Popularity Weighted Ranking for Academic Digital Libraries

Yang Sun and C. Lee Giles

Information Sciences and Technology
The Pennsylvania State University
University Park, PA, 16801, USA

Abstract. We propose a popularity weighted ranking algorithm for academic digital libraries that uses the popularity factor of a publication venue overcoming the limitations of impact factors. We compare our method with the naive PageRank, citation counts and HITS algorithm, three popular measures currently used to rank papers beyond lexical similarity. The ranking results are evaluated by discounted cumulative gain(DCG) method using four human evaluators. We show that our proposed ranking algorithm improves the DCG performance by 8.5% on average compared to naive PageRank, 16.3% compared to citation count and 23.2% compared to HITS. The algorithm is also evaluated by click through data from CiteSeer usage log.

Keywords: weighted ranking, citation analysis, digital library.

1 Introduction

Effectively indexing and retrieving information from large document databases continues to be a challenging task. Automated digital libraries make it easier for users to access and use these databases. In Web search, PageRank [14] and HITS [9] algorithms created to measure importance or authority as a ranking factor showed a great success compared to lexical similarity measures. Citation count is also widely used in evaluating the importance of a paper. However, unweighted citation counting often does not accurately describe the impact of papers [11].

The publication process of academic papers makes the citation graph much different from the Web graph. First, publication date and content of papers usually do not change over time whereas those of the Web pages can. Second, the typical citation graph of academic papers is an acyclic digraph without loops (there are rare exceptions to this). It is also common that a paper does not cite future papers. (Except in unusual cases where papers can cite unpublished work that is published later. In that case these papers can be treated as multiple versions.) Thus, the interpretation of the naive PageRank algorithm would be problematic[13,17]. We introduce a popularity factor weighted ranking algorithm based on PageRank with significantly improved performance for ranking academic papers. Our contributions are as follows:

- We define a new popularity factor that reflects the influence of publication venues and overcomes the limitations of a venue's impact factor.
- A popularity factor weighted ranking score of a paper in the proposed ranking method is defined by the weighted citations from other papers and the popularity factor of its publication venue and is implemented on the CiteSeer metadata.
- A user study with four evaluators shows the improved performance of this new algorithm. We also use clickthrough data from CiteSeer usage log to validate the results.

2   Weighted Ranking

According to information foraging theory[15], users of information retrieval systems will evaluate the value of documents by information cues (such as title, author, venue, citation count, publication date of a paper in academic digital libraries) and follow the most valuable document. The more cues they encounter, the better they can evaluate the value. Lexical similarity only shows a limited information cue about a document. Citation count as an information cue is usually considered to be strongly correlated to academic document impact [12]. Although it is widely used in academic evaluation, citation count has limitations which make it less precise. The citation count of an individual paper by itself does not reflect the different citation structure in each discipline[19]. The papers with high impact and the ones with low impact are treated the same in citation count [4].

2.1   Weighted Ranking Method

In our research users are modeled as optimal information foragers who evaluate the cost of pursuing a document by information cues and follow the most valuable document [15]. According to this user model, the reference in a high impact paper will have a high probability to be followed by users. The citations of a paper should be viewed as weighted links. Not only the count of citations but also the impact of citations matters in this sense. Furthermore, the quality of the publication venue where a document is published is also an important information cue for users to evaluate the value of a document.

Popularity Factor. All serious research publication venues have a peer review process for publishing papers. It is fair to consider that the impact of a paper is partially reflected by where the paper is published. Impact factors of journals are widely used in evaluating the quality of publication venues[3]. There are limitations about the definition and the usage of impact factors. The impact factor is not normalized across research areas[6]. The calculation of an impact factor only considers a 3 year period. But important papers may receive many citations after 3 years[18]. Conferences are not considered in the calculation. Conferences often play very important roles in computer and information science research because of their timeliness and popularity.

We introduce the popularity factor to consider venue as an information cue and to reflect the influence of a publication venue. The popularity factor is defined based on citation analysis of publication venues. Note that the popularity factor does not distinguish journals from conference or workshop proceedings. The popularity factor of a publication venue v in a given year is defined by Equation 1:

$$PF(v,t) = \frac{n_v}{N} * \sum_{i \in P} \frac{PF(i,t) \times w(i)}{N(i)}. \tag{1}$$

where $PF(v,t)$ is the popularity factor of publication venue v in a given year t, P is the set of publication venues i which cite v in that year, and n_v is the number of papers published in venue v in that year. If N is the total number of papers published in that year in a particular discipline, n_v/N represents the probability of a reader having access to a paper published in the publication venue v. Let $w(i)$ be the weight which represents the frequency that venue i cites venue v. $N(i)$ is the total number of references generated by venue i, and $PF(v,t)$ is normalized so their squares sum to 1 for reasons of convergence: $\sum_v (PF(v,t))^2 = 1$ and has a range from 0 to 1 with larger values ranked higher. In our definition, a discipline is considered as a collection of related publication venues. The number of total papers in a discipline can be obtained by counting all the papers in all venues of the discipline. Multiple popularity factor values for one venue may occur in different disciplines.

