Citation Analysis and Keyword Mining based on Fulltext Extraction of Scientific Literature

Jinsong Zhang
College of Transportation Management
Dalian Maritime University
Dalian, China. 116026
zhangjinsong85@163.com

ABSTRACT
Citation analysis as a meaningful research tool has been studied for a long time for domain information visualization, information retrieval, and bibliometric analysis. This paper proposes three steps of mining keyword relationships using citation graph analysis based on the fulltext of scientific literature in the scientific publication database. First, the method Citation Probability Distribution Distance (CPDD) was proposed to generate domain knowledge graphs based on domain and domain context. We then introduce three rules to merge them into a new graph to improve the performance. Secondly, we use a topic modeling method (Labeled LDA) to improve CPDD to avoid the significant hypothesis in the aforementioned method by analyzing the distribution of citation over keywords. In this way, we can find the topic distribution for each citation and establish the keyword relationship graph by citations. Last but not least, we will use optimized PageRank algorithm to evaluate the ranking results of the selected publications, not only taking into account the citation counts, but also considering the keyword relationships generated by citation analysis based on fulltext extraction.

Categories and Subject Descriptors
H.3.7 [**Digital Libraries**]: Digital Libraries

General Terms
Design; Experimentation.

Keywords
Citation Analysis; CPDD; LLDA Model; Keywords Relationship; Fulltext Extraction; PageRank.

1. INTRODUCTION
With abundant journals and conferences in scientific database, it is difficult at times to know which article is best, which represents the highest standards for research in the field. Bibliometric analysis can provide varies methods for analyzing the relationships among the publications, i.e., direct citation, bibliographic coupling, and co-authorship analysis. The researchers thought the citation analysis is a meaningful research tool. Because the graph generated by citation analysis can represent a domain knowledge graph modeling, which is one of the most important methods to describe the significant characters of a selected domain, as well as essential for domain information visualization, information retrieval, and bibliometric analysis. Therefore, the study of generating a domain knowledge graph and analyzing the interrelation of publications is a meaningful research direction.

JCDL'12, June 10–14, 2012, Washington, DC, USA.
ACM 978-1-4503-1154-0/12/06.

Keywords given by the authors can be seen as the topics or the concentrative summary of the publications, so the citation always occurs because some of topics between the papers are interrelated. In this paper, we will treat the keywords in the papers given by authors as the topics of the publications, i.e., the paper has three keywords and citing five papers, it can be regarded that each citation at least related to one or more topics. So, hopefully, generating a citation graph based on keywords relationship will improve the performance of citation analysis, as well as output a ranking result for the papers in a specified topic.

This dissertation research will study the questions as follows.

First of all, we will introduce a method, Citation Probability Distribution Distance (CPDD), used for generating two graphs: domain graph and domain citation graph. After that three rules was introduced to merge them into a new graph, which can enhance the performance.

In addition, the aforementioned CPDD should be conducted on one hypothesis that treated all keywords between the citing and cited papers are related. So in second part of this research, we will take into account the fulltext extraction, and use citation context to generate the keywords distribution by Labeled LDA (LLDA) [1] model, so the actual relationships between keywords can be found not only by citation analysis, but also by the similarity as citing cause. Therefore, the graph will be re-generated by the principle: the keywords between two papers are related only when they are linked at least by one citation, as well as the similarity of these two keywords is higher than others.

Last but not least, we will present an evaluaion method based on PageRank algorithm [8] to estimate the publications in the graphs generated by the method before, the evaluate method will not only take the citation count, but also calculate each paper's score by distribution of keywords. Hopefully, a list of ranking score for each publication in the scientific literature graph will be calculated for assessing the most significant papers in a special field.

2. RELATED WORK
2.1 Citation Analysis of Research Papers
Academic publications are often scrupulously reviewed. The significant characteristics of academic publications can be described 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 as the most significant field of bibliometrics has been studied for a long time and there has been a great deal of work on it. It is used as a way to analyze relationships between publications and their relative influence.

