TAPRank: A Time-aware Author Ranking Method in Heterogenous Networks

Xiangjie Kong[1,2], Jinmeng Zhou[1], Jun Zhang[1], Wei Wang[1], Feng Xia[1]
[1]*School of Software, Dalian University of Technology, Dalian 116620, China*
[2]*Key Laboratory of System Control and Information Processing, Ministry of Education, Shanghai 200240, China*
xjkong@ieee.org

Abstract—**Measuring the impact of authors can not only be a good guidance for new researchers, but also provide a standard for academic foundations and awards. Heterogeneous networks can capture more information about the interactions between entities and they are more and more widely used for the measurement of author impact. However, most of the existing researches take all the papers into the networks as equal, although they have different importance levels. In this paper, we propose a new model: TAPRank, which calculates author impact in author-paper network with considering the PageRank scores of papers for the first time. The PageRank algorithm is implemented in paper citation network, taking the time of publication of each paper into consideration. In addition, the experiments on DBLP dataset show a better performance of TAPRank than other state-of-the-art models.**

Index Terms—**Author impact, heterogeneous networks, pageRank algorithm**

1. Introduction

With the rapid evolution of modern science, more and more scholar data are produced in various academic activities. And more and more search engines and digital libraries are focused on these scholar data, which makes us get easy access to a great deal of academic information. However, it also makes exploring meaningful knowledge be a difficult task for academic researchers.

Researchers often need to know the most influential entities in a certain field at the beginning stage of their research careers. Finding experts with influential papers or a list of outstanding scholars, are particularly important to understand the development of the related areas. Besides, measuring the contribution of each scholar can give respect to the influential experts as well as provide a standard for academic awards. In this context, many ranking methods are proposed to measure the academic impact of scholars. These methods are mainly classified into two categories: citation based methods and network based methods.

Citation can be seen as a way that the authors of the citing paper give credit to the authors of the cited paper. Citation counts is the most traditional indicator of academic influence. On this basis, a lot of citation based index methods have been proposed to measure scholars' academic impact, like h index [1]. These methods consider all the citations as equal and do not take the quality of citations into account. Some researchers have argued that citations from papers or authors of greater significance should be more important and highly weighted [2].

The entities' positions in the networks may greatly affect the measurement of their impact. A large number of researchers also attempt to qualify and rank the importance of scholars based on the topological structure of academic networks. The typical two kinds of author networks are coauthorship networks which means that there exit bilateral links between authors with common papers and citation networks which means links in the network point to cited authors from citing authors. Two authors can also related in other forms such as they are cited together by other works, or they have common references in in their oeuvre [3]. As for the method, centrality measurements are widely used to evaluate scholars' academic impact, including Katz-bonacich [4], Eigenvector [5] and some hybrid centrality measures [6] which combine two or more centrality measures to better understand the position and the importance of the entities in the network. PageRank [7] algorithm is another network based method. It can compute a globe ranking result for all the nodes in the network, which is very fit to calculate scholars' scientific impact. The extensional PageRank algorithms mainly focus on developing new weighted forms of the nodes [8], [9] and constructing novel relationships between the nodes [10].

Both centrality measures and PageRank algorithm are constructed in homogenous networks. That is, there is only one single type of nodes and one single type of relations in a network. While researchers have proved that the employment of hybrid or heterogeneous networks has a better performance with capturing the complex research communications and interactions [11]. All the entities such as authors, papers, and venues can be a part of a network [12]. For example, authors can be related to the papers they have written, and papers can have links to the venues or terms they belong to. In the network, authors and papers reinforce each other on the impact scores [13]. Highly ranked authors tend to publish more highly ranked papers, and the high scores of papers affect the impact of their authors in return. In similar, the published venues and the citations also have the same effect on author ranks through the connection of papers. Therefore, many recent studies devote to measure scholar impact in heterogenous networks

[14], [15]. These methods are trying to define more types of nodes and relations in one network. They contribute a lot to the researches of academic impact.