According to our definition, this popularity factor differs from the impact factor by considering the impact of all publication venues, recent to long ago papers, and the probability of reader access. These differences overcome several shortcomings of the impact factor and provide a robust and reliable measure for publication venues. The popularity factor is computed with a simple iterative algorithm and achieve convergence point after 18 iterations [9,14]. Top 5 venues ranked by popularity factors in 2004 based on our database (includes primarily Computer Science and Engineering papers) is listed in Table 1.

Table 1. Popularity factors for computer science venues in 2004. Conferences and journals are both included.

Popularity factor	Name
0.05868	INFOCOM
0.04277	ACM SIGPLAN
0.04027	ACM SIGCOMM
0.02731	Human factors in computing systems
0.02622	Mobile computing and networking

Ranking Score. A paper will nearly always be cited after it is published. A citation relationship in a published document should not change over time (revisions to technical reports may be an exception). Figure 1 illustrates the temporal effect of citation graphs.
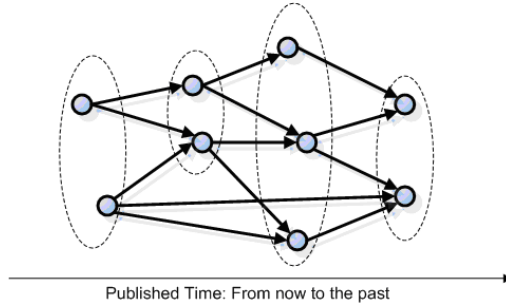
Published Time: From now to the past

Fig. 1. A schematic illustration of a citation graph. Each circle represents a paper. Arrows represent citations. It can be seen that citation graphs typically do not have backward citations. They are acyclic digraphs.

With the popularity factor and temporal effect, we define the ranking score $R(d_T)$ of an academic paper d at a previous time T in Equation 2 as

$$R(d_T) = PF(v_{d_T}) + \sum_{t>T, d_t \in D} \frac{R(d_t)}{N(d_t)}. \tag{2}$$

where $R(d_t)$ is the ranking score of a paper dt which is published at time t and cite paper d_T. D is the set of papers which cite d_T. $N(d_t)$ is the number of references in paper d_t. $PF(v_{d_T})$ is the popularity factor of the publication venue v where paper d_T is published. The ranking score has a range of 0 to very large numbers. The vector PF is considered as an initial score of a paper when there is no citation record for this paper. This ranking assumes that the ranking score of a previously published paper will not have any impact on later published ones. Our algorithm does not permit bidirectional citations. If a paper is cited before it is published (no a common case), the paper will be considered to have two versions and the two versions will have separate rankings. The citation graph can be constructed as a function of publication time because of the temporal property of papers. The adjacency matrix of the graph can be then sorted and form a strict upper triangular matrix. Then, the equation of ranking scores can be written to a system of n linear equations with n unknowns, which has a single unique solution. As such there are no convergence issues. Notice that the ranking function has a computational complexity of $O(nm)$ where n is the number of papers in the database and m is the average citations to a paper. m typically ranges from 0 to less than 50000 with a power law distribution for academic papers [16], making the ranking algorithm scalable for large digital libraries. The evaluation method and the results are discussed in the next section in order to demonstrate how our algorithm improves the ranking performance.

3   Evaluation

Evaluating the quality of ranking algorithms used in an information retrieval system is a nontrivial problem[5,7]. Most evaluation methods involve human

evaluators judging the relevance of the retrieved documents to specific queries. In this research we compare our ranking method to naive PageRank, citation count and HITS using discounted cumulative gain method with four human evaluators. Two of the evaluators were graduate students from computer science department whose research interest is in data mining and Web search. One is a research programmer who has experience on machine learning program development. The other evaluator is a software engineer with Master's degree in computer science. The four ranking algorithms are implemented in a basic information retrieval system which indexes 536,724 papers' metadata from the CiteSeer database[1] using Lucene[2] in which *tf-idf* is used as the basic ranking function. The evaluation is expected to compare the performance of the re-ranking functions beyond lexical similarity. Top 5 ranked papers by our algorithm with comparison to naive PageRank, citation count and authority scores of HITS are shown in Table 2.

