Full-Text Citation Analysis: Enhancing Bibliometric and Scientific Publication Ranking

Xiaozhong Liu
School of Library and Information Science
Indiana University Bloomington
liu237@indiana.edu

Jinsong Zhang
Dalian Maritime University
Dalian, China
zjs.dlmu@gmail.com

Chun Guo
School of Library and Information Science,
Indiana University Bloomington
chunguo@indiana.edu

ABSTRACT

The goal of this paper is to use innovative text and graph mining algorithms along with full-text citation analysis and topic modeling to enhance classical bibliometric analysis and publication 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. Based on 104 topics (labeled with keywords) and their review papers, the cited publications of each review paper are assumed as "important publications" for ranking evaluation. The result shows that full text citation and publication content prior topic distribution along with the PageRank algorithm can significantly enhance bibliometric analysis and scientific publication ranking performance for academic IR system.

Categories and Subject Descriptors

H.3.3 [[Information Storage and Retrieval]]: Information Search and Retrieval

General Terms

Algorithms, Experimentation, Measurement

Keywords

Bibliometrics, Publication Ranking, Citation Analysis, Topic Modeling, PageRank, Prior Knowledge

1. INTRODUCTION AND MOTIVATION

Bibliometrics is a set of methods to quantitatively analyze the relatedness of scientific publications [3], i.e. scholarly networks, publication or venue importance, and co-authorship analysis. Citation analysis along with graph mining is a commonly used bibliometric method. In most previous works, while various methods were used to characterize the citation network, the basic assumption was easy and straightforward: either *Publication1 cites Publication2*, or *Author1 cites Author2*, regardless of sentiment, reason, topic, or motivation. More recent studies have shown, however, that this assumption is oversimplified. The reason or motivation for a citation matters. Taking the SDR model [9] as an example, it proposed the Structural Descriptive Referential model to capture the structural knowledge of citation, i.e., research question or methodology. However, most of these studies stay 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.

The combination of citation bibliometrics and text mining provides a synergy unavailable with each approach taken independently [7]. In our research, instead of classifying citations, we used a supervised topic modeling algorithm, Labeled LDA (LLDA), 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 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, the enhanced PageRank can generate an authority vector, and each score in the vector tells the publication's topical importance.

In the remainder of this paper, we: 1) introduce our novel methods for constructing a bibliometric citation graph with prior distributions via full-text topic modeling, 2) review relevant literature and methodology for bibliometric analysis, topic modeling, and network mining, 3) describe the experiment setting and evaluation results, and 4) discuss the findings and limitations of the study and identify subsequent research steps.

2. RESEARCH METHODS

Most previous bibliometrics studies share a common assumption: if $paper_1$ cites $paper_2$, then $paper_1$ and $paper_2$ 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

CIKM'12, October 29–November 2, 2012, Maui, HI, USA.
Copyright 2012 ACM 978-1-4503-1156-4/12/10...$15.00.

paper) topic probability distribution, and citation topic probability distribution. 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 to different scientific topics. In addition, topic authorities can be non-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 but not least, because we utilize full text citation analysis, one paper can have more than one citation edge with the other paper. For instance, if $paper_1$ cites $paper_2$ three times, there will be three distinct edges on the citation network between these two papers. Hence, the accumulated transitioning probabilities between $paper_1$ and $paper_2$ can be higher than others, resulting in more accurate PageRank random walk modeling.

2.1 Topic Modeling with Labels

Blei et al., [1] proposed Latent Dirichlet Allocation (LDA) as a promising unsupervised topic modeling algorithm. LDA employs a generative probabilistic model in the hierarchical Bayesian framework. 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 $\sum \alpha_j = 1$, and $\alpha_1, \alpha_2 \ldots, \alpha_T$ 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 assign a label to each statistical topic automatically. 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, 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 labeled LDA (LLDA) algorithm [11] 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 $\beta_{key_{i,}}$, over all vocabulary words. 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 θ_{paper,key_i} and the label's preference for the word $\beta_{key_i,w}$. Figure 1 visualizes the LLDA generative process. For each topic (keyword) key_k, one draws a multinomial distribution β_{key_k} from the symmetric Dirichlet prior η. Then, for each publication, one builds a label set Λ paper for the deterministic prior Φ. Finally, one selects a multinomial distribution θ_{paper} over the labels Λ paper from Dirichlet prior α.

