On the Temporal Dimension of Search

Philip S. Yu
IBM T.J. Watson Research Center
19 Skyline Drive
Hawthorne, NY 10532
psyu@us.ibm.com

Xin Li
Department of Computer Science
University of Illinois at Chicago
851 South Morgan Street,
Chicago, IL 60607-7053
xli3@cs.uic.edu

Bing Liu
Department of Computer Science
University of Illinois at Chicago
851 South Morgan Street,
Chicago, IL 60607-7053
liub@cs.uic.edu

ABSTRACT
Web search is probably the single most important application on the Internet. The most famous search techniques are perhaps the PageRank and HITS algorithms. These algorithms are motivated by the observation that a hyperlink from a page to another is an implicit conveyance of authority to the target page. They exploit this social phenomenon to identify quality pages, e.g., "authority" pages and "hub" pages. In this paper we argue that these algorithms miss an important dimension of the Web, the temporal dimension. The Web is not a static environment. It changes constantly. Quality pages in the past may not be quality pages now or in the future. These techniques favor older pages because these pages have many in-links accumulated over time. New pages, which may be of high quality, have few or no in-links and are left behind. Bringing new and quality pages to users is important because most users want the latest information. Research publication search has exactly the same problem. This paper studies the temporal dimension of search in the context of research publication search. We propose a number of methods deal with the problem. Our experimental results show that these methods are highly effective.

Categories & Subject Descriptors
H.3.3 [**Information Storage AND Retrieval**]: Information Search and Retrieval - Search process, Selection process.

General Terms
Algorithms, Experimentation.

Keywords
Temporal dimension of search, publication search, Web search.

1. INTRODUCTION
The most successful Web search techniques, PageRank [2] and HITS [6], exploit the hyperlink structure of the Web to find quality pages such as "authorities" and "hubs" [1] [3] [4] [5] [6] [7]. However, an important factor that is not considered by these techniques is the timeliness of search results. In this paper, we study search from the temporal dimension. This dimension is essential when users look for the latest information. We believe that the temporal dimension of search is of great importance to the future developments of search technology. In this paper, we take the first step towards this direction. We investigate this problem in the context of research publication search [8] because of two reasons. First of all, results in the research publication domain can be objectively evaluated, as all the citation data of the paper collection are available. Secondly, concepts in both academic citation and Web domains are largely the same. For example, a research paper corresponds to a Web page, and a citation of a research paper corresponds to a hyperlink in a Web page.

We present a number of methods to incorporate time dimension in paper search. These methods are evaluated experimentally. The results show that the proposed methods are highly effective.

WWW 2004, May 17-22, 2004, New York, New York, USA.
ACM 1-58113-912-8/04/0005.

2. THE PROPOSED TECHNIQUES
There are many factors that contribute to the potential importance of a paper, such as the citations it has received, the date of these citations, its authors, and the publication journal. PageRank only includes the first factor, the citations that a paper receives. To integrate the time dimension, we add timing factors in the PageRank and propose the TimedPageRank algorithm.

2.1 TimedPageRank
Since we are interested in both a paper's current importance and its potential, we modify PageRank by weighting each citation according to the citation date. The system calculates the time-weighted PageRank (PR^T) value for each paper as follows:

$$PR^T(A) = (1-d) + d \times \left(\frac{w_1 \times PR^T(p_1)}{C(p_1)} + \ldots + \frac{w_n \times PR^T(p_n)}{C(p_n)}\right) \quad (1)$$

where
$PR^T(A)$ is the time-weighted PageRank score of paper A,
$PR^T(p_i)$ is the time-weighted PageRank score of paper p_i that links to paper A,
$C(p_i)$ is the number of outbound links of paper p_i and
d is a damping factor, which is set to 0.85.

Equation (1) is a modified version of the original PageRank. In this equation, we introduced the timed weight for each citation, w_i. Its value reduces exponentially with the citation age; the base of the exponential function is *DecayRate*. *DecayRate* is a parameter (we use 0.5 in our experiments). Note that if *DecayRate* is 1, the time-weighted PageRank algorithm will be the same as the original PageRank algorithm. In addition, the *DecayRate* parameter can be tuned according to the nature of a dataset/the user.