Table 2. Ranking scores of the top 5 papers in computer science

Title	Weighted	PageRank	Citation	HITS
The Anatomy of a Large-Scale Hypertextual Web Search Engine - Brin (1998)	0.13504	0.11924	521	0.89659
Boosting a Weak Learning Algorithm By Majority - Freund (1995)	0.07568	0.06363	174	0.27771
A Tutorial on Learning With Bayesian Networks - Heckerman (1996)	0.04321	0.04048	203	0.20106
Irrelevant Features and the Subset Selection Problem - John (1994)	0.04097	0.05434	290	0.14429
Dynamic Itemset Counting and Implication Rules for Market Basket Data - Motwani (1997)	0.03944	0.04392	232	0.43546

3.1   Evaluation Method

Discounted cumulative gain (DCG) [8,20] measure considers the ranking of papers in addition to the relevant score from evaluators. The assumption in DCG is the lower the ranked position of a paper the less valuable it is for the user because the less likely it is going to be examined by users. Human evaluators are required to rate the ranked papers with a score from 0 to 2, where 0 represents non relevance, 1 represents marginal relevance, and 2 represents high relevance. The DCG value is computed separately at each relevance level. The agreement among evaluators is examined by kappa statistics to show the confidence level by using the evaluation results. The kappa statistic is a standard measure of agreement with categorical data to access the degree to which two or more raters agree in assigning data to categories [10].

 Fifty queries were selected from the CiteSeer query log. Papers are ranked by the four ranking algorithms separately in addition to the *tf-idf* lexical similarity measure. Top twenty papers are mixed and presented to four human evaluators (using four evaluators is considered enough for DCG evaluation methods [8,20]).

3.2   Results

Kappa Measure. The agreement among the four evaluators was examined for each of the 50 queries. The average kappa agreement among the four evaluators is 53% which is considered to be in the level of moderate agreements [10].

Precision Recall. The precision-recall curves at different relevance levels for the four algorithms using standard methods [8] are presented in Figure 2.
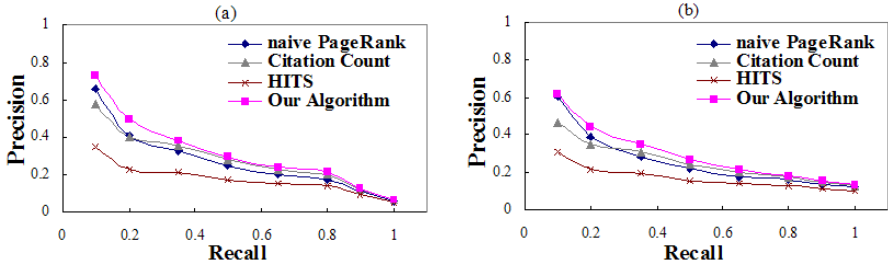

Fig. 2. Precision-Recall of the four ranking algorithms at relevance level 1(a) and 2(b)

DCG. The DCG vector curves for ranks 1-20 is shown in Figure 3. The DCG scores are shown in Table 3. Both the curve and statistics show that our algorithm significantly outperforms the other three algorithms for documents ranked after rank 10.
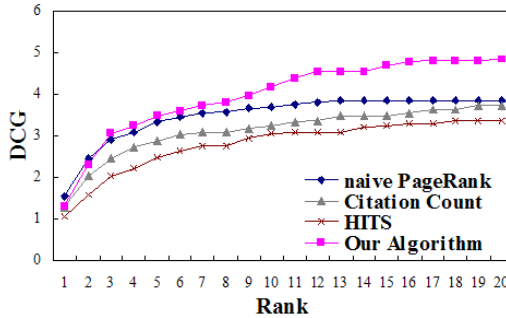

Fig. 3. DCG at various document rankings

3.3   Validity of Results

To validate the results, a pairwise measure of each ranking algorithm is calculated using the clickthrough data extracted from CiteSeer usage log. For any pair of papers, if the clickthrough rate of the high-ranked paper is larger than the low-ranked paper, we consider this pair is correctly ranked. The average pairwise accuracy of each algorithm is listed in Table 4.

