Time-Aware Weighted PageRank for Paper Ranking in Academic Graphs

Chin-Chi Hsu [*], Kuan-Hou Chan [†], Ming-Han Feng [‡], Yueh-Hua Wu [§], Huan-Yuan Chen [†],
Sz-Han Yu [†], Chun-Wei Chen [†], Ming-Feng Tsai [¶], Mi-Yen Yeh [*], Shou-De Lin [†]
[*]Institute of Information Science, Academia Sinica, Taipei, Taiwan
[†]Dept. of Computer Science and Information Engineering,
National Taiwan University, Taipei, Taiwan
[‡]Graduate Institute of Networking and Multimedia,
National Taiwan University, Taipei, Taiwan
[§]Dept. of Electrical Engineering, National Taiwan University, Taipei, Taiwan
[¶]Dept. of Computer Science, National Chengchi University, Taipei, Taiwan

ABSTRACT

Given a citation network, where each node is a paper and each directed link models the citation relationship, we propose to use a time-aware weighted PageRank model to rank the importance of each paper node. Our ranking model is based on PageRank while each node is associated with a weight integrating paper-related features including the publication time, venue, and authors and their affiliations. The result of the WSDM Cup Ranker challenge show that our proposed model not only eliminates the time bias omitted by traditional PageRank but also provides an effective paper importance ranking.

Keywords

WSDM Cup; PageRank; Paper Ranking

1. INTRODUCTION

The challenge in the 2016 WSDM Cup is to rank the importance of scholarly articles in Microsoft Academic Graph, which includes a large-scale paper citation network and rich information of the papers. To show the rank result, each paper should be associated with a probability score, the higher the more important it is.

To rank the importance of each node in a graph, PageRank [5], which is originally designed for ranking webpages in the World Wide Web, is widely used for various applications [1, 3, 4, 6, 7]. According to the basic concept of PageRank, a webpage is given a high score if it is linked by either a large number of other pages or a highly scored page. However, we find there will be a problem if applying the same idea in our citation networks. Since a paper will be cited only by papers published at a later time, an early published paper will be easier to get more citations compared to the lately published one, or to get more citations from papers that have more citations. As a result, given two papers with the same number of citations, PageRank may give a higher score to the paper that publishes earlier while in fact the importance of the later one is not necessarily lower.

To diminish the time bias that is ignored by traditional PageRank, in this competition, we propose to use a time-aware weighted pagerank model, which is based on PageRank while associating each paper with a weight taking both the time factor and the rich paper information into consideration at the same time. Initially, each paper is assigned a time-dependent weight. Then, we further configure the weight by integrating features including paper venue, authors, and the author corresponding affiliations. The corresponding weight of the paper can represent our prior knowledge of its importance.

The remainder of the paper is organized as follows. We introduce our proposed solution in Section 2, followed by the corresponding experiments in Section 3. Section 4 concludes the paper and provides future work.

2. METHODOLOGY

In this section, we first introduce our newly designed time-aware weighted PageRank model, followed by the weight configuration of each paper.

2.1 Weighted PageRank

Our weighted PageRank is defined as follows. Let $PR(p)$ be the PageRank score of paper p. It is iteratively updated using the formula below.

$$PR(p) = (1-d)\frac{W(p)}{\sum_p W(p)} + d \sum_{q \in \mathrm{Pred}(p)} \frac{PR(q)W(p)}{\sum_{r \in \mathrm{Suc}(q)} W(r)}, \tag{1}$$

where $W(p)$ is the weight of paper p, d is the damping factor, $\mathrm{Pred}(p)$ is the set of predecessors of p (papers citing p), and

[*]{chinchi, miyen}@iis.sinica.edu.tw
[†]{r04922139, r04922009, r04922007, r04922050, sdlin}@ntu.edu.tw
[‡]b00902001@ntu.edu.tw
[§]b02901078@ntu.edu.tw
[¶]mftsai@cs.nccu.edu.tw

WSDM'16, February 22–25, 2016, San Francisco, CA, USA.
© 2016 ACM. ISBN 978-1-4503-3716-8/16/02…$15.00
DOI: http://dx.doi.org/XXXX.XXXX

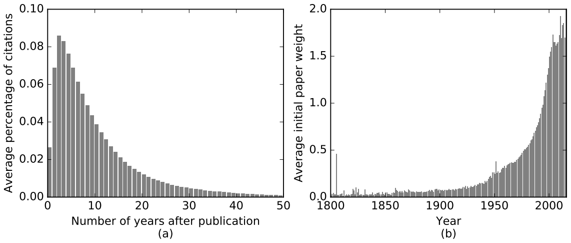

Figure 1: (a) The average percentage of citations a paper receives after publication. (b) The initial paper weight distribution over years.