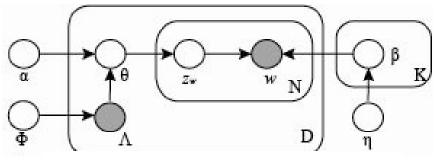

Figure 1. LLDA Algorithm

2.2 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 was missing or authors didn't provide any), and 6.13% had only 1 or 2 keywords, which is probably not enough to cover the whole paper.

To cope with this problem, we used greedy match, where we assumed that author-assigned keywords were not enough to cover the semantics of the paper, to expand the paper topic space. First, we loaded all possible keyword (topic label) strings into memory, and we 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 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. All topics not in this collection were ignored.

For this approach, a subset of keywords (topics) from the training LLDA model was used to infer the paper topic distribution. The topic scores for key_i, i.e., $P(z_{key_i}|paper)$, were normalized for future experiments, where $\sum_{i=1}^{n} P(z_{key_i}|paper) = 1$.

2.3 Citation Topic Inference

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 words, +n 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 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 in the citation network. As with publication topic inference, citation distributions for this method were normalized.

2.4 Citation Network with 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 $\{p_{z_{key_1}}(v), p_{z_{key_2}}(v), \ldots p_{z_{key_n}}(v)\}$, where $p_{z_{key_t}}(v)$ is the prior probability of vertex v for topic z_{key_t} and $\sum_{i=1}^{|V|} p_{z_{key_i}}(v) = 1$. 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}.

For a given publication u, we used $S_{in}(u)$ and $S_{out}(u)$ to represent a set of incoming and outgoing edges (citations) to node u, with "in" degree $d_{in}(u) = |S_{in}(u)|$ and "out" degree $d_{out}(u) = |S_{out}(u)|$. Thus, $\sum_{i=1}^{d_{out}(v_j)} p_{z_{key_t}}(v_i|v_j) = 1$. For example, if a publication cites only 3 papers and, for a specific topic, the transitioning probabilities to these 3 papers are 0.1, 0.1, and 0.8, 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:

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

and each edge's (i.e., each citation's) transitioning probability:

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

where $P(z_{key_t}|paper_v)$ is the publication topic inference score from section 2.2, and $P(z_{key_t}|citation_{j,i})$ is the citation topic inference score from section 2.3.

Unlike classical PageRank, a citation graph with vertex and edge priors permits non-uniformly-distributed random jumps. Based on section 2.1, topic distributions for each publication could be sparse for the greedy match assumption, and for a given topic, z_{key_t}, the vertex prior probability, $p_{v,z_{key_t}}$, 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" approach, where all the topics are considered, and root vertices $R = V$. We used the PageRank with prior algorithm [13] 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 $\{A_{z_{key_1}}(v), A_{z_{key_2}}(v), \ldots A_{z_{key_n}}(v)\}$. Each authority score in the vector indicates the publication topic importance with respect to both paper topic and full text citation priors. And we can get *n* ranking lists as a result.

2.5 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 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 keyword. In order 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 only focused on one topic (keyword). We assume that if a publication is cited by a review paper, and if this review paper concentrates on keyword key_i, then this publication is important for key_i. Since 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 key_i cited $paper_1$ twice and $paper_2$ once, then, $Importance_{key_i}(paper_1) = 2$ and $Importance_{key_i}(paper_2) = 1$. 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 0. We also assume that if a paper is cited 4 or more times by the review paper, then its importance is equal to 4.

The goal of this evaluation is to compare the performance of our approach against the 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. ***TFIDF/BM25/Language Model***: For these methods 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 TFIDF + 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. ***Language Model + PageRank***: For this baseline we combined the language model with PageRank (without topic priors), using random walk probability as the model prior.

We used two indicators to measure ranking algorithm performance: Mean Average Precision (MAP), and normalized Discounted Cumulative Gain (nDCG) [6]. nDCG estimates the 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.

3. PREVIOUS RESEARCH

In this section, we survey existing studies focusing on two fields: PageRank analysis for citation network and bibliometrics for scientific publications.