Table 3. The DCG score for the four ranking algorithms

	Our Algorithm	naive PageRank	Citation Count	HITS
DCG @ rank 1	1.3	1.54	1.28	1.05
DCG @ rank 5	3.49	3.33	2.87	2.49
DCG @ rank 10	4.18	3.71	3.24	3.06
DCG @ rank 15	4.69	3.86	3.49	3.25
DCG @ rank 20	4.86	3.86	3.71	3.36
Ave. DCG(1-10)	3.28±0.87	3.12±0.68	2.69±0.62	2.35±0.64
Ave. DCG(10-20)	4.68±0.16	3.85±0.04	3.54±0.14	3.25±0.11

Table 4. Average pairwise accuracy based on Clickthrough data

	Our Algorithm	naive PageRank	Citation Count	HITS
Ave. accuracy	74.18%	70.29%	67.1%	67.13%

4   Discussion and Conclusions

A new weighted ranking score of a paper was defined by the weighted citations from other papers and the popularity factor of its publication venue. A ranking system based on the Lucene index and Citeseer metadata was built and was used to evaluate our algorithm with comparison to other popular ranking algorithms.

The algorithm is evaluated by DCG method using human evaluators and we compare its results to the ranking of naive PageRank, citation counts and HITS. The comparison results show that the weighted ranking algorithm improves the DCG ranking performance by 8.5% compared to naive PageRank, 16.3% compared to citation count and 23.2% compared to HITS. We also use the clickthrough data from CiteSeer usage log to validate the results. This leads us to believe that our weighted ranking algorithm is more accurate than those currently being used.

Our evaluation experiment shows that the user agreement on paper rankings is not very high. Effective personalized ranking algorithms would most likely satisfy the diversity of most user information needs.

References

1. Citeseer, http://citeseer.ist.psu.edu.
2. D. Cutting, "The Lucene Search Engine," http://lucene.apache.org/, 2006.
3. E. Garfield, "The impact factor," Current Contents, 25, 3-7, 1994.
4. S. Harnard, "The New World of Webmetric Performance Indicators: Mandating, Monitoring, Measuring and Maximising Research Impact in the Open Access Age," Proc. of the 1st ECSP in Biomedicine and Medicine, 2006.
5. D. Hawking, N. Craswell, P. Thistlewaite, and D. Harman, "Results and challenges in Web search evaluation," Proc. of the 8th International World Wide Web Conference, 1321-1330, 1999.

6. F. Hecht, B. Hecht, and A. Sandberg, "The journal "impact factor": a misnamed, misleading, misused measure," Cancer Genet Cytogenet, 104(2), 77-81, 1998.

7. D. Hull, "Using statistical testing in the evaluation of retrieval experiments," Proceedings of the 16th annual international ACM SIGIR Conference, 329-228, 1993.

8. K. Jarvelin and J. Kekalainen, "IR evaluation methods for retrieving highly relevant documents," Proc. of the 23rd SIGIR conference, 41-48, 2000.

9. J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," Journal of ACM, 48, 604-632, 1999.

10. R. J. Landis and G. G. Koch, "The measurement of observer agreement for categorical data," Biometrics, 33, 159-174, 1977.

11. S. Lehmann, B. Lautrup, and A. D. Jackson, "Citation networks in high energy physics," Physical Reivew E68, 026113 (2003).

12. F. Narin, "Evaluative bibliometrics: The use of publication and citation analysis in the evaluation of scientific activity," Cherryhill, N.J.: Computer Horizons, 1976.

13. Z. Nie, Y. Zhang, J. Wen, and W. Ma, "Object-Level Ranking: Bringing Order to Web Objects," Proc. of the 14th International World Wide Web Conference, 2005.

14. L. Page and S. Brin, "The PageRank citation ranking: bringing order to the web," tech. report SIDL-WP-1999-0120, Stanford University, Nov. 1999.

15. P. Pirolli and S. Card, "Information foraging in information access environments," Proc. of the SIGCHI conference, 51 - 58, 1995.

16. S. Redner, "How Popular is Your Paper? An Empirical Study of the Citation Distribution," European Physical Journal B, 4, 131-134, 1998.

17. M. Richardson, A. Prakash, and E. Brill, "Beyond PageRank: Machine Learning for Static Ranking," Proc. of the 15th International World Wide Web Conference, 2006.

18. P. Seglen, "Why the impact factor of journals should not be used for evaluating research," British medical journal, 314(7079), 498-502, 1997.

19. Thomson and Corporation, "In Cites: Analysis Of" http://www.in-cites.com/analysis/, 2005.

20. E. Voorhees, "Evaluation by Highly Relevant Documents," Proc. of the 24th annual international ACM SIGIR conference, 74-82, 2001.