ISBN 978-981-11-7861-0
Proceedings of 2018 the 8th International Workshop on Computer Science and Engineering (WCSE 2018)
Bangkok, 28-30 June, 2018, pp. 136-140
doi: 10.18178/wcse.2018.06.024

A Weighted PageRank for Scientific Paper Ranking

Xiao Liu [1 +]

[1] College of Electronics and Information Engineering, Tongji University, Shanghai 201800, China

Abstract. Ranking scientific papers is a challenging but important task. This paper focuses on three issues: (1) how to use publication time information of papers to capture the dynamics of an evolving scientific literature network; (2) how to use metadata information of papers to get better ranking results; (3) how to use topic information to more accurately rank a paper with a specific topic. In response to these problems, we propose a weighted PageRank algorithm which uses citations, publication time, topics, authors, venues and other relevant information collaboratively. We conduct experiments on two public datasets. The results show that our method ranks scientific papers more accurately than existing methods and topic-based PageRank vectors can produce more accurate rankings than a global PageRank vector.

Keywords: PageRank, Topic sensitive ranking, Heterogeneous networks, Topic-based rank

1. Introduction

Ranking scientific papers is a very challenging task, partly due to the diversity of research topics and partly due to the dynamic nature of the citation network. This paper focuses on three issues. First, we study how to use time information in the citation network to get better ranking results. Second, we study how to leverage different kinds of information simultaneously to improve paper ranking results. Third, we study how to generate topic-based rankings.

For the first issue, we use two time-related strategies. First, we use a personalized PageRank vector, which promotes higher scores to newer papers. Second, we develop time-aware weights for the edges of the citation network.

To address the second issue, we reweight the edges of the citation network using the metadata of a paper such as author of the paper and venue (conferences or journals) of the paper. Since journals and authors can provide additional information for papers, a more accurate ranking list can be generated, especially for newly published papers with only a few references.

For the third issue, we calculate a set of topic-based PageRank vectors by using a series of representative topics. Each topic-based PageRank vector can capture the notion of importance for a particular topic more accurately.

We conduct experiments on two public-available datasets. One is Cora dataset, which contains papers in the field of computer science. The other one is Aminer dataset, which includes papers in various research fields. The results show that compared with other methods, the proposed algorithm can generate more accurate rankings.

2. Related Work

Ranking scientific papers is a challenging and important task. Page introduced the PageRank algorithm to estimate the academic impact of the paper[1]. After that, many researchers applied the PageRank algorithm on the citation network to rank scientific papers[2][3][4]. Their results confirm that the number of

[+] Corresponding author. Tel.: +8613120933982
E-mail address: 0xiaoliu@tongji.edu.cn

citations of papers shows popularity and PageRank finds the prestige of papers. However, the results of PageRank are biased toward older papers because it does not capture the dynamics of the citation network. For recent papers, few references can be found, so they are underestimated by the PageRank algorithm.

To get better paper ranking results, some researchers tried to use come additional information related to papers. Walker proposed a method called CiteRank[5], which takes into account publication time information and uses a random walk process to rank papers by predicting the number of future references. This model promotes recent papers to higher scores in order to reduce the time-bias problem.

Sayyadi used the same strategy of CiteRank and introduced a method called FutureRank[6]. Besides publication time information, it also takes into account author information of papers. In FutureRank, the use of authorship provides additional information when ranking recent publications. Zhou also found that using the author information can rank the authors and their publications more accurately[7].

A Topic-sensitive ranking is more meaningful than a global ranking because researchers usually only interested in the areas that they focus on. Jardine proposed a model called TPR[8], which extracts papers' topics by applying LDA(Latent Dirichlet Allocation)[9] to analyze textual content. They used the topic information to weight the transition matrix of PageRank to get the topic-sensitive ranking. Zhang introduced a model called CTPM[10] that adopts the correlation topic model[11] instead of LDA to extract papers' topics. Papers' venue information and topic information are used to weight the transition matrix of PageRank in CTPM.

