Yet Another Paper Ranking Algorithm Advocating Recent Publications

Won-Seok Hwang
Dept. of Electronics and Computer Engineering
Hanyang University, Korea
hws23@hanyang.ac.kr

Soo-Min Chae
Dept. of Electronics and Computer Engineering
Hanyang University, Korea
aesem@hanyang.ac.kr

Sang-Wook Kim
Dept. of Electronics and Computer Engineering
Hanyang University, Korea
wook@hanyang.ac.kr

ABSTRACT

In this paper, we propose a new paper ranking algorithm that gives a high rank to papers which is credited by other authoritative papers or published in premier conferences or journals. Also, the proposed algorithm solves a problem that recent papers are rated poorly due to few citations.

Categories and Subject Descriptors: H.3.3 [Information Storage and Retrieval] Information Search and Retrieval

General Terms: Algorithms, Experimentation

Keywords: Ranking, PageRank, Impact Factor, Citation Analysis

1. INTRODUCTION

Owing to the development of the Internet and Web technology, most academic papers are being searched by paper search engines in the web rather than in libraries. DBLP, CiteSeer, Google Scholar, and Libra are the typical examples of paper search engines.

Since a large number of papers could be matched to a query, ranking is crucial in paper search engines. The inherent properties of paper ranking compared with web page ranking are two-fold: (1) A paper can cite only those papers published earlier than itself, and cannot modify the citations once done; (2) In addition to citations, there are various types of information to be used in ranking such as titles, abstracts, contents, references, keywords, authors, publication dates, and publication venues. Property (1) causes recent papers to hardly get high scores in citation-based ranking. Our goal is to develop a paper ranking algorithm that exploits these two properties.

There have been several algorithms in the literature that partially satisfy the properties. PopRank [3] utilizes the author-paper relationship and the publication venue-paper relationship apart from citations. The Browsing-Based Model [5] also utilizes the author-paper relationship. These two algorithms consider the quality of papers independently of the relevance to queries. Authority-Based Ranking [2] determines ranking by simultaneously taking citations, authors, publication venues, and relevance to queries into account. CiteRank [4], different from others, considers the recency of papers in ranking by exploiting the publication dates. However, none of these algorithms satisfy those properties above completely.

WWW 2010, April 26–30, 2010, Raleigh, North Carolina, USA.
ACM 978-1-60558-799-8/10/04.

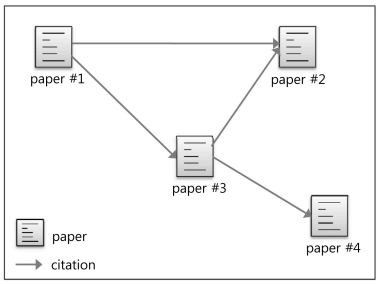

Figure 1: A graph modeled for scientific literature.

In this paper, we propose a new paper ranking algorithm that achieves the following goals.

- G1: To give a high rank to papers credited by a number of good papers.
- G2: To give a high rank to papers published in premier publication venues.
- G3: To solve the distortion in ranking due to publication dates, which causes recent papers never ranked high.

2. THE ALGORITHM

To acheive (G1), we employ *Random Walk with Restart* (RWR) on a graph where nodes are papers and edges are citations among papers. Figure 1 is an example of a graph modeled for a scientific literature database. The edge from p_1 to p_2 means that p_1 cites p_2.

$$\mathbf{r}_{i+1} = (1-\alpha)(C^T + \mathbf{w} \times \mathbf{d}^T) \times \mathbf{r}_i + \alpha \mathbf{w} \qquad (1)$$

Equation 1 represents the concept of RWR with the modeled graph. C is an adjacency matrix for the graph. $\mathbf{w}$ is a vector whose elements w_i is set to a uniform value $1/N$, where N is the number of nodes in the graph. The score vector $\mathbf{r}_i$ stores the scores of all the nodes at step i. The parameter α controls the ratio of the random walk and the restart. The vector $\mathbf{d}^T$ represents dangling nodes (1 for the dangling and 0 for the others).

By computing Equation 1 iteratively, vector $\mathbf{r}_i$ converges to a certain vector, which is called the authority score vector. The *authority score vector* provides the ranks of papers and we will call this result $Ranking_{\mathrm{G1}}$.

In order to achieve (G2), we utilize the *reputation* of publication venues where papers appear. For computing the reputation of publication venue v, we use *impact factor* of v at a certain year y defined as follows:

$$IF(v, y; t) = \frac{Cited(\cup_{i=1..t} V_{y-i}, y)}{|\cup_{i=1..t} V_{y-i}|} \quad (2)$$

In Equation 2, V_y is a *set of papers* published in venue v at year y and t denotes the size of the time unit considered. The function $Cited(A, y)$ counts the number of citations of the papers in A from all the papers published in year y. The original impact factor uses $t = 2$, but Yan [5] showed that it normally takes 5 years to get sufficient amount of citations. So, we set t as 5 to reflect this.

To reflect impact factors in ranking, we modify and normalize vector $\mathbf{w}$. Let's consider p_i that was published in venue v_i at year y_i. w_i, the element corresponding to p_i in $\mathbf{w}$, is set to $\frac{IF(v_i,y_i;5)}{\sum_{j=1}^{N} IF(v_j,y_j;5)}$. For example, if the impact factors of venues for p_1, p_2, p_3 and p_4 in Figure 1, are 2.5, 2.5, 0.5 and 0.5, respectively, $\mathbf{w}$ is set to $(0.417, 0.417, 0.083, 0.083)^T$. This modification results in new ranking $Ranking_{G2}$.