When calculating author impact in heterogenous networks, current researches take all the nodes as equal at the beginning of the process. However, the papers are of different importance and their starting values should be distinguished as a matter of course. In this paper, we firstly propose to value the papers with their PageRank scores, which are obtained in paper citation network.

Another point that we focus on is the publishing time of the papers. When calculating the PageRank scores of the papers in citation network, recently published papers generally obtain lower scores because the growth of citations is a gradual process of accumulation. On the contrary, people are less concerned about the past information and new researchers typically start studies from rather recent publications [16]. In another word, authors with recent publications and authors cited by recent publications attract more attention. So the publishing time of the papers should be taken as a factor to give higher values to recent papers and their references in the PageRank algorithm.

In our paper, we present an author ranking algorithm named TAPRank, which is a time-aware model based on the author-paper network. We make the following contributions:

- We distinguish the papers by the PageRank scores before the random walk process in a heterogeneous network for the first time.
- A time-aware function is taken into consideration in the course of PageRank algorithm and gives higher values to recently published papers as well as their references.
- We perform the experiments on DBLP dataset and confirm the most suitable time function whose values decay with the age of publications.

The rest of this paper is organized as follows: in the next section, we describe the proposed TAPRank Model. The third section presents the evaluation by the experiment results on DBLP dataset. At last, we expound the conclusion and the future work.

2. Methods

As heterogenous networks have been proved to be reliable for measuring author impact, our proposed method is based on the paper-author network. Most of the current researches focusing on paper-author network take all the nodes as equal at the beginning of the impact measuring process. However, papers are of different importance in reality. We can get such importance scores of papers through a PageRank model at first. So there are two steps in our TAPRank model:

(1) Calculate the time-aware PageRank scores of all the papers in paper citation network. As we have mentioned above, people are more concerned about recent information and new researchers typically start studies from rather recent publications. Scholars with high influence at present catch more attention than those before so that they have more power to lead the trend of academic researches. As well, when measuring academic achievement to determine the awards and foundations, recent influence should be given more attention. So we take a time-aware function into consideration in the course of paper citation PageRank method to give higher values to the recently published papers and their references.

(2) Get the rank result of author impact through a random walk process between authors and papers. We construct an author-paper network in which papers are distinguished by the PageRank scores and in the random walk process the scores of papers have an effect on the impact scores of their authors. In this way, we can take notice of the activities of the authors by the papers linked to them.

In this section, we will expound the model in detail.

2.1. PageRank in Paper Citation Network

PageRank is an algorithm used to order the importance of each node in a network from an iterative process. The node would have high rank score if the nodes pointing to it have high scores. The following equation can well present the mechanism of PageRank:

$$PR(p_i) = \frac{1-d}{N} + d\sum_{j=1}^{m} \frac{PR(p_j)}{L(p_j)} \qquad (1)$$

where p_i is one of the nodes in the network and N is the total number of the nodes, p_j is the node that links to p_i, $L(p_j)$ is the sum of outgoing links of p_j. $PR(p_i)$ is the visiting probability of node p_i, which represents the influence of p_i, and the same for $PR(p_j)$. d is the damping factor (set as 0.85 as in [7]), which is the probability that the node p_i is visited following the links pointing to it. The remain $1-d$ is the probability that the visit does not follow the links but randomly choose p_i as the node. In every iteration of PageRank, we use Eq. (1) to update the PageRank score of each node. At last the scores of all the nodes converge to a stable state. Then we can get the rank result of them.

We use the PageRank algorithm to obtain the influence scores of the papers in paper citation network. Paper citation network can be denoted by $G_p = (V_P, E_P)$, where V_P is the set of papers and E_P is the set of citation relations. There exists a directed edge when a paper cites another. But it is obvious that newly published papers get less citations than those who have been published for a long period of time, which causes the unfairness of new publications in the citation network. PageRank without considering the time of publications can not notice this feature. So we develop a time-aware PageRank algorithm.

