6): 1) **KGM+CC**: keyword greedy matching (publication) + citing and cited paper topics (citation); 2) **KGMS+CC**: keyword greedy matching with smoothing (publication) + citing and cited paper topics (citation); 3) **ALL+ALL**: all topics (publication) + all topics (citation); 4) **ALLS+ALL**: all topics and smoothing (publication) + all topics (citation).

All four methods generated publication topic prior distributions. We used these priors along with the language model for recommendation ranking. For comparison, we used two baseline algorithms: 1) **LM**: the language model without priors; and 2) **LM + PageRank**: the language model with PageRank priors. For all applications of the language model in this study we used the Dirichlet smoothing technique (Erosheva et al., 2004). MAP and nDCG results are presented in Tables 2.

	LM	LM+ PageRank	KGM + CC	KGMS+ CC	ALL + ALL	ALLS+ ALL
MAP@all	**0.1211**	**0.1226**	**0.1218**	**0.1601**	**0.1536**	**0.1641**
nDCG@10	0.1183	0.1740	0.1712	0.2015	0.2032	0.2137
nDCG@30	0.1424	0.1951	0.1929	0.2281	0.2317	0.2411
nDCG@50	0.1586	0.2091	0.2088	0.2447	0.2460	0.2563
nDCG@100	0.1774	0.2297	0.2290	0.2670	0.2648	0.2756
nDCG@300	0.2022	0.2539	0.2520	0.2897	0.2907	0.3035
nDCG@500	0.2100	0.2615	0.2607	0.2989	0.3007	0.3109
nDCG@1000	0.2202	0.2711	0.2704	0.3078	0.3108	0.3208
nDCG@3000	0.2326	0.2847	0.2832	0.3199	0.3241	0.3335
nDCG@5000	0.2375	0.2890	0.2886	0.3236	0.3290	0.3374

nDCG@all	**0.2647**	**0.3157**	**0.3148**	**0.3445**	**0.3470**	**0.3550**

Table 2. Different publication and citation inference methods

In terms of result comparison, we found that the recommendation ranking performance of the topic prior based algorithms significantly ($p < 0.001$) outperformed both the language model (LM) and language model + PageRank (LM + PageRank) approaches, except for KGM+CC. The keyword greedy matching approach, KGM+CC, was just a little worse than the PageRank approach, but not significantly so, which is highly likely *because of the zero probabilities in the publication topic distribution*, A.K.A., the (author-provided) keywords cannot fully cover the topics of the paper or citation. We also found the purely content-based algorithm based on the language model to be the worst approach in this experiment, which substantiates our hypothesis that centrality-based network analysis (like PageRank), as prior, can significantly enhance content-based recommendation performance. In addition, publication (vertex) and citation (edge) topic characterization can also significantly enhance recommendation performance as compared with the classical PageRank algorithm, and appropriate smoothing techniques are important to improve recommendation performance. In particular, ALLS + ALL topic inference is better than all other approaches in this evaluation.

MAP for this experiment tells whether the recommended citation is correct or not. nDCG@n, in this evaluation, is a more important indicator, for it tells the degree of citation importance. If the nDCG score is large, the target algorithm can prioritize the most important candidate citations on the ranking list. In Table 2, it's clear that topic priors is always better than PageRank + language model and all other baseline methods, especially for nDCG performance. It is clear that publication topical prior, based on publication and citation distributions + citation relations, outperforms classical topic-independent publication prior.

5 Conclusion

In this study, from a language model perspective, we have enhanced classical content similarity-based citation recommendation by adding different kinds of publication prior probabilities. Initially, citation relationship based on PageRank is used to generate publication priors. We then propose a more sophisticated method of characterizing each publication (vertex) and citation (edge) in the citation network by utilizing a supervised topic modeling algorithm, where each topic is labeled by an author-provided keyword. This citation network is then used to generate a publication topic prior vector. Further, for each candidate citation, we calculate a dynamic citation prior, $P_{\theta_T}(pub_i)$, by using the publication topic prior vector and a user's working context topic distribution. By using supervised topic modeling, we can find out the topic number, k, and, later in the future, the topical importance can be used for topic-based citation recommendation, where the topic label is given to facilitate user interpretation.

Based on MAP and nDCG@n evaluation, we find that all kinds of publication priors can significantly improve the recommendation performance comparing with content based language model. However, simply using keywords with greedy matching (publication) + citing and cited papers' topics (citation) cannot surpass classical topic-independent PageRank. One reason is that greedy matching (pseudo-keyword) and keyword metadata can hardly cover all the topics of the target publication. Another reason is that citations important for a topic are not necessarily related to it. For example, some natural language processing studies are important for information retrieval topic.

Both topic inference and smoothing techniques can significantly enhance citation recommendation performance, because all publication and citation topic probabilities are non- zero. When we integrate these two methods together (ALLS + ALL, all topic inference plus smoothing), recommendation performance outperforms all other methods.