Since the 1900s, scientists and librarians were convinced of the growth of the research literature. Garfield [2] briefly reviews work in this field, i.e., "College libraries and chemical education" [3], seen as the progenitors of the field of citation analysis, with the purpose to assist the student whose major is chemistry to estimate the significant books for his research judged by experts from 1,600 books. In much earlier studies for citation analysis, the paramount approach is exploited to rank by the frequency of citations. Garfield [4] described a method to evaluate the journals by frequency and impact of citations for science policy studies based on Science Citation Index (SCI) and Institute for Scientific Information (ISI).

Based on these classics bibliometrics papers, more and more scholars put their research focus on citation frequency or citation impact and used it in different domains. Harhoff et al. [5] judged the value of patented inventions by the citation frequency, "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 [6], and drawn a conclusion that annual citation rates of ecological papers are affected by many factors, i.e., the hypothesis tested, article length, authors' information, which cast doubt for the validity of using citation counts for academic evaluation.

With the deeper study of citation analysis, more and more researches found the problems of citation count [7] and doubt if it is reasonable enough for simply counting the number of times an article or author is cited with the general assumption that the number of citations reflects an article's influence. On the other hand, full-text analysis is to some extent compensate for lack of citation count and offers new opportunity for the citation analysis.

2.2 Citation Analysis with PageRank

Recently, PageRank has become a significant method for evaluating the most important nodes in complex graphs analysis. Examples include social networks, Web graphs, telecommunication networks, and biological networks. 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 and deciding "which entities are most important in the network relative to a particular individual or set of individuals." The PageRank algorithm, first proposed by Page, Motwani and Winograd [8] and used in Google Search, is a method for computing a ranking score for every web page based on the graph structure of the web to measure the relative importance of web pages. Distinguished from the traditional method of simple backlink counts, PageRank utilizes the graph to re-calculate 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 [9].

Although more and more publications focus on PageRank, most prior research for improving the ranking of search-query results computes a single vector using a link structure of the network which is independent of particular search queries. Haveliwala [10] 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.

3. PROPOSED METHOD

3.1 Graphs based on CPDD

Most of the existing efforts of domain knowledge graphs are based on a selected domain publication corpus, such as the particular core journals or conference proceedings, and this sampling strategy can be biased. Especially when a large number of domain literatures are not available, it is difficult to generate or estimate comprehensive and accurate domain knowledge. Fortunately, utilizing "domain context" publications, i.e., closely related journals or domain cited publications, can be able to solve the disconnection problem between the domain and its context.

First, we will propose a method based on Citation Probability Distribution Distance, which will take into account the probability distribution distance (K-L Divergence) that each keyword cite the domain or domain context publications. The process of generating the graphs is shown as follows:

(1) Compute the probability of keyword and citation, according to:

$$P(C_i \mid Key_i) = \frac{count_{revelant}(C_i, Key_i)}{count(Key_i)}$$

where $\sum \mathrm{count}(\mathrm{Key_i})$ and $\sum count(C_i, Key_i)$ denote frequency of each keyword appeared in papers and the frequency of each keyword appeared in each citation, respectively.

(2) Generate an array for keyword and citation:

$$\begin{array}{c} \\ keyword_1 \\ \vdots \\ \vdots \\ keyword_n \end{array} \begin{array}{c} \begin{array}{cccc} cited_1 & \cdots & \cdots & cited_m \end{array} \\ \begin{bmatrix} p_1 & \cdots & \cdots & p_m \\ \vdots & \ddots & & \vdots \\ \vdots & & \ddots & \vdots \\ p_n & \cdots & \cdots & p_s \end{bmatrix} \end{array}$$

(3) Generate the weight for keywords by K-L Divergence.

$$D_{kl}(K_1 \parallel K_2) = \sum_i K_1(i) \cdot \ln \frac{K_1(i)}{K_2(i)}$$

K-L divergence is a non-symmetric measure of the difference between two probability distributions K_1 and K_2, so the weight for K_1 and K_2, can be computed as:

$$weight(K_1, K_2) = \frac{D_{kl}(K_1 \| K_2) + D_{kl}(K_2 \| K_1)}{2}$$

3.2 Optimized CPDD based on Fulltext

But, there is an obvious limitation for the research methodology mentioned before, the potential hypothesis here means that every citation has the equal relationship with every topic (keyword) for each paper, i.e., paper1 citing paper2 just because all the topics of paper1 have equal relationships with paper2. Distinctly, there is a bias which is not corresponds to what happens in reality. In fact, citation occurs between two papers only because of one or limited number of reasons (topics), but hardly related to all possible topic pairs. So, taking into account the fulltext analysis will generate the

citation context and extract the citation cause, then enhance the improvement of the algorithm.

