Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence

Ranking Scientific Articles by Exploiting Citations, Authors, Journals, and Time Information

Yujing Wang, Yunhai Tong *
Key Laboratory of Machine Perception
Peking University, Beijing 100871, China
kellyweiyangwang@gmail.com,yhtong@pku.edu.cn

Ming Zeng
Sun Yat-Sen University
Guangzhou 510620, China
mingtsang.zm@gmail.com

Abstract

Ranking scientific articles is an important but challenging task, partly due to the dynamic nature of the evolving publication network. In this paper, we mainly focus on two problems: (1) how to rank articles in the heterogeneous network; and (2) how to use time information in the dynamic network in order to obtain a better ranking result. To tackle the problems, we propose a graph-based ranking method, which utilizes citations, authors, journals/conferences and the publication time information collaboratively. The experiments were carried out on two public datasets. The result shows that our approach is practical and ranks scientific articles more accurately than the state-of-art methods.

Introduction

Ranking scientific publications is an important task, which helps researchers find related works of high quality. However, the dynamic nature of the evolving publication network makes the task very challenging. Traditional ranking methods view the article collection as a static citation network and leverage the PageRank algorithm (Page et al. 1998) to calculate the prestige of articles. However, these methods do not capture the dynamic nature of the network and are often biased to old publications. For newly published articles, little citations can be found so that they are hard to be recommended by citation-based systems.

To tackle this problem, many efforts have been made to explore additional information for help. For example, Walker et al. (2007) introduced an algorithm called CiteRank, which considers the publication time of scientific articles and utilizes a random walk model to predict the number of future citations for each article. The model reduces the bias of time to some extent, because recent articles will be promoted to higher scores. Nevertheless, since only citations and time information are used, the method is still unable to obtain reasonable comparison between recent publications.

Sayyadi and Getoor (2009) defined a model, FutureRank, which estimates the future PageRank score for each article by using citations, authors, and time information collaboratively. In the model, the usage of authorship provides additional information to rank recent publications. If an author is authority (i.e., publishing many prestigious papers previously), the new publications of him/her can be expected to have good quality. Moreover, P-Rank (Yan, Ding, and Sugimoto 2011) constructs a scholarly network consisting of different entities (publications, authors and journals) and performs a propagation on the network to rank the entities jointly.

In this paper, we mainly focus on two issues. First, we are aiming to propose a PageRank+HITS framework that exploits different kinds of information simultaneously and examines their usability in tasks of ranking scientific articles. Second, we study how to capture time information in the evolving network to obtain a better ranking result.

To address the first issue, we construct a heterogeneous network which contains three sub-networks (citation network, paper-author network, and paper-journal network). Our approach conducts the HITS and PageRank algorithm collaboratively to generate the ranking list of scientific articles. The PageRank algorithm is beneficial to explore the global structure of the citation network, while the HITS algorithm (Kleinberg 1999) is helpful to leverage the local structure by distinguishing the entities as authorities and hubs, and calculating their scores in a mutual reinforcing way. Therefore, the PageRank and HITS algorithm are applied collaboratively to the citation network to calculate the prestige of articles. The HITS-style algorithm can also be applied to the paper-author and paper-journal network by viewing the authors and journals as different types of hub nodes and the papers as authority nodes. As authors and journals provide additional information for articles, we can get a more reasonable ranking result, especially for recent papers whose citations are rare.

For the second issue, we explore two time-relevant strategies. First, in each iteration of the propagation method, recent articles are promoted to higher prestige scores, because they are always underestimated due to the lack of citations. At the same time, instead of using the same weight for all edges between hubs and authorities, we define a time-aware weight to each edge in the network.

The proposed algorithm is evaluated on two public datasets. One is the arXiv dataset, containing articles published on high energy physics. Another is the Cora dataset, which comprises papers in the computer science field. The

*Corresponding author: Yunhai Tong

experiments demonstrate that our approach predicts the future citations of scientific articles more accurately, compared with the baseline approach and other state-of-art methods.