2.2. Time-aware PageRank

Figure 1 shows the average number of citations a paper obtains in each year within DBLP dataset. In the year the paper published (past years = 0), a paper got fewer citations because it needs time for this paper to diffuse the influence.

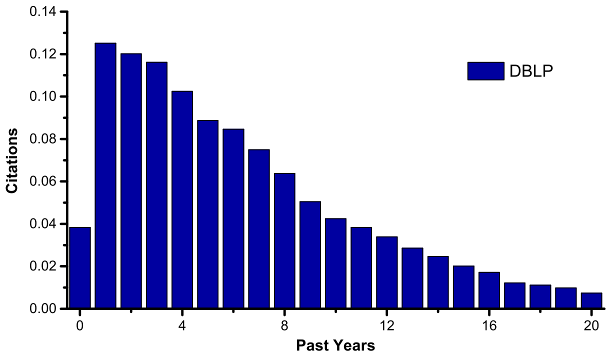

Figure 1: Average Number of Citations

The maximum number of citations a paper got occurs in the year after publication (past years = 1), and the citations in each individual year sharply decrease. Therefore, we think the newer the paper is, the more citations it will obtain in the future. As well, it will attract more attention to its references.

But on the contrary, newly published papers certainly get less impact scores in the paper citation network, because the growth of citations is a gradual process of accumulation. It is unfair for new coming researchers and newly published papers. While the aggregate citation counts increases, the citations that an article obtains in each individual year generally decrease, which represents the attention the authors get through their papers is in the same situation. And recent citations to a paper can also refresh the attention to its authors. We think authors with recent publications and citations are more active and should get more academic impact in the research area.

Under this theory, a time-aware function $f(age)$ is added to our time weighted PageRank. It can be expressed as Eq. (2):

$$\begin{aligned} PR(p_i) = (1-d)\frac{f(age_i)}{\sum_{k=1}^{N} f(age_k)} \\ + d\sum_{j=1}^{m}\frac{PR(p_j) * f(age_j)}{L(p_j)} \end{aligned} \tag{2}$$

where age_i is the age of paper p_i after publication and the same with paper p_j. Let T_c be the current time of the year and T_{p-i} represents the publication year of p_i. We get the age of a paper p_i as Eq. (3):

$$age_i = T_c - T_{p-i} \tag{3}$$

$f(age)$ is a function whose values decay with the increase of age and the values range from 0 to 1. We present the $f(age)$ function which conforms to the requirement as Eq. (4):

$$f(age) = exp^{-\alpha * age} \tag{4}$$

We can see that in the first part of Eq. (2), $\sum_{k=1}^{N} f(age_k)$ is the sum of the values assigned to each node by the function of age. Nodes with high values of $f(age)$ will get higher chance to be visited. And in the second part of Eq. (2), the score that p_i gains is more greatly affected by newly published citations. Then with the effect of $f(age)$, recently published papers can get a chance to higher scores and a paper can get more weight from recently published papers that cited it.

As we have introduced, instead of considering each article as being of equal weight, this time-aware PageRank method gives more attention to recently published papers and expands to their references. We value the papers in the author-paper network by the scores calculated by this time-aware PageRank algorithm. Then we can calculate author impact through a random walk process between the authors and the papers.

2.3. Random Walk Process in Author-Paper Network

In the author-paper network, authors and the papers they have written are interconnected. There are two rules when we calculate the author impact in this network:

Rule 1: The score of an author is influenced by the scores of all the papers he has written.

Rule 2: The score of a paper is influenced by the scores of its authors.

With these two rules, papers and authors reinforce each other: the reputations of authors depend on the quality of papers they have written; Authors with high reputations tend to write high quality papers. Our random walk process is constructed in the author-paper network under these two rules. It is implemented as the following steps:

Step 1: In the author-paper network, we equally set the initial score of each author as $1/N_a$, where N_a is the total number of authors. And the papers take the time-aware PageRank scores as their initial values.