3.3 PageRank Algorithm based on Fulltext

After fulltext analysis for the experiment of citation analysis, we can investigate the semantic keywords relationships by exploring the citing reason, and, we can enhance the publication ranking function by leveraging this relationship in the domain and domain context graph. A further research direction can be represented as: estimating the significance of papers based on fulltext analysis in the graph by modified PageRank algorithm.

PageRank [8] method has been widely used in the web page analysis, taking into account this idea, this paper will put the focus on the citation relationships, which seen the keywords as the page link, and will calculate the rank for each paper by the cited paper contribution.

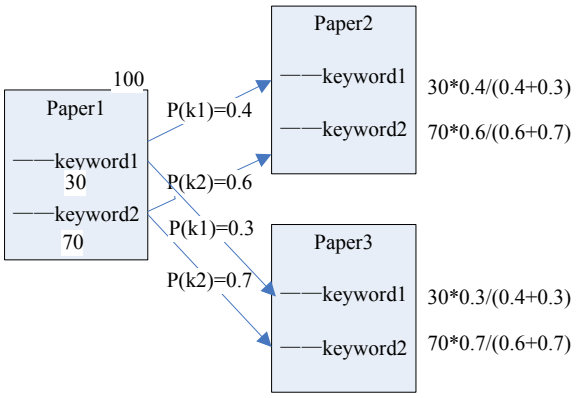

Figure 1. Optimized PageRank sample.

As shown in the figure above, the initial score for each paper is 100, and the paper1 has two citations: paper 2 and paper 3. The distribution for keyword1 and keyword2 are 0.3 and 0.7, respectively. So the topical scores for keyword1 and keyword2 for paper1 are 30 and 70.

For paper1, the score of keyword1 can divide into two parts, 40% to paper2, and 30% to paper3, which is estimated by the citation context. So for paper2, it is means that the score of keyword1 in paper2 can be added from paper1 is about 30*0.4/(0.4+0.3)=17.14. It is same for keyword2 in paper2 and the score is 32.31, so the added score for paper2 is: 100+17.14+32.31=149.45. The new score for paper3 by the same method is 150.55.

Table 1 Ranking score for citation papers

	PageRank	Optimized PageRank
Paper1	100	100
Paper2	150	149.45
Paper3	150	150.55

By now, we observed that if the analysis only by citations count, the paper2 is as important as paper3, but actually the paper3 is more significant than paper2 by considering the citation cause.

4. EXPERIMENTS

4.1 Graphs Generated by CPDD

As the method described in the 3.1, we will present two graphs based on CPDD for Domain and Domain Context.

4.1.1 Experiments Data

For Domain graph, we will select 26 journals and conferences, including 6171 papers as the domain for the selected field (information retrieval). And 50617 papers treated as the domain context, which are closely related journals, conference or domain cited publications.

4.1.2 Experiments Result

Using method described above in 3.1, we can generate two graphs: domain citation keyword graph and domain context citation keyword graph.

Table 2. Graphs parameters based on CPDD

	Domain Citation	Domain Context Citation
Vertices	376	5460
Edges	1570	63738
AVG. PATH length	1.978	2.964
Network Diameter	2	8
Graph density	0.022	0.043

Compare the two graphs mentioned above, the density of Domain Context citation graph is a little higher than the other graph, while it has more vertices and edges, which yielded some noisy nodes. The network becomes denser as the cutoff value becomes lower, whereas it becomes sparser as the cutoff value becomes higher. So the analyzer has to select a reasonable value of the cutoff, made the structure of network becomes clearly visible. Therefore, we will present three rules for combing the two graphs, in order to get a new graph, which can be clear visible and reasonable value of the cutoff, and showed as the table 2.

a. Re-weight edges (by using linear combination)

b. Integrate additional edges from domain context

c. Integrate additional vertexes from domain context (via connectivity)

Table 3 Combined graph parameters based on CPDD

	New Citation Graph
Vertices	517
Edges	11601
AVG. PATH length	1.977
Network Diameter	3
Graph density	0.087

4.1.3 Evaluation Method

First, we trained the LLDA [1] topic model for each keyword based topic (the keyword is assumed as the topical label for each publication) by using publication abstract, and generate a probability distribution for each keyword $P(word_x|keyword_y)$.