3. Our Proposed Model

3.1. Weighted PageRank

Let $\mathbf{r}$ be the PageRank vector of all the papers. $\mathbf{r}$ can be iteratively updated using the equation below.

$$\mathbf{r} = \mathbf{s} + (1-\alpha)\mathbf{M}\cdot\mathbf{r} + (1-\alpha)^2\mathbf{M}^2\cdot\mathbf{r} + \cdots (1-\alpha)^k\mathbf{M}^k\cdot\mathbf{r} \tag{1}$$

where $\mathbf{s}$ is a personalized PageRank vector, $\mathbf{M}$ is the transition matrix of the PageRank algorithm and α is a pre-defined constant. Let s_i be the value of the ith paper p_i in $\mathbf{s}$. We set s_i to s_i=exp(-$\tau(Y_{\max} - Y_i)$), where $(Y_{\max} - Y_i)$ is the age of the i-th paper and τ is a constant[5][6]. The intuition of this setting is that most of the papers' citations appear in the first few years after their publication and the generation speed of new references will decrease after that. Therefore, we promote higher scores to newer papers.

The CiteRank algorithm is a special case of our weighted PageRank if we set $\mathbf{M}_{ij} = 1/k_i^{out}$ if j-th paper cites i-th paper and 0 otherwise. Our weighted transition matrix is defined as follows.

$$\mathbf{M}_{ij} = w(p_i) / \sum_j w(p_j) \tag{2}$$

where $w(p_i)$ is the designed weight for each paper p_i. We define $w(p_i)$ as follow.

$$w(p_i) = \begin{cases} |\text{Pred}(p_i)|/(Y_{\max} - Y_i + 1) & \text{if } |\text{Pred}(p_i)| > 0 \\ \varepsilon & \text{otherwise} \end{cases} \tag{3}$$

where Pred(p_i) is the set of predecessors of p_i in the citation network(papers citing p_i) and ε is a small constant. We set ε=10^{-6} in this paper. The intuition behind our weight configuration is that a new paper is inspired mostly by the papers with high weights. In other words, our weight configuration tends to propagate more scores to the recently published and highly cited papers.

A lot of literatures also use author and venue information to design the weights of papers[8][10][12]. The intuition is that a paper is more likely to be of high quality if this paper is written by a well-known author, or published in a top-level venue, or reported by a famous affiliation.

We adopt the way used in [12]. Let $w_0(p_i)$ be the initial weight of paper p_i which is defined in equation (3). We define the weight of venue v_i as below.

$$w(v_i) = \frac{\sum_{p\in P(v_i)} w_0(p)}{|P(v_i)|} \tag{4}$$

where $P(v_i)$ is the set of papers that are published in venue v_i. The weight of venue v_i is the average of weights of the papers that are published in it. Then, we define the weight of author a_i as below.

$$w(a_i) = \frac{\sum_{p \in P(a_i)} (w_0(p) + w(v)) / |A(p)|}{|P(a_i)|} \tag{5}$$

where $P(a_i)$ is the set of papers that are written by author a_i, $A(p)$ is the set of authors who write paper p and $w(v)$ is the weight of the venue that paper p is published in. Finally, we can define the weight of paper p_i as the sum of the initial weight in equation (3), the weight of the venue that it is published in and the average of the weights of the authors who write it.

$$w(p_i) = w_0(p_i) + w(v) + \frac{\sum_{a \in A(p_i)} w(a)}{|A(p_i)|} \tag{6}$$

3.2. Topic-sensitive ranking

To generate the topic-sensitive rankings, we pre-compute a set of PageRank vectors for all topics. When calculating a PageRank vector for a topic c, we set the personalized PageRank vector as follow.

$$\mathbf{s}_c = \mathbf{s} \circ \mathbf{q}_c \tag{7}$$

where $\mathbf{q}$ is a column vector, $[pr(c|p_1); pr(c|p_2); \ldots]$, of which each element is the topic distribution of a paper and $\circ$ represents the element-wise multiplication. The topics of each paper can be extracted from the text content of the paper using the methods proposed in [8][10] The topic distribution of each paper can be calculated using the correlation topic model as proposed in [10]. Then, for each topic, we can generate a topic-based PageRank vector $\mathbf{r}_c$ using equation (1) and equation (7).