Related Work

Ranking scientific articles is an important task which has been studied for many years. In 1972, Garfield (1972) proposed a measure, namely the *Impact Factor*, to estimate the prestige of journals. In the following years, many efforts have been made to improve the afore-mentioned measure. For instance, Garfield (1983) applied the same idea to ranking scientific authors. Pinski and Narin (1976) noticed that the citations from more important journals should be given higher weights when calculating the influence of other journals. Thus, they introduced a method to calculate the influence of journals based on their cross citing matrix. The motivation of this work is similar to the well-known Page-Rank algorithm which was proposed later.

Page et al. (1998) introduced the famous PageRank algorithm, which provides a way to explore the global information in the ranking procedure. It has been adopted by various applications. For example, Liu et al. (2005) utilized the PageRank algorithm on the co-authorship network to calculate the influence of scientists. (Ma, Guan, and Zhao 2008); (Bollen, Rodriquez, and de Sompel 2006); and (Chen et al. 2007) applied the PageRank approach to the citation network for ranking scientific articles. However, these methods are biased to old articles because recent articles do not have sufficient citations. Moreover, as analyzed in (Maslov and Redner 2008), the PageRank algorithm holds some other pitfalls when extending to the citation network. For instance, the number of citations is ill suited to compare the impact of papers from different scientific fields. In addition, Kleinberg (1999) proposed another famous algorithm, HITS, which makes distinction between hubs and authorities and computes their prestige in a mutually reinforcing way.

In order to capture the dynamic nature of the publication network, Walker et al. (2007) introduced the CiteRank method, which considered the publication time information and leveraged a random walk procedure to predict the article's future citations. In this model, recent articles will be promoted to higher scores so that the bias is reduced to some extent. The CiteRank scores $\vec{S}$ are given by:

$$\begin{aligned} \vec{S} &= I \cdot \vec{p} + (1-a)W \cdot \vec{p} + (1-a)^2 W^2 \cdot \vec{p} + ... \\ p_i &= e^{-(T_{current} - T_i)/\tau} \end{aligned} \quad (1)$$

where p_i represents the probability of initially selecting the ith paper, and $T_{current} - T_i$ is the number of years since the ith paper was published. W is the citation matrix, τ and a are constant parameters.

Some researchers also utilized the authorship and conference information together with the citations to improve the article's ranking result. Zhou and Orshanskiy (2007) presented a method to co-rank authors and their publications using a reinforcement algorithm. This approach benefits from three networks simultaneously: the social network connecting the authors, the citation network connecting the publications, as well as the authorship network that ties the authors and publications together. (Zhou, Lu, and Li 2012) and (Das, Mitra, and Giles 2011) also accommodated the source of authorship as well as an iterative algorithm to obtain better ranking results for scientists and their publications.

Sayyadi and Getoor (2009) proposed a method, FutureRank, which estimated the expected future prestige scores of articles by the following information: the citation network, the authorship information, and the publication date for each article. Yan, Ding, and Sugimoto (2011) constructed a heterogeneous scholarly network and used a random walk method called P-Rank to rank authors, journals, and publications simultaneously. The method performs the following two steps alternately until convergence is encountered: (1) performing random walk on two bipartite graphs, namely the paper-author graph and the paper-journal graph; (2) conducting PageRank propagation on the citation network. In addition, Hwang, chae, and Kim (2010) also proposed an algorithm that ranks publications by considering the importance of corresponding authors and journals.

Article Ranking Model

In this section, we introduce our proposed article ranking model. Each article can be represented by four attributes, namely the citation list, the author list, the journal/conference name and the publication date. We build a network structure based on these attributes and utilize a graph-based propagation algorithm to calculate the prestige scores of articles. Our method makes use of four kinds of information (the citations, authorship, journal/conference information and publication date) collaboratively. The motivation can be described as follows:

- Important articles are often cited by many other important publications.
- A good hub paper is one that points to many good authorities; a good authority paper is one that is pointed by many good hubs.
- Authors with high-reputation are more likely to publish papers of high-quality.
- Good papers are more likely to appear on top journals and conferences.
- The publication time is useful for ranking scientific articles. For example, recent papers have little citations so that they should be given some promotion in the propagation procedure.