Then, we calculated the cosine similarity for each two keywords pair, and using the similarity between two keywords as the real-world topical distance. To validate the accuracy of a specific keyword based knowledge graph, we assume if, on the graph, two keywords are directly connected, the cosine similarity (with LLDA model) score should be high. Otherwise, if they are not directly connected, the similarity score should be low. Based on this assumption, we calculated the graph precision, recall and F-measure.

In the preliminary experiment, we find the new combined graph can enhance the F score, which is presented in Table 3.

Table 4 Evaluation results for 3 graphs

Graph name	precision	Recall	F-Score
Domain citation	0.935	0.126	0.111
Context citation	0.291	0.37	0.162
Combine citation	0.813	0.313	0.226

4.2 Optimized CPDD and PageRank by Fulltext Extraction

Recently, we can present a graph based on the relationships of keywords and find out the most popular topics in the first experiment, but one limitation is that all the results analyzed by title and abstract for the papers. So what will be presented if taking into account the fulltext extraction and citation reason.

4.2.1 Experiments Data

In this experiment, 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.

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 (*$paper_1$ 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 contexts.

4.2.2 Experiments Method

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 prior.

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

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 $\{p_{z_{key_1}}(v_i|v_j), p_{z_{key_2}}(v_i|v_j), \dots p_{z_{key_n}}(v_i|v_j)$, 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}.

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

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

We can also calculate 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, and $P(z_{key_t}|citation_{j,i})$ is the citation topic inference score.

Unlike classical PageRank, a citation graph with vertex and edge priors permits non-uniformly-distributed random jumps. 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 $R = V$.

4.2.3 Evaluation Method

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. We clearly understand that review papers don't cover all important publications for the target area, 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, i.e., traditional PageRank without prior, Language Model, Vector Space.

5. CONCLUSION

By now, the first experiment has got a positive preliminary results, the domain graph and domain context graph were generated by CPDD method, some rules were proposed for combing a new graph, which can enhance the performance well. The second and third experiment based on fulltext extraction, citation analysis based on LLDA model and PageRank algorithm for generating ranking score are working in progress. The goal of the future experiments based on full text citation analysis can optimize the first experiment conclusion and present an effective way for evaluate the publication's contribution for a special domain knowledge graph.

6. REFERENCES

[1] 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 *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing* (Singapore, August 06 - 07, 2009). EMNLP '09. ACM, New York, NY, 248-256.

[2] Garfield, E. 1995. New International Professional Society Signals the Maturing of Scientometrics and Informetrics. Scientist *9*, 11-11.

[3] Gross, P.L., and Gross, E.M. 1927. College Libraries and Chemical Education. Science *66*, 385-389.

[4] Garfield, E. 1972. Citation Analysis as a Tool in Journal Evaluation - Journals Can Be Ranked by Frequency and Impact of Citations for Science Policy Studies. Science *178*, 471.

[5] 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*, 511-515.

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

[7] MacRoberts, M.H., and MacRoberts, B.R. 1996. Problems of citation analysis. Scientometrics *36*, 435-444.

[8] L. Page, S.B., R. Motwani, and T. Winograd. 1998. the pagerank citation ranking bringing order to the web. Technical report, Stanford Digital Library Technologies Project.

[9] Y. Sun and C. Giles. Popularity Weighted Ranking for Academic Digital Libraries. In Proc. of the 29th European Conference on Information Retrieval (ECIR 2007), pages 605–612, 2007.

[10] Haveliwala, T.H. 2003. Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search. Knowledge and Data Engineering, IEEE Transactions on, 15, 4, 784-796.