Step 2: Update the author scores through their papers by Eq. (5):

$$RW(a_i) = \frac{1-d}{N_a} + d\sum_{j=1}^{m}\frac{RW(p_j)}{C(p_j)} \tag{5}$$

where $RW(a_i)$ is the random walk score of author a_i, $RW(p_j)$ is the score of paper p_j published by a_i, N_a is the author number in the network, $C(p_j)$ is the total numbers of authors of p_j and d is the damping factor. The authors have $1-d$ probability to be randomly visited and d probability to be visited following their papers.

Step 3: Update the paper scores through their authors by Eq. (6):

$$RW(p_j) = \frac{1-d}{N_p} + d\sum_{i=1}^{n}\frac{RW(a_i)}{C(a_i)} \tag{6}$$

where $RW(a_i)$, $RW(p_j)$ have the same meaning as in Eq. (5), and N_p is the number of papers in the network, $C(a_i)$ is the total number of papers written by author a_i. d is the damping factor which has the same value as in Eq. (5).

Step 4: Repeat step 2 and step 3 until the random walk scores of authors are converged. The condition is that the sum of the difference between the scores for all the authors computed at two successive iterations falls below a threshold.

After these four steps, we can get the rank scores for all the authors and order them by their influence.

3. Experiments

In this section, we apply our TAPRank model to the experiments and determine the suitable parameter α for the time function of the model. Besides, we evaluate its performance by the comparison with some other models.

3.1. Dataset and Models

We do the experiments on DBLP dataset. DBLP is a computer science bibliography on major computer science journals and proceedings. We download the DBLP dataset from Aminer [17], with citation information captured from ACM Digital Library. The dataset contains 23365 papers from 1990 to 2013 and covers two domains (Data Mining and Artificial Intelligence).

The models used for comparison are as follows:

a. RWRank model. It is the basic random walk process between papers and related authors.

b. PRWRank model. It is the PageRank based method. PRWRank model takes the PageRank scores of papers into account at the beginning of the random walk process. The PageRank scores are calculated in paper citation network.

c. TAPRank model. This is our proposed model which values the papers by the results of the time-aware PageRank algorithm before the random walk process.

d. PAveRank model. It obtains an author score based on the average value of PageRank scores of his papers. This is a method presented in [18] which also measures author impact using the PageRank scores of papers.

We assume that the authors highly ranked by their impact will get more citations in the future. So we set the data before 2008 to the experiments. And the rank of authors by the citation counts in the recent 5 years (2009-2013) is taken as the test to verify the performance of each model, which named as FutCit rank.

3.2. Results

As we have assumed, the authors of high impact attract more attention at present and will get more citations in the future. Namely, top ranked authors in FutCit rank may have better rank results in our proposed method. In our experiments, we rank authors from lower positions. For example, order at 1 is a better result than at 100. We get the top 100 ranked authors by FutCit. Figure 2 depicts the boxplots of the rank positions of this 100 authors by TAPRank with different values of α. We can see the best, the median, and the worst rank values by different methods. When $\alpha = 0.3$, TAPRank gets the best result with the consideration of all the best, the median and the worst rank positions. So in the following of this paper, we use $\alpha = 0.3$ in our TAPRank model. And FutCit rank is also the ground truth.

TABLE 1: Common Members with FutCit Rank

TopN	RW	PRW	TAP	PAve
20	0	6	**7**	5
100	16	37	**45**	37
1000	317	412	**452**	385

As shown in Table 1, our proposed TAPRank model gets the most common members with FutCit when we take a look at the top lists of each method. There are 7 common members in the top 20 list, 45 common members in the top 100 list, and 452 in the top 1000 list.

Pearson Correlation Coefficient can calculate the correlation between two rank results. Its values vary from -1 to 1 with correlation ranging from the most negative to the most positive. The second column in Table 2 presents the Pearson Correlation Coefficient between the authors' scores in the FutCit rank and in each model. It is obvious that TAPRank makes a great improvement compared to the RWRank model and PRWRank model with a score of 0.63637. This is also a better result than PAveRank model.