Network Structure

There are three types of nodes in the network, i.e., papers, authors, and journals/conferences. In addition, we have three types of edges. The *citation edge* is a directed edge, which links from the original paper to one of its citing papers. The *authorship edge* is an undirected edge between a paper and an author, which indicates that the specific author is in the paper's author list. The *journal/conference edge* is an undirected edge between a paper and a journal/conference, which denotes that the paper is published on the specific journal or conference.

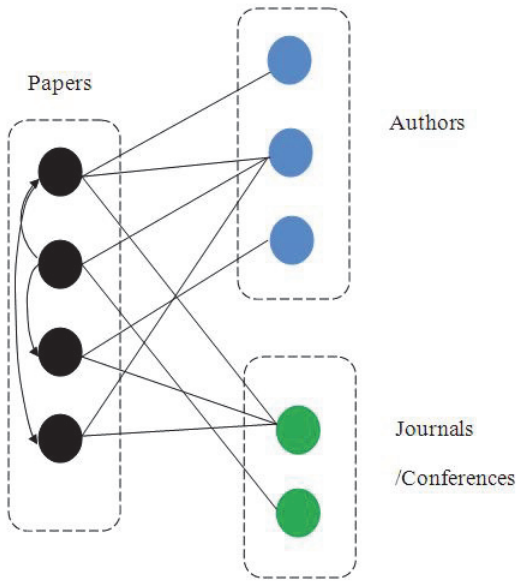

Figure 1: A demonstration of the network structure

Figure 1 is a demonstration of the network structure. To get a clear understanding of the network structure, we can view the network as a composition of three sub-networks, namely the *citation network*, the *paper-author network*, and the *paper-journal network*. The *citation network* contains all the paper nodes and citation edges between them, constructing a graph structure where the traditional PageRank and HITS algorithm can be used jointly to estimate the ranking result. The *paper-author network* is a bipartite graph, containing both the paper nodes and author nodes. It contains only one type of edges, i.e., the *authorship edges* between papers and authors. According to (Sayyadi and Getoor 2009), the article nodes can be taken as authorities and the author nodes as hubs; thus, the HITS-style reinforcement algorithm can be applied to this kind of network. The *paper-journal network* has a similar structure, consisting of two types of nodes (the article nodes and the journal/conference nodes) and one type of edges (the edges between journals/conferences and papers), which form a bipartite graph.

Ranking Algorithm

We adopt a reinforcement procedure to calculate the prestige scores of articles, as well as the scores of authors and journals/conferences. The algorithm is conducted by the following steps:

1. Initially, all the authority scores of papers are assigned to be $\frac{1}{N_p}$, where N_p is the number of all papers in the collection.
2. Calculate the hub scores of authors by the *paper-author network*.
3. Calculate the hub scores of journals/conferences by the *paper-journal network*.
4. Calculate the hub scores of papers by the *citation network*.
5. Update the authority scores of papers, using five types of information, i.e., the PageRank score propagated from citations, the score transferred from authors, the score transferred from journals/conferences, the score transferred from hub papers, and the time-aware score calculated by publication date.
6. Perform step 2-5 iteratively until convergence is encountered.

Calculating hub scores The hub scores can be calculated by collecting the authority scores from corresponding papers. Different from the traditional HITS method, we normalize the score of a hub by its number of links. It helps to obtain a more appropriate estimation for the quality of hubs (e.g., conference that publishes many papers may have low average quality). Thus, the hub score of an author is calculated by

$$H(A_i) = \frac{\sum_{P_j \epsilon Neighbor(A_i)} S(P_j)}{|Neighbor(A_i)|} \tag{2}$$

where $H(A_i)$ is the hub score of author A_i, $S(P_j)$ is the authority score of paper P_j, $Neighbor(A_i)$ is the collection of papers which correspond to Author A_i, and $|Neighbor(A_i)|$ is the number of papers in the collection. After calculation, the sum of hub scores for all the authors is normalized to 1. That is, $\sum_i H(A_i) = 1$.