4. Experiments

4.1. Datasets

We evaluate our approach on two public-available datasets. One is Cora provided by McCallum. We remove all the papers that do not have publication time. We also change the name of conference to remove time information. For instance, “Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining (KDD)” will be renamed as “Proceedings of the ACM SIGKDD international conference on knowledge discovery and data mining (KDD)”. In all, we collect 16,459 papers, 15,748 authors, 7,597 journals/conferences and 52,634 citations in this dataset.

Another dataset is AMiner dataset, which is provided for 2017 Open Academic Data Challenge. We remove the papers which do not appear in the citation graph. In all, we have 973,513 papers, 654,051 authors, 6,550 journals/conferences and 5,429,945 citations in this dataset. In this dataset, 5,854 authors are labeled with their research interests. We view these labels as the research fields of their papers and get 119,632 papers with labels in total.

4.2. Experimental setup

We split the dataset into query set and test set. For Cora dataset, we split the dataset at the time point, 1996-01-01. For AMiner dataset, we split the dataset at the time point, 2012-01-01. For evaluation, we rank the papers according to the number of references on test set and use the result as ground truth ranking. Then, we apply the ranking algorithm on the query set to generate predicted ranking.

We use two metric to evaluate the ranking results. One is the spearmans rank correlation coefficient. The other is the average precision. We compute the average precision for the top k papers in the ground truth ranking.

We evaluate various methods for comparison: (1) pagerank[1]; (2) FutureRank[6]; (3) CiteRank[5]; (4) THRank[13]; (5) TPR[8]; (6) CTPM[10]. The random jumping probability is set to 0.15 for all algorithms. The damping factor is set to 0.85 for the algorithms which do not make use of author and journal information. k is set to 5 and α is set to 0.85 in our method.

We choose machine learning, computer vision, software engineering, information retrieval, distributed systems and wireless sensor network as labeled topics to evaluate the topic-sensitive ranking on AMiner dataset.

4.3. Results

Table 1: Global ranking results on two datasets

	Cora dataset			AMiner dataset		
Algorithm	correlation	AP@10	AP@100	correlation	AP@10	AP@100
PageRank	0.2427	0.2595	0.1307	0.3018	0.1208	0.0685
FutureRank	0.4060	0.2750	0.0422	0.4452	0.3000	0.2542
CiteRank	0.4136	0.5583	0.2149	**0.4772**	0.2750	0.2537
THRank	0.2383	0.3000	0.0400	0.3093	0.2600	0.1649
TPR	0.3520	0.0827	0.0241	0.3870	0.0167	0.0123
CTPM	0.1828	0.0000	0.0319	0.2330	0.0000	0.0000
Our Method	**0.4142**	**0.5750**	**0.2642**	0.4360	**0.3667**	**0.3018**

Table 2: The results of topic-sensitive rankings on 6 different topics

	Machine Learning			Computer Vision		
Algorithm	correlation	AP@10	AP@100	correlation	AP@10	AP@100
THRank	0.2618	0.4248	0.4042	0.2942	0.4633	0.3200
CTPM	0.3415	0.0693	0.0627	0.3431	0.0111	0.0421
Our Method-g	0.3436	0.0000	0.0491	0.3323	0.0000	0.0386
Our Method-t	**0.6141**	**0.6778**	**0.6292**	0.6180	**0.5490**	**0.5345**
	Software Engineering			Information Retrieval		
Algorithm	correlation	AP@10	AP@100	correlation	AP@10	AP@100
THRank	0.1857	0.2796	0.0845	0.2378	0.5600	0.2367
CTPM	0.2753	0.0000	0.0278	0.3107	0.0367	0.0678
Our Method-g	0.2456	0.0786	0.0542	0.3293	0.0536	0.0305
Our Method-t	**0.5666**	**0.6700**	**0.3364**	0.6228	**0.7889**	**0.5238**
	Distributed Systems			Wireless Sensor Network		
Algorithm	correlation	AP@10	AP@100	correlation	AP@10	AP@100
THRank	0.1783	0.2900	0.1556	0.1147	**0.3648**	0.2299
CTPM	0.2530	0.0000	0.0319	0.2849	0.0347	0.1593
Our Method-g	0.2904	0.0667	0.0345	0.3429	0.1000	0.1052
Our Method-t	**0.5742**	**0.5857**	**0.4050**	**0.5374**	0.2171	**0.5030**