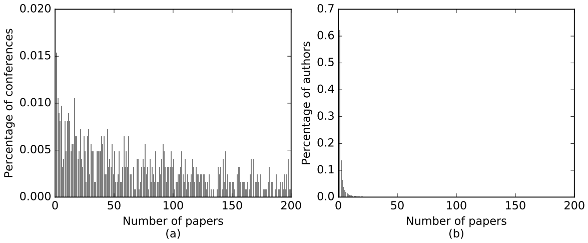

Figure 2: The two charts count how many percentage of (a) conferences and (b) authors are relevant to different numbers of papers. The distribution of journals to paper numbers is similar to that of conferences and the distribution of affiliations to paper numbers is similar to that of authors.

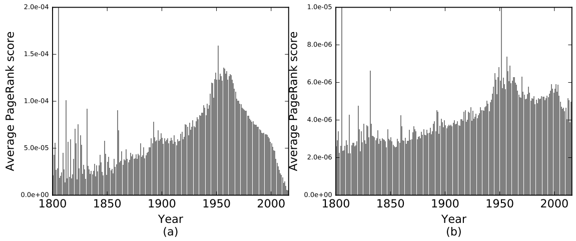

Figure 3: The average score distribution over years for (a) traditional PageRank and (b) weighted PageRank with complete paper weights.

$\mathrm{Suc}(q)$ is the set of successors of q (papers cited by q). The original PageRank score is a special case of our weighted PageRank if we set $W(p) = 1 \; \forall \; p$.

In Eq. (1), we ensure $W(p) > 0 \; \forall \; p$ to guarantee the convergence of the PageRank update process. Theoretically, PageRank can be formulated as a Markov chain model. Then we see the scores PR as the probability distribution of stopping states in a Markov process. To converge to a unique stationary distribution after repeated update process, *Irreducibility* is one of the conditions, which requires that there must exist a direct path from a state to any other one in the state graph. If $W(\bar{p}) = 0$ for some paper $\bar{p}$, then we obtain $PR(\bar{p}) = 0$ by Eq. (1). In other words, there is no path from $\bar{p}$ to another paper, or from another paper to $\bar{p}$, which violates the irreducibility condition.

Note that previous models have also applied the weight concept in computing the PageRank score. However, they considered only the first term [2] or the second term [8] of Eq. (1). However, our model takes both weights into consideration. To our knowledge, we are the first to propose this variant of weighted PageRank.

2.2 Weight Configuration

The essence of our proposed weighted PageRank lies in the weight design. It is composed of two parts. The first part is related to the citation relationship and the publication time. The second part is related to other paper-related features such as paper venue, authors and their affiliations.

2.2.1 Initial Paper Weights

We discover that the average number of citations per year works well especially. For each paper p, we define its weight $W_0(p)$ as follows:

$$W_0(p) \equiv \begin{cases} \frac{|\mathrm{Pred}(p)|}{\max\limits_{p \in P} Y(p) + 1 - Y(p)} & \text{if } |\mathrm{Pred}(p)| > 0 \\ \epsilon & \text{otherwise} \end{cases}, \qquad (2)$$

where P is the set of all papers, $|\mathrm{Pred}(p)|$ is the in-degree (the number of citations) of p in the citation network, and $Y(p)$ is the publication year of p. Recall that we have mentioned to ensure the positive weight values in Section 2.1 for convergence. In response, we assign a very small positive value ϵ for those papers without any citation. Last but not least, for the time factor we select only the publication year because it is the only temporal feature with no missing values in the given `Papers.txt`.