Similarly, the hub score of a journal or conference can be calculated by

$$H(J_i) = \frac{\sum_{P_j \epsilon Neighbor(J_i)} S(P_j)}{|Neighbor(J_i)|} \tag{3}$$

where $H(J_i)$ is the hub score of the journal/conference J_i, $Neighbor(J_i)$ is the collection of papers published on J_i, and $|Neighbor(J_i)|$ is the number of papers in $Neighbor(J_i)$. The scores are also normalized such that $\sum_i H(J_i) = 1$.

The hub score of a paper is computed by

$$H(P_i) = \frac{\sum_{P_j \epsilon Neighbor(P_i)} S(P_j)}{|Neighbor(P_i)|} \tag{4}$$

where $H(P_i)$ is the hub score of paper P_i, $Neighbor(P_i)$ is the collection of papers which P_i links to, and $|Neighbor(P_i)|$ is the number of papers in $Neighbor(P_i)$. Similarly, the hub scores of papers are normalized such that $\sum_i H(P_i) = 1$.

Calculating authority scores Given the authority and hub scores of all the papers, as well as the hub scores of authors and journals/conferences, the authority score of each paper can be updated by:

$$\begin{aligned} S(P_i) = {} & \alpha \cdot PageRank(P_i) \\ & + \beta \cdot Author(P_i) \\ & + \gamma \cdot Journal(P_i) \\ & + \delta \cdot Citation(P_i) \\ & + \theta \cdot P_i^{Time} \\ & + (1 - \alpha - \beta - \gamma - \delta - \theta) \cdot 1/N_p \end{aligned} \tag{5}$$

where α, β, γ, δ and θ are constant parameters which range in (0, 1). As shown in the equation, the authority score $S(P_i)$ is calculated by the linear combination of the following five scores:

- $PageRank(P_i)$ is the traditional PageRank score of paper P_i calculated by the *citation network*:

$$PageRank(P_i) = \sum_{P_j \epsilon In(P_i)} \frac{1}{|Out(P_j)|} \cdot S(P_j) \quad (6)$$

  where $In(P_i)$ contains the papers which link to the paper P_i and $|Out(P_j)|$ is the number of papers which link out from paper P_j. $S(P_j)$ is the original authority score of paper P_j before updating.

- $Author(P_i)$ is the authority score of paper P_i propagated from the corresponding authors in the *paper-author network*. It can be computed by

$$Author(P_i) = \frac{1}{Z(A)} \cdot \sum_{A_j \epsilon Neighbor(P_i)} H(A_j) \quad (7)$$

  where $Neighbor(P_i)$ is the author list corresponding to paper P_i, and $H(A_j)$ is the hub score of author A_j. $Z(A)$ is a normalized value, which equals to the sum of scores transferred from all the authors to papers.

- $Journal(P_i)$ is the authority score of paper P_i propagated from the corresponding journal/conference in the *paper-journal network*. The formula for calculation is similar to equation 7:

$$Journal(P_i) = \frac{1}{Z(J)} \cdot \sum_{J_j \epsilon Neighbor(P_i)} H(J_j) \quad (8)$$

  where $Neighbor(P_i)$ contains the corresponding journal/conference (only one for each paper) of paper P_i, and $H(J_j)$ is the hub score of journal/conference J_j. $Z(J)$ is the sum of scores transferred from all the journals/conferences to all the papers.

- $Citation(P_i)$ is the authority score of paper P_i collected from hub papers, which can be obtained similarly:

$$Citation(P_i) = \frac{1}{Z(P)} \cdot \sum_{P_j \epsilon Neighbor(P_i)} H(P_j) \quad (9)$$

  where $Neighbor(P_i)$ contains the hub papers which links to paper P_i, and $H(P_j)$ is the hub score of paper P_j. $Z(P)$ is the sum of scores transferred from all the hub papers.