To achieve (G3), we should consider the ages of papers. Compared to old papers, young (recent) papers have little chance to be cited by others, thereby being always ranked low. To overcome this distortion, we define the *age damping factor* ρ_p for each paper p as in Equation 3.

$$\rho_p = e^{-age(p)/\tau}/\tau \quad (3)$$

Equation 3 is a slight modification of the probability function of CiteRank [4], where τ denotes the *characteristic decay time* and $age(p)$ denotes the age of p. According to our experiments, 4 and 8 are reasonable for τ.

To reflect the age damping factor ρ_{p_i} of paper p_i in vector $\mathbf{w}$, the impact factor of the publication venue is multiplied by ρ_{p_i}: i.e., w_i is set to $\frac{IF(v_i,y_i;5)\times\rho_{p_i}}{\sum_{j=1}^{N} IF(v_j,y_j;5)\times\rho_{p_j}}$. We will call the result ranking $Ranking_{G3}$.

3. EXPERIMENTS

In our experiments, we used DBLP data, which was downloaded in March 2009, and the citation information was obtained from Libra. Our data has 1,071,973 papers and the average number of citations per paper is 7.67.

Figure 2 shows the average authority scores of papers over a year obtained by $Ranking_{G2}$ and $Ranking_{G3}$. For $Ranking_{G2}$, the average authority score of old papers is much higher than that of recent papers. However, for $Ranking_{G3}$, the gap of the average authority scores of the two groups is small. Figure 2 also shows the effect of τ; taking the lower value of τ pulls up the authority score of young papers in $Ranking_{G3}$.

Table 1 compares the precisions of the paper ranking algorithms, CiteRank [4], Browsing-Based Model (shown as BBM) [5], PopRank [3], and the proposed algorithm. The parameter α is set to 0.15 for all the algorithms. For our algorithm, τ is to 4 and 8.

We selected 6 queries (with popular keywords related to data mining: "clustering," "sequential pattern mining," "graph pattern mining," "spatial databases," "web mining," and "multirelational data mining"), looked for the top n papers ($n = 10, 20, 30$), and then compared them with references in the corresponding chapters of a famous data mining book [1]. Table 1 shows that our proposed algorithm is more accurate than the previous ones. The accuracy is slightly higher when τ is 8 than 4.

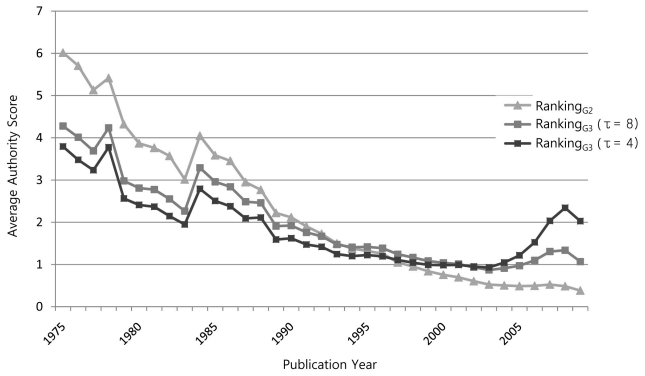

Figure 2: Effect of the age damping factor.

	CiteRank	BBM	PopRank	Proposed Alg.	
				$\tau = 4$	$\tau = 8$
Top 10	0.243	0.200	0.229	0.257	0.257
Top 20	0.171	0.136	0.207	0.200	0.207
Top 30	0.138	0.148	0.176	0.181	0.186

Table 1: Precisions of paper ranking algorithms.

4. CONCLUSIONS

In this paper, we proposed a new paper ranking algorithm which balances the impacts of old papers and new papers. To credit the recent papers, we defined the age damping factor for the papers. The age damping factor ρ has a special parameter τ denoting the characteristic decay time. According to the experimental results, the new algorithm is more accurate than existing algorithms when τ is between 4 and 8. Tuning the optimal value of τ could be an interesting issue remained.

5. ACKNOWLEDGMENT

This work was supported by NHN Corp. Any opinions, findings, and conclusions or recommendations expressed in this material are the authors' and do not necessarily reflect those of the sponsor.

6. ADDITIONAL AUTHORS

Additional author: Gyun Woo (Dept. of CSE, Pusan National Univ., email: `woogyun@pusan.ac.kr`)

7. REFERENCES

[1] J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2nd Edition, 2006.

[2] V. Hristidis, H. Hwang, and Y. Papakonstantinou. Authority-based keyword search in databases. *ACM Trans. Database Syst.*, 33(1), 2008.

[3] Z. Nie, Y. Zhang, J.-R. Wen, and W.-Y. Ma. Object-level ranking: bringing order to web objects. In *WWW*, pages 567–574. ACM, 2005.

[4] D. Walker, H. Xie, K.-K. Yan, and S. Maslov. Ranking scientific publications using a simple model of network traffic. *Journal of Statistical Mechanics*, 2007, June 2007.

[5] S. Yan and D. Lee. Toward alternative measures for ranking venues: a case of database research community. In *JCDL*, pages 235–244. ACM, 2007.