Drawing on classic bibliometrics papers, many scholars have focused their research on citation frequency and citation impact and applied it in different domains. Harhoff, Narin, Scherer and

Vopel [4] judged the value of patented inventions by citation frequency and concluded, "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 articles and various characteristics of journals, articles, and authors [8] and concluded that annual citation rates of ecological papers are affected by many factors, including the hypothesis tested, article length, and authors' information.

Traditionally, citation analysis treats all citations equally. However, in reality, not all citations are equal. Some scholars consider location to be a factor affecting the relative importance of a citation. Herlach [5] 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 others that have been mentioned only once. The stylistic aspects of a citation also matter. Bonzi [2] distinguished between a number of broad categories of citations, i.e., those not specifically mentioned in the text and those barely mentioned direct quotation.

PageRank [10] has become a significant method for evaluating the most important nodes in complex graphs analysis. From the point of citation analysis in bibiometrics, PageRank is also an efficient way to evaluate a paper's ranking score in a specific domain. White and Smyth [13] first proposed the priors idea in their formalization of a relative-rank extension to both PageRank and HITS. They experimentally evaluated different properties of some algorithms on toy graphs and demonstrated how the approach could be used to study relative importance in real-world networks. Rodriguez and Bollen [12] 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.

4. EXPERIMENTS

4.1 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 and abstract 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. In a total of 223,810 references, we successfully identified 94,051 references, which accounted for 42.0% of all references.

Then, we sampled 10,000 publications (with full text) to train the LLDA topic model. Author-provided keywords were used as topic labels. For instance, this paper has 6 author provided keywords. Thus, our LLDA training would have assumed that this paper is a multinomial distribution over these 6 topics.

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 < 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 a 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 in section 2, 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.

We then used our approach to compare with other baseline methods, including PageRank, TFIDF, BM25, language model, and PageRank + Language Model. The results are presented in the following tables. The best performing algorithm is highlighted for each row, and these results are visualized in Figures 2 and 3.

Clearly, for baseline ranking methods, PageRank + language model achieved the best performance, and PageRank alone (topic independent) performed the worst. For MAP@n, PageRank + language model was better than our method (PageRank with prior, highlighted) when n ≤ 50. But for n ≥ 100, PageRank with priors was better than all other methods. We also used significant testing to compare PageRank with prior and PageRank + language model (*t < 0.01, **t < 0.005, ***t < 0.001). After n ≥ 1000, PageRank with priors is significantly better than all other baseline methods.

nDCG@n is a more important indicator in this research, for it tells the degree of (publication topic) importance. If nDCG score is large, the target algorithm can prioritize the most important on the ranking list. In Table 2 and Figure 3, it's clear that PageRank with priors is always better than PageRank + language model and all other baseline methods. After n ≥ 10, the results are significant.

Table 1: Different publication and citation inference methods (MAP)

	PageRank	TFIDF	BM25	Language Model	LM + PageRank	**PageRank with prior**
MAP@10	0.0168	0.1551	0.1637	0.163	**0.2039**	0.1955
MAP@30	0.0192	0.1387	0.1397	0.1498	**0.1872**	0.1728
MAP@50	0.0186	0.1295	0.1254	0.138	**0.1702**	0.1581
MAP@100	0.0182	0.1171	0.1151	0.1198	0.1424	**0.144**
MAP@300	0.0162	0.0918	0.0904	0.0935	0.1106	**0.1207**
MAP@500	0.0145	0.0858	0.0851	0.0864	0.1001	**0.1144**
MAP@1000	0.011	0.0754	0.0756	0.0759	0.0918	**0.1064***
MAP@3000	0.0072	0.064	0.0652	0.0672	0.081	**0.1011***
MAP@5000	0.006	0.0614	0.0626	0.0646	0.078	**0.1004***
MAP@ALL	0.0037	0.0415	0.0418	0.0438	0.0542	**0.0816****

Table 2: Different publication and citation inference methods (nDCG)