TABLE 2: Comparison of Each Method

Methods	Correlation	AUC
RW	0.32743	0.61630
PRW	0.53512	0.72823
TAP	**0.63637**	**0.78436**
PAve	0.60751	0.72761

In addition, we draw the receiver operating characteristic (ROC) curves respectively. To each model, we take higher ranked authors as positive entities and lower ranked authors as negative entities from the rank result. And the FutCit rank is used to classify the sample. Authors with higher citation counts from 2009 to 2013 are more likely to be judged as positive. Using the result, Figure 3 shows the ROC curves for each model. In this figure, $1 - Specificity$ represents the false positive rate, while $Sensitivity$ represents the true positive rate. The more the curve closes to the top left corner, the better the rank result confirms to the FutCit rank. TAPRank performs better than other models in ROC curves. We can also discover the same trend from the area the curves cover (AUC), which means the accuracy for each model. We can see the AUC values in Table 2. TAPRank model gets the highest value of 0.78436. PAveRank model even gets a worse result than PRWRank model.

With the results of the experiments, we have proved the reliability of the TAPRank model from different aspects.

4. Conclusion

In this paper, we devote to measuring author impact based on heterogeneous networks. We propose the TAPRank model, which calculates author impact in author-paper network with papers valued through a time-aware PageRank algorithm. Higher scores are given to newly published papers

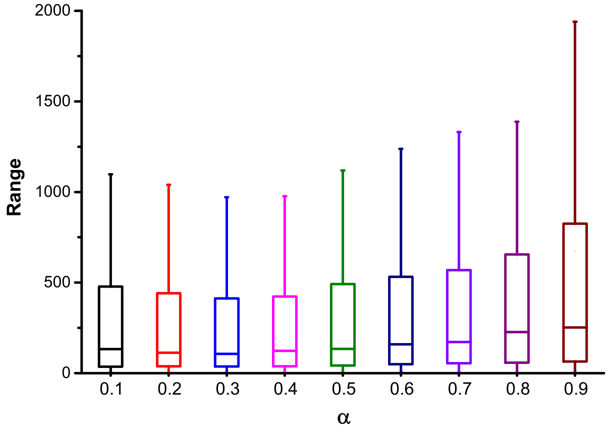

Figure 2: Boxplots of Relative Ranks for Top Authors

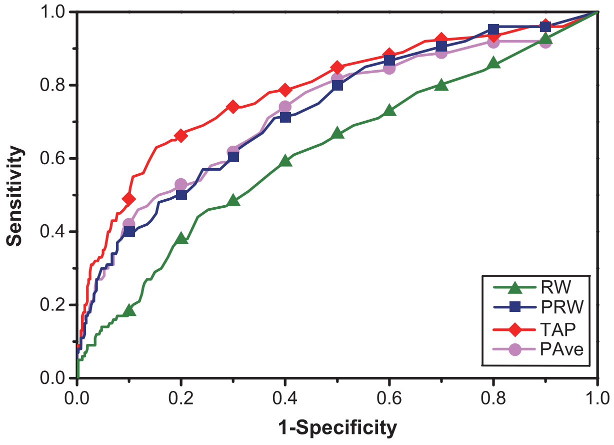

Figure 3: ROC Curves of Each Method

as well as their references by this PageRank algorithm. We constructed the experiments on DBLP dataset to confirm the time function used in the PageRank algorithm. Besides, we prove that our TAPRank performs better than other related models.

We only take the evaluation on DBLP dataset in the experiments, which is not sufficient to some degree. Different resources of dataset should be applied to test the effectiveness of the TAPRank model. As well, more factors in the time-aware PageRank algorithm and more types of nodes in the network should also be explored.

Acknowledgment

This work was partially supported by the Natural Science Foundation of China under Grant No. 61572106, No. 61174174 and No. 61203165, the Foundation of Key Laboratory of System Control and Information Processing, Ministry of Education, China, under Grant No. SCIP2012001, and the Fundamental Research Funds for the Central Universities under Grant DUT15YQ112 and No.DUT15QY06.