The rationale behind our weight design is as follows. As a rule of thumb, we use in-degree as an indicator to evaluate the importance of a paper. We put it to the weight as prior knowledge of the PageRank model as Xing and Ghorbani [8] did. In addition, we have the following observation: Most citations to a paper p appear in the first few years after the publication of p and the generation speed of new citations will decrease afterwards. Therefore, given two papers with different publication time, the newer one is more likely to have a higher average number of citations per year. In other words, our weight configuration tends to score newer papers higher. We draw the citation distribution over the past years for the given citation network, as shown in Figure 1. The figure surely implies the reduction of the average number of citations for a paper. Coincidentally, Sayyadi and Getoor [6] reports a similar distribution for high energy physics publications. We also show the weight distribution over years in Figure 1. Newer papers are clearly given higher weights on average. In conclusion, the traditional PageRank scores older papers higher while our weighted version highly evaluates newer papers given the same average number of citations per year. Therefore, the time-aware weighted PageRank is thus better than the traditional version because of the cancellation of the two paper scoring preferences.

2.2.2 Complete Paper Weights

We further improve the weight by considering the following paper features, the paper venue (conferences or journals), authors, and author affiliations. Intuitively, if a paper is published in a top conference or a top journal, or written by a famous author, or reported by a well-known affiliation, then we could be internally convinced of the quality. We first define how to compute the weight of a conference c, a journal j, an author a, and an affiliation f. The weights of these features are derived from the initial paper weight $W_0(p)$ as follows.

$$W(c) \equiv \frac{1}{|P(c)|} \sum_{p \in P(c)} W_0(p), \tag{3}$$

$$W(j) \equiv \frac{1}{|P(j)|} \sum_{p \in P(j)} W_0(p), \tag{4}$$

$$W(a) \equiv \frac{1}{|P(a)|} \sum_{p \in P(a)} \frac{W_0(p) + W(c) + W(j)}{|A(p)|}, \text{ and} \tag{5}$$

$$W(f) \equiv \frac{1}{|P(f)|} \sum_{p \in P(f)} \frac{W_0(p) + W(c) + W(j)}{|F(p)|}, \tag{6}$$

where $P(c)$, $P(j)$, $P(a)$, and $P(f)$ are the sets of papers having the corresponding feature value c, j, a, f. Also, $A(p)$ is the set of authors writing p and $F(p)$ is the set of affiliations of $A(p)$. We should first compute $W(c)$ and $W(j)$ using $W_0(p)$ and then calculate $W(a)$ and $W(f)$ using $W_0(p), W(c)$ and $W(j)$. A few experiment results support that $W_0(p) + W(c) + W(j)$ should be divided by the number of authors $|A(p)|$ or the number of affiliations $|F(p)|$ of paper p before we calculate their mean values.

Now, for a given paper p, let $c(p)$ and $j(p)$ be the conference or journal that paper p is published. Then, we have the complete weight configuration of p as follows.

$$\begin{aligned} W(p) \equiv\ & W_0(p) + W(c(p)) + W(j(p)) \\ & + \sum_{a \in A(p)} \frac{W(a)}{|A(p)|} + \sum_{f \in F(p)} \frac{W(f)}{|F(p)|}, \end{aligned} \tag{7}$$

where $W(c(p)), W(j(p)), W(a), W(f)$ are defined in Eq. (3)-(6), respectively.

Note that some feature values may be missing for a given paper. For example, the same paper cannot be published in both conference and journal so one of this value must be missing. In addition, sometimes these values may just be missing because the imperfectness of the given data. In this case, we substitute the missing value with the average value of that feature (but not just zero) in the given dataset. To be more specific, we use an example, where the conference value of paper p is missing, to explain how the substitution is done. Suppose in the given dataset we observe three unique conferences, say c_1,c_2, and c_3, in total. Then, the $W(c(p))$ term in Eq. (7) will be assigned as $W(\bar{c}) = \frac{1}{3}\sum_{i=1}^{3} W(c_i)$ as a pseudo conference weight of p. The treatment of missing values of other features are the same. The results on the leaderboard also support that we should fill the mean value instead of just 0.