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

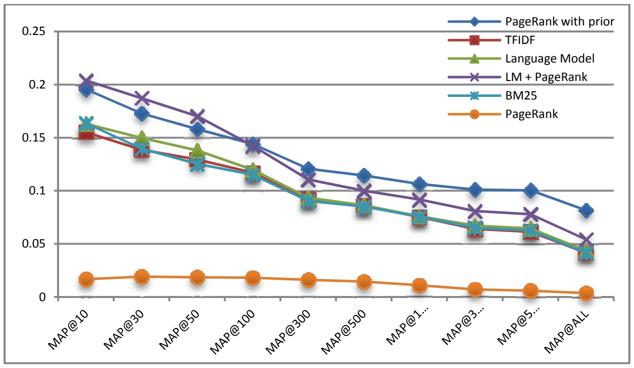

Figure 2: Different publication and citation inference methods (MAP)

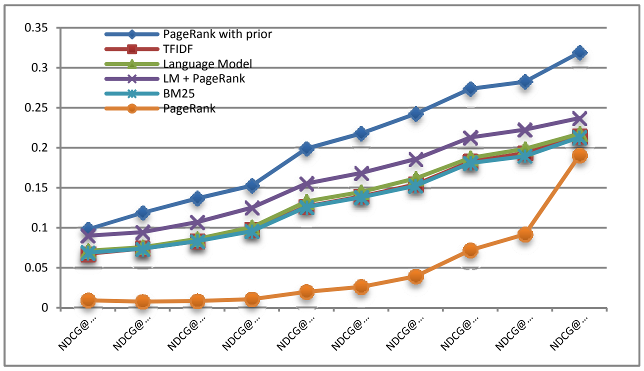

Figure 3: Different publication and citation inference methods (nDCG)

5. ANALYSIS AND CONCLUSION

Based on the MAP@n and nDCG@n evaluation, we find that PageRank with publication priors and citation transitioning probability distributions extracted from full-text data can produce reliable, high quality topic ranking results, which significantly outperform a list of baseline algorithms. Meanwhile, as the topics extracted from publications are labeled with author provided keywords, the result (publication topic importance) is interpretable, which is important for bibliometrics analysis.

Another interesting finding of this research is that considering full text publication and citation transitioning probabilities for bibliometric analysis can help us to find the most significant publications for each topic. This new method may favor publications that make significant contributions but which have not yet received many citations.

6. LIMITATIONS AND FUTURE WORK

The 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 only extracted 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 in section 4, 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. Another problem is that we only identified 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 existing well-established bibliometric algorithms, like 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 plug our method into other bibliometric methods for better scientific publication, author, and venue characterization, e.g., by introducing topical h-index or topical impact factors.

7. REFERENCES

[1] Blei, D.M., Ng, A.Y. and Jordan, M.I. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*. 3, 993-1022.

[2] Bonzi, S. 1982. Characteristics of a literature as predictors of relatedness between cited and citing works. *Journal of the American Society for Information Science*. 33, 4, 208-216.

[3] De Bellis, N. 2009. Bibliometrics and citation analysis: from the Science citation index to cybermetrics. *Scarecrow Press*.

[4] Harhoff, D., Narin, F., Scherer, F.M. and Vopel, K. 1999. Citation frequency and the value of patented inventions. *Review of Economics and statistics*. 81, 3, 511-515.

[5] Herlach, G. 1978. Can retrieval of information from citation indexes be simplified? Multiple mention of a reference as a characteristic of the link between cited and citing article. *Journal of the American Society for Information Science*. 29, 6, 308-310.

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

[7] Kostoff, R.N., del Rio, J.A., Humenik, J.A., Garcia, E.O. and Ramirez, A.M. 2001. Citation mining: Integrating text mining and bibliometrics for research user profiling. *Journal of the American Society for Information Science and Technology*. 52, 13, 1148-1156.

[8] Leimu, R. and Koricheva, J. 2005. What determines the citation frequency of ecological papers? *Trends in Ecology & Evolution*. 20, 1, 28-32.

[9] Liu, X., Qin, J. and Chen, M., 2011. ScholarWiki system for knowledge indexing and retrieval. In *Proceedings of the American Society for Information Science and Technology*. 1-4.

[10] Page, L., Brin, S., Motwani, R. and Winograd, T. 1999. The PageRank citation ranking: Bringing order to the web. Technical Report. Stanford InfoLab

[11] Ramage, D., Hall, D., Nallapati, R. and Manning, C.D., 2009. Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *EMNLP '09 Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics*, 248-256.

[12] Rodriguez, M.A. and Bollen, J. 2006. Simulating network influence algorithms using particle-swarms: Pagerank and pagerank-priors.

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