Table 1 shows the results of the global ranking on two datasets. We can find that the results of our method and CiteRank are much better than the results of other methods on Cora dataset. And our method performs better than CiteRank when ranking the top k papers. On AMiner dataset, our method performs slightly worse than FutureRank and CiteRank when considering the spearmans rank correlation. However, the ranking of the top k papers of our method is better than that of FutureRank and CiteRank. In practise, we believe that the ranking of the top k papers is much more important that the ranking of all the papers.

The results of the topic-sensitive rankings on 6 different topics are shown in Table 2. In the table, "Our Method-g" represents using the global PageRank vector generated by our method to rank the papers. And "Our Method-t" represents using the corresponding topic-based PageRank vector to rank the papers. The proposed method significantly outperforms other topic-sensitive methods on most of the topics and the topic-based PageRank vectors generate more accurate results than the global PageRank vector.

5. Conclusion

In this paper, we propose a weighted PageRank algorithm for ranking scientific papers. We design the weight of the transition matrix of our PageRank algorithm by taking into account time information, author information and venue information. Moreover, we extend our method to support the topic-sensitive ranking. We evaluate our method on two public-available datasets. The experimental results show our method can generate better results than many existing methods and the topic-based PageRank vectors give more reasonable ranking results than the global PageRank vector.

6. References

[1] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank citation ranking: Bringing order to the web. *Technical report*. Stanford InfoLab, 1999.

[2] J. Bollen, M. Rodriquez, and H. V. Sompel. *Journal status. Scientometrics*. 2006, 69(3): 669-687.

[3] P.Chen, H. Xie, S. Maslov, and S. Redner. Finding scientific gems with Google. *Journal of Informetrics*. 2007, 1: 8-15.

[4] N. Ma, J. Guan, and Y. Zhao. Bringing pagerank to the citation analysis. *Information Processing & Management*. 2008, 44(2):800-810.

[5] 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*. 2007(06): 6-10.

[6] H. Sayyadi, and L. Getoor. Futurerank: Ranking scientific articles by predicting their future pagerank. *In Proceedings of the Ninth SIAM International Conference on Data Mining (SDM09)*. 2009.

[7] Y. B. Zhou, L. Lu, and M. Li. Quantifying the influence of scientists and their publications: distinguishing between prestige and popularity. *New Journal of Physics*. 2012, 14.

[8] J. G. Jardine, and S. Teufel. Topical PageRank: A model of scientific expertise for bibliographic search. *EACL*. 2014: 501-510.

[9] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*. 2003, 3: 993-1022.

[10] Y. Zhang, J. Ma, Z. Wang, B. Chen, and Y. Yu. Collective topical PageRank: a model to evaluate the topic-dependent academic impact of scientific papers. *Scientometrics*. 2018, 114: 1345-1372.

[11] D. M. Blei, and J. D. Lafferty. Correction: A correlated topic model of science. *Annals of Applied Statistics*. 2007, 1(2): 634-634.

[12] A. D. Wade, K. Wang, Y. Sun, and A. Gulli. *WSDM Cup 2016 – Entity Ranking Challenge. Proceedings of the 9th ACM Conference on Web Search and Data Mining*. San Francisco, CA. 2016.

[13] T. Amjad, Y. Ding, A. Daud, J. Xu, and V. Malic. Topic-based heterogeneous rank. *Scientometrics*. 2015, 104: 313-334.