Here we further justify the design of our weight configuration. Our empirical tests conclude that we have better to compute the weights of authors $W(a)$ and affiliations $W(f)$ using the features of conference and journal. Otherwise, the addition of author and affiliation features is helpless for PageRank. We draw Figure 2 as an explanation. Due to page limits, we report only the distributions of conferences and authors to the number of papers. There is high percentage of authors writing only one paper. In contrast, a conference often publishes more than one papers. We believe that the author weight $W(a)$ will not be accurate if it is obtained from only one paper weight. Therefore, we define $W(a)$ additionally with the conference weight $W(c)$ and the journal weight $W(j)$. $W(a)$ could be influenced by not only the papers p written by author a, but also by other papers of the same conference and journal as p.

Table 1: Statistics of G_0, G_1, and G_2. In fact, G_1 is actually the giant component of the sample graph, but we find that there is no performance difference with finding the giant component.

	G_0	G_1	G_2
Number of nodes	50,011,348	3,245,737	25,326,487
Number of edges	757,462,733	47,006,967	162,131,793

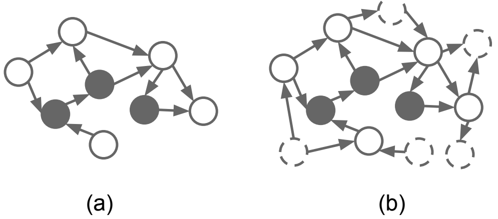

Figure 4: Illustration of two sample graphs for evaluation. The solid circles represent seed papers. (a) G_1 is the reduced subgraph composed of the seed papers along with the set of their 1-hop in- and out-neighbors (hollow circles), and (b) G_2 is the reduced subgraph composed of the seed papers and their 2-hop in and out neighbors (dashed circles) in the given citation network.

We show the effect of whether to use the weights or not for PageRank in Figure 3. The scores generated by traditional PageRank are very time-biased. On average, papers in 1950s is commonly judged important than those in 2000s. Traditional PageRank hence fails to compare two papers of different decades. We notice that the highest average PageRank score appears in 1950s, not the earliest 1800s. That is because there are only 1.7% of papers, 1.2% of citations before 1950s. We also present the change after we assign a weight to each paper. There is no large time bias among papers, especially for the period from 1950s to 2000s. The above results show that our proposed weighted PageRank enables us to determine the ranking of paper importance by eliminating the dependence on time.

3. EXPERIMENTS

3.1 Generating Sample Graphs

There are about 50 million papers with more than 750 million citations in the given file `PaperReferences.txt`. Apparently, it is very time consuming to just read this citation network into memory. If we can samle a small but representative subgraph from this large dataset, it will be more efficient to evaluate our new idea.

In order to find a good sample graph, we carefully examine the rules and the given data sets. First, on the Rules page[1] of WSDM cup we read a sentence in the section of Evaluation Measures: “A group of Computer Science researchers

[1] https://wsdmcupchallenge.azurewebsites.net/Home/Rules

Table 2: Results reported from the online leaderboard. Row 2 to 7 are the results of our weighted PageRank with different weight configurations.

Model	G_1	G_2	G_0
Traditional PageRank	0.687	0.685	0.683
$W(p) \equiv \lvert \text{Pred}(p) \rvert$	0.695	0.695	0.691
$W(p) \equiv W_0(p)$	0.709	0.715	0.713
$W(p) \equiv W_0(p) + W(c) = W_1(p)$	0.729	0.733	0.731
$W(p) \equiv W_1(p) + W(j) = W_2(p)$	0.721	0.727	0.731
$W(p) \equiv W_2(p) + \sum_{a \in A(p)} \frac{W(a)}{\lvert A(p) \rvert} = W_3(p)$	0.733	0.747	0.745
$W(p) \equiv W_3(p) + \sum_{f \in F(p)} \frac{W(f)}{\lvert F(p) \rvert} = W_4(p)$	0.733	0.747	0.745

are invited by the organizers to conduct pairwise ranking of papers in the fields they actively conduct research." This sentence implies that most of the labeled papers should be in the Computer Science field. Second, by observing the conference names in `ConferenceSeries.txt`, we found that all the conferences are likely to be in the Computer Science (CS) field (although`Journals.txt` contains other academic domains).