- P_i^{Time} is a time-aware value for paper P_i. As we have mentioned, recent publications are of great importance but they are always underestimated by citation-based algorithms due to the lack of citations. Therefore, we use this personalized score to promote the prestige of new articles. According to (Sayyadi and Getoor 2009), we define the function as follows:

$$P_i^{Time} = e^{-p*(T_{current} - T_i)} \quad (10)$$

  where T_i is the publication time of paper P_i, $T_{current} - T_i$ is the number of years since the paper P_i was published. p is a constant value, which is set to be 0.62 in our experiments (Sayyadi and Getoor 2009). The sum of P_i^{Time} scores for all the papers is normalized to 1.

- $(1 - \alpha - \beta - \gamma - \delta - \theta) \cdot \frac{1}{N_p}$ denotes the probability of random jump, where N_p is the number of papers in the network.

In the propagation algorithm, the initial authority scores of all the papers are set to be $\frac{1}{N_p}$. For papers which do not cite any other articles, we assume that they have links to all the other papers. Thus, the sum of authority scores for all the papers will keep to be 1 in each iteration. The convergence is judged by the following rule: if the difference between current and previous scores of each article is less than a threshold, then the convergence is encountered. In practice, the threshold is set to be 0.0001 experimentally.

Time-Aware Weights of Edges

In the previous section, the propagation algorithm uses an assumption that all the weights of edges are equal to 1. However, the weights should be time-relevant to capture the dynamic nature of the evolving network. Therefore, we define a time-aware strategy for weight calculation. In practice, we find that older articles often get more accurate predictions than recent ones. This observation is easy to be explained as for older articles, the citation information is more sufficient and reliable. Therefore, when computing the scores of hubs, we give the edges associated with older authority papers higher weights, because their scores are more reliable than those of new articles. Thus, the hub score of an author can be calculated as follows (see function 2 for comparison):

$$H(A_i) = \frac{\sum_{P_j \epsilon Neighbor(A_i)} w_{ap}(i,j) \cdot S(P_j)}{\sum_{P_j \epsilon Neighbor(A_i)} w_{ap}(i,j)} \quad (11)$$

where $w_{ap}(i,j)$ is the weight of edge from author A_i to paper P_j. The scores are normalized such that $\sum_i H(A_i) = 1$. The value of $w_{ap}(i,j)$ is defined as

$$w_{ap}(i,j) = a^{(T_{current} - T_i)} \quad (12)$$

where $T_{current} - T_i$ is the number of years between the publication date of article P_i and the current time; a is a constant parameter greater than 1. In our experiments, we set $a = 2$.

The hub scores of journals/conferences and papers can be calculated similarly. On the other hand, the authority papers published recently are more important; so they should receive higher scores from the hubs. Thus, the authority score of paper P_i received from the corresponding authors can be calculated by (see function 7 for comparison)

$$Author(P_i) = \frac{1}{Z(A)} \cdot \sum_{A_j \epsilon Neighbor(P_i)} w_{pa}(i,j) \cdot H(A_j) \quad (13)$$

where $w_{pa}(i,j)$ is the weight of edge from paper P_i to author A_j. The weight $w_{pa}(i,j)$ is defined as

$$w_{pa}(i,j) = \frac{1}{1 + b \cdot (T_{current} - T_i)} \quad (14)$$

where b is set to be 1 experimentally. The authority scores received from journals/conferences and hub papers can be computed in the same way.

Table 1: Distribution of papers over years

arXiv dataset													
	before 1992	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003
# of papers	41	1273	2055	2397	2344	2617	2663	2767	2779	3135	3153	3296	1035
Cora dataset													
	before 1988	1988	1989	1990	1991	1992	1993	1994	1995	1996	1997	1998	1999
# of papers	342	140	233	414	728	1115	1777	2596	2731	2851	2402	1088	39

Experiments

Datasets

We use two public datasets in our experiments. One is the arXiv (hep-th) dataset[1], which is provided for the 2003 KDD Cup competition. The collection contains articles published on high energy physics from 1992 to 2003. The dataset consists of approximately 30,000 articles, with 350,000 citations. In the dataset, each article is associated with a publication date. For each article, we extract the authors and journal information from the description. Two authors or journals will be considered the same if they are exact match. Although the simple rule sometimes leads to mistakes, we find that it works well in most of the cases. By this rule, we find about 16,000 authors and 400 journals in total.