The limitation of this work is mainly from the test corpus. We cannot access full-text data for all papers and only extracted 67.7% of the papers' full text. When full text was unavailable, we used the title and abstract as a compromise, but this can be biased. This problem could be fixed by using image-based text recognition in the future.

6 References

Basu, C., Hirsh, H., Cohen, W. W., & Nevill-Manning, C. (2002). Technical paper recommendation: A study in combining multiple information sources.*Journal of Artificial Intelligence Research*, *14*, 241-262.

Bernstam, E. V., Herskovic, J. R., Aphinyanaphongs, Y., Aliferis, C. F., Sriram, M. G., & Hersh, W. R. (2006). Using citation data to improve retrieval from MEDLINE. *Journal of the American Medical Informatics Association*,*13*(1), 96-105.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993-1022.

Chandrasekaran, K., Gauch, S., Lakkaraju, P., & Luong, H. P. (2008, July). Concept-based document recommendations for citeseer authors. In *Adaptive hypermedia and adaptive web-based systems* (pp. 83-92). Springer Berlin Heidelberg.

Conry, D., Koren, Y., & Ramakrishnan, N. (2009, October). Recommender systems for the conference paper assignment problem. *Proceedings of the third ACM conference on Recommender systems*, 357-360.

Dumais, S. T., & Nielsen, J. (1992, June). Automating the assignment of submitted manuscripts to reviewers. *Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, 233-244.

Dietz, L., Bickel, S., & Scheffer, T. (2007, June). Unsupervised prediction of citation influences. *Proceedings of the 24th international conference on Machine learning* , 233-240.

Erosheva, E., Fienberg, S., & Lafferty, J. (2004). Mixed-membership models of scientific publications. *Proceedings of the National Academy of Sciences*,*101*(suppl 1), 5220-5227.

Gerrish, S., & Blei, D. M. (2010, June). A Language-based Approach to Measuring Scholarly Impact. In *ICML*, 375-382.

Guo, C., Zhang, J., & Liu, X. (2013). Scientific metadata quality enhancement for scholarly publications.

Haveliwala, T. H. (2003). Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. *Knowledge and Data Engineering, 15*(4), 784-796.

He, Q., Pei, J., Kifer, D., Mitra, P., & Giles, L. (2010, April). Context-aware citation recommendation. *Proceedings of the 19th international conference on World wide web*, 421-430.

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

He, Q., Kifer, D., Pei, J., Mitra, P., & Giles, C. L. (2011, February). Citation recommendation without author supervision. *Proceedings of the fourth ACM international conference on Web search and data mining*,. 755-764.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems (TOIS)*, *20*(4), 422-446.

Kraaij, W., Westerveld, T., & Hiemstra, D. (2002, August). The importance of prior probabilities for entry page search. *Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, 27-34)

Lao, N., & Cohen, W. W. (2010). Relational retrieval using a combination of path-constrained random walks. *Machine learning*, *81*(1), 53-67.

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

Liu, X. (2013). Generating metadata for cyberlearning resources through information retrieval and meta-search. *Journal of the American Society for Information Science and Technology*, *64*(4), 771-786.

Liu, X., Zhang, J., & Guo, C. (2012, October). Full-text citation analysis: enhancing bibliometric and scientific publication ranking. *Proceedings of the 21st ACM international conference on Information and knowledge management*, 1975-1979.

Liu, X., Zhang, J., & Guo, C. (2013). Full-text citation analysis: A new method to enhance scholarly networks. *Journal of the American Society for Information Science and Technology*, *64*(9), 1852-1863.

Liu, X., Yu, Y., Guo, C., Sun, Y., & Gao, L. (2014, September). Full-text based context-rich heterogeneous network mining approach for citation recommendation. *Proceedings of the 14th ACM/IEEE-CS Joint Conference on Digital Libraries*, 361-370.

Liu, Y., Niculescu-Mizil, A., & Gryc, W. (2009, June). Topic-link LDA: joint models of topic and author community. *Proceedings of the 26th annual international conference on machine learning* , 665-672.

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

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The PageRank citation ranking: bringing order to the web.

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009, August). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1-Volume 1*, 248-256.

Ritchie, A., Robertson, S., & Teufel, S. (2008, October). Comparing citation contexts for information retrieval. *Proceedings of the 17th ACM conference on Information and knowledge management*, 213-222.

Shi, X., Leskovec, J., & McFarland, D. A. (2010, June). Citing for high impact. *Proceedings of the 10th annual joint conference on Digital libraries*, 49-58.

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

White, S., & Smyth, P. (2003, August). Algorithms for estimating relative importance in networks. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining* , 266-275.

Xia, H., Li, J., Tang, J., & Moens, M. F. (2012, April). Plink-lda: Using link as prior information in topic modeling. *Database systems for advanced applications* , 213-227.

Zhai, C., & Lafferty, J. (2001, September). A study of smoothing methods for language models applied to ad hoc information retrieval. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, 334-342.