References

[1] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National academy of Sciences of the United States of America*, vol. 102, no. 46, pp. 16 569–16 572, 2005.

[2] L. Li, X. Wang, Q. Zhang, P. Lei, M. Ma, and X. Chen, "A quick and effective method for ranking authors in academic social network," in *Multimedia and Ubiquitous Engineering*, 2014, pp. 179–185.

[3] J.-P. Qiu, K. Dong, and H.-Q. Yu, "Comparative study on structure and correlation among author co-occurrence networks in bibliometrics," *Scientometrics*, vol. 101, no. 2, pp. 1345–1360, 2014.

[4] Y. Li, C. Wu, X. Wang, and P. Luo, "A network-based and multi-parameter model for finding influential authors," *Journal of Informetrics*, vol. 8, no. 3, pp. 791–799, 2014.

[5] J. D. West, M. C. Jensen, R. J. Dandrea, G. J. Gordon, and C. T. Bergstrom, "Author-level eigenfactor metrics: Evaluating the influence of authors, institutions, and countries within the social science research network community," *Journal of the American Society for Information Science and Technology*, vol. 64, no. 4, pp. 787–801, 2013.

[6] A. Abbasi, "h-type hybrid centrality measures for weighted networks," *Scientometrics*, vol. 96, no. 2, pp. 633–640, 2013.

[7] S. Brin and L. Page, "Reprint of: The anatomy of a large-scale hypertextual web search engine," *Computer networks*, vol. 56, no. 18, pp. 3825–3833, 2012.

[8] Y. Ding, "Applying weighted pagerank to author citation networks," *Journal of the American Society for Information Science and Technology*, vol. 62, no. 2, pp. 236–245, 2011.

[9] D. Fiala, "Time-aware pagerank for bibliographic networks," *Journal of Informetrics*, vol. 6, no. 3, pp. 370–388, 2012.

[10] J. Liu, Y. Li, Z. Ruan, G. Fu, X. Chen, R. Sadiq, and Y. Deng, "A new method to construct co-author networks," *Physica A: Statistical Mechanics and its Applications*, vol. 419, pp. 29–39, 2015.

[11] Q. Meng and P. J. Kennedy, "Discovering influential authors in heterogeneous academic networks by a co-ranking method," in *Proceedings of the 22nd ACM international conference on Conference on information & knowledge management*, 2013, pp. 1029–1036.

[12] T. Amjad, Y. Ding, A. Daud, J. Xu, and V. Malic, "Topic-based heterogeneous rank," *Scientometrics*, vol. 104, no. 1, pp. 313–334, 2015.

[13] A. Pal and S. Ruj, "Citex: A new citation index to measure the relative importance of authors and papers in scientific publications," in *IEEE International Conference on Communications*, 2015. DOI: arXiv:1501.04894.

[14] Y.-p. Du, C.-q. Yao, and N. Li, "Using heterogeneous patent network features to rank and discover influential inventors," *Frontiers of Information Technology & Electronic Engineering*, vol. 16, no. 7, pp. 568–578, 2015.

[15] Z. Liu, H. Huang, X. Wei, and X. Mao, "Tri-rank: An authority ranking framework in heterogeneous academic networks by mutual reinforce," in *IEEE International Conference on Tools with Artificial Intelligence (ICTAI)*, 2014, pp. 493–500.

[16] H. Xu, E. Martin, and A. Mahidadia, "Contents and time sensitive document ranking of scientific literature," *Journal of Informetrics*, vol. 8, no. 3, pp. 546–561, 2014.

[17] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: extraction and mining of academic social networks," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 990–998.

[18] M. Nykl, K. Ježek, D. Fiala, and M. Dostal, "Pagerank variants in the evaluation of citation networks," *Journal of Informetrics*, vol. 8, no. 3, pp. 683–692, 2014.