After the current importance of a paper is evaluated, we also want to know how the importance changes in the future year. Our experimental data show an obvious trend that a new paper is more likely to draw citations than an old paper. Therefore, another parameter called the *agingfactor*, *Aging*(A) (which is in [0, 1]), is introduced. In our experiments a brand new paper's *agingfactor* is 1. After the publication, it declines linearly with

Table 1: Comparison results of different methods using all papers

1	2	3	4	5	6	7	8	9	10	11	12
No. of top papers	Original PageRank		TPR		TPR (AJEval)		LR		LR (AJ Eval)		Best citation count
10	2516	44%	3982	70%	4212	74%	4219	75%	4136	73%	5661
20	3406	46%	5258	72%	5435	74%	5371	73%	5457	74%	7345
30	4024	48%	6144	73%	6306	75%	6385	76%	6519	78%	8406

time, and its range is from 0.5 to 1. Thus, a paper A's final TimedPageRank (TPR) is computed as follows:

$$TPR(A) = Aging(A) * PR^T(A) \quad (2)$$

2.2 Source Evaluation: Author and Journal

Although TimedPageRank is able to boost the rank of emerging quality papers, it is not sufficient for all the papers because new papers only have a few or no citation. To assess the potential importance of a new paper, its source information, its authors and the journal, are useful.

We compute author evaluation by averaging the time-weighted PageRank values of all the past papers of the author, and journal evaluation by averaging the time-weighted PageRank values of all the papers published in the journal. We can evaluate paper based on its journal evaluation, or its author evaluation, or the combination of both.

2.3 Linear Regression

Another simple technique that can be used to rank a paper is linear regression. That is, one can use citation count of the paper received in the past few years to perform a linear regression to predict the citation count in the coming year. This predicted citation count can be used as the score of the paper for final ranking. This method is fairly straightforward and will not be discussed further.

Similar to TimedPageRank technique, linear regression will not be accurate if a paper is published only recently, and has only a few or no citation. In this case, we again used author and journal evaluation to score the new paper. In this case, author or journal evaluation can be done by using actual citation counts of all the papers of the author or journal. After they are computed, we can use the journal evaluation, or the author evaluation, or the combination of both to make the final source evaluation.

3. EMPIRICAL EVLAUATIONS

In our experiments, we used the research publication dataset from the KDD CUP 2003. The documents are from an archive of High Energy Particle Physics publications. Our system ranks the papers that are relevant to a given query and presents to the user. For the purpose of this research, we assume that a paper is relevant to a query as long as it contains all the query words. We experimented with 25 queries. To evaluate the performance of proposed techniques, we do not compare their rankings directly. Instead, we compare the number of citations that the top ranking papers receive in the following year to evaluate our algorithms.

Table 1 presents the experiment results. Only the results for the top 30 papers are given. The results are presented in 3 rows. Each row gives the total citation counts for a group of top papers based on different evaluation methods. Column 2 gives the total citation count for each group of retrieved top papers by Original PageRank for all 25 queries, Column 3 gives the ratio of the total citation count for this method and the total citation count of the ideal ranking (called *best citation count* given in Column 12). Similar to columns 2 and 3, column pairs, 4-5, 6-7, 8-9, and 10-11 show the citation counts of retrieved papers, and the citation percentage for different techniques respectively.

From the data in table 1, we can draw the conclusion that both TimedPageRank and Linear Regression perform significantly better than the original PageRank. Moreover, the author and journal evaluation is helpful in improving the prediction results in most cases.

4. CONCLUSIONS

This paper studies the temporal dimension of search. So far, little research work has been done to consider time in either publication search or Web search. In this paper, we made an attempt to study this problem. A number of techniques to remedy the situation were proposed. Experiments have also been conducted to evaluate these techniques. Our results show that the proposed techniques are highly effective. Furthermore, the proposed methods can be conveniently adapted to Web search because concepts in the two domains are largely parallel. In our future work, we plan to conduct this research.

5. REFERENCES

[1] A. Arasu, J. Cho, H. Garcia-Molina, A. Paepcke, and S. Raghavan. Searching the Web. ACM Transactions on Internet Technology, 1(1), 2001.

[2] S. Brin, L. Page. The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems, 30, 1998.

[3] A. Borodin, J. S. Rosenthal, G. O. Roberts, and P. Tsaparas, Finding authorities and hubs from link structures on the world wide web. WWW-2001.

[4] S. Chakrabarti, B. Dom, D. Gibson, J. Kleinberg, P. Raghavan, and S. Rajagopalan. Automatic resource compilation by analyzing hyperlink structure and associated text. WWW-1998.

[5] T. Haveliwala. Topic-sensitive PageRank. WWW-2002.

[6] J. Kleinberg. Authoritative sources in a hyperlinked environment. ACM-SIAM Symposium on Discrete Algorithms, 1998.

[7] J. Kleinberg, S. R. Kumar, P. Raghavan, S. Rajagopalan, and A. Tomkins. The Web as a graph: measurements, models, and methods. International Conference on Combinatorics and Computing, 1999.

[8] S. Lawrence, K. Bollacker, and C. L. Giles. Indexing and retrieval of scientific literature. CIKM-99.