Based on the above two observations, we first choose the CS-related conferences from the conferences listed in `ConferenceSeries.txt`. Then, we select the papers published in these conferences from the `Papers.txt` file. There are 500 thousand papers selected in total as our *seed papers*. Based on these seed papers, we include the papers cite them and the papers they cite in the given citation network. Usually a paper tends to cite papers in the same field. Therefore, by walking from the seed papers in the citation network, we are able to catch almost all CS-related papers even if we do not know whether a paper is published in some CS-related journals or if the venue feature of a paper is missing. In this way, we generate two different sample graphs, $G_1 and G_2$, as illustrated in Figure 4. In essence, G_1 is the graph of seed papers along with their one-hop neighbors and G_2 further includes the two-hop neighbors in the citation network, respectively. For ease of exposition, we denote the originally given citation network by G_0. The statistics of the three graphs are shown in Table 1.

3.2 Results

We answer the following two questions to show the effectiveness of our proposed time-aware weighted PageRank: (1)Does weighted PageRank outperform PageRank? and (2) Does the weight configutation work as expected?

Table 2 demonstrates that time-aware weighted PageRank significantly raises our score on the leaderboard. We evaluate different weight configurations of our weighted PageRank on the leaderboard. Through these experiments, the damping factor $d = 0.5$ returns the most positive results, as claimed in [1, 3, 4, 7].

From the experiment results, we have the following remarks. First, we discover that the two sample graphs effectively capture most of the labeled papers in the original citation network. In addition, our time-aware weighted PageRank keeps stable performance for all the tree graphs. The effectiveness of our sample graphs further implies that the possibility of our model overfits a single graph is very low. Second, our time-aware weighted PageRank with complete weight configuration outperforms the traditional PageRank significantly. Most of the weight configuration proposals enhance the overall performance. Only two feature information, journal and affiliation, failed to help improve the ranking results. The performance of adding the journal information falls down on the two sample graph. However, further empirical results show that PageRank cannot gain such a large improvement when added authors and affiliations without journal information. Adding the journal feature itself does not work well, but the combination of multiple features can reach a better performance. When adding the affiliation information, all the reported scores remain the same without improvment. Nonetheless, we believe that the results deliver a positive message: Considering affiliations is likely to bring a more robust PageRank model to avoid potential overfitting to the Evaluation set on the leaderboard. Since adding the affiliation information does not degrade performance, we positively assume its helpfulness to predict the ranks in the hidden Test set in phase 2.

4. CONCLUSION

We have demonstrated the time-aware weighted PageRank-based model is effective and competitive for the paper importance ranking problem. The experience of this competition tells us that the citation relationship among papers still serves as an good indication of paper importance. The features shared among papers also reflect the implicit relationships among papers without citations. Despite the achievement in the competition, we have some questions left for future research. For instance, we are interested in exploring if there exist more theoretical reasons supporting the effectiveness of the proposed weight configuration. Also, we want to explore the possibility of fusing other paper features, such as title and study field, to further improve the PageRank performance. Last but not least, we seek to know if it is possible to learn the weights rather than defining them.

5. REFERENCES

[1] P. Chen, H. Xie, S. Maslov, and S. Redner. Finding Scientific Gems with Google. *ArXiv Physics e-prints*, Apr. 2006.

[2] Y. Ding. Applying weighted pagerank to author citation networks. *J. Am. Soc. Inf. Sci. Technol.*, 62(2):236–245, Feb. 2011.

[3] N. Ma, J. Guan, and Y. Zhao. Bringing pagerank to the citation analysis. *Inf. Process. Manage.*, 44(2):800–810, Mar. 2008.

[4] S. Maslov and S. Redner. Promise and Pitfalls of Extending Google's PageRank Algorithm to Citation Networks. *ArXiv e-prints*, Jan. 2009.

[5] L. Page, S. Brin, R. Motwani, and T. Winograd. Technical Report 1999-66, November.

[6] H. Sayyadi and L. Getoor. Future rank: Ranking scientific articles by predicting their future pagerank. In *SDM'09*.

[7] D. Walker, H. Xie, K.-K. Yan, and S. Maslov. Ranking scientific publications using a model of network traffic. *Journal of Statistical Mechanics: Theory and Experiment*, 6:10, June 2007.

[8] W. Xing and A. Ghorbani. Weighted pagerank algorithm. In *Communication Networks and Services Research, 2004. Proceedings. Second Annual Conference on*, pages 305–314, May 2004.