Another dataset is Cora[2] provided by McCallum. It contains 19,396 research papers in the computer science field with 46,714 citations. Because the evaluation procedure depends on the time information, we remove the papers which do not include publication date. Thus, there remain 16,456 papers and 35,463 citations. Moreover, we remove the time information in the conference name. For example, "*The Twenty-Seventh AAAI Conference on Artificial Intelligence (AAAI-13)*" will be renamed as "*The AAAI Conference on Artificial Intelligence (AAAI)*". In all, we collect 13,232 authors and 8,324 journals/conferences in the dataset. The distribution of papers over years is demonstrated in Table 1.

Evaluation Metrics

The evaluation procedure is challenging because we do not have the ground truth of the article's real rank. Sayyadi and Getoor (2009) used the future PageRank score as ground truth. However, it is not very appropriate because the PageRank algorithm itself is biased to old articles. Some old papers become unfashionable, but they still get high PageRank scores because they received many citations in the past. Therefore, we adopt the future citation number as ground truth instead. We view the system from a historical time point and calculate the number of future citations (cited after the historical time point) for each paper. The ground truth list is then obtained by ranking the papers according to the future citation number. The estimated rank list can be generated by applying the algorithm on the historical time point. Finally, the result can be reported by the similarity of two rank lists. We assess the similarity by the Spearman's rank correlation coefficient Myers and Well (2003):

$$\rho = \frac{\sum_i (R_1(P_i) - \bar{R}_1)(R_2(P_i) - \bar{R}_2)}{\sqrt{\sum_i (R_1(P_i) - \bar{R}_1)^2 \sum_i (R_2(P_i) - \bar{R}_2)^2}} \quad (15)$$

where $R_1(P_i)$ is the position of paper P_i in the first rank list; $R_2(P_i)$ is the position of the specific paper in the second rank list; $\bar{R}_1$ and $\bar{R}_2$ are the average rank positions of all papers in the first and second rank list respectively. In the case of ties, the rank position is set to be the average rank of all the ties.

Experimental Setup

We evaluate various configurations of the proposed algorithm. Depending on whether and how to use different types of sub-networks, there are 8 kinds of settings:

- **PageRank**: uses the traditional PageRank algorithm on the *citation network* for rank calculation.
- **Author-relevant**: uses PageRank on the *citation network* and HITS on the *paper-author network* to estimate the future prestige of articles.
- **Journal-relevant**: uses PageRank on the *citation network* and HITS on the *paper-journal network*.
- **Hub-relevant**: uses PageRank and HITS on the *citation network* collaboratively.
- **AJ-relevant**: utilizes PageRank on the *citation network*, together with HITS on two networks: the *paper-author network* and the *paper-journal network*.
- **AH-relevant**: utilizes PageRank on the *citation network*, together with HITS on two networks: the *paper-author network* and the *citation network*.
- **JH-relevant**: utilizes PageRank on the *citation network*, together with HITS on two networks: the *paper-journal network* and the *citation network*.
- **All-relevant**: performs PageRank on the *citation network*, together with HITS on all the three networks: the *paper-author*, *paper-journal*, and *citation network*.

Depending on how to use time information, there are 3 kinds of settings:

- **No-time**: does not use time information in the graph propagation procedure.
- **Time**: uses time information to promote recent articles in the graph propagation procedure.
- **Time-weighted**: uses time information to promote recent articles in the graph propagation procedure; at the same time, uses time-aware weighting strategy to calculate the weights of edges in the network (see function 12 and 14).

[1] http://www.cs.cornell.edu/projects/kddcup/datasets.html

[2] http://people.cs.umass.edu/ mccallum/data.html

Table 2: Results on two datasets

arXiv dataset (2000-01-01)											
	Page-Rank	Author-relevant	Journal-relevant	Hub-relevant	AJ-relevant	AH-relevant	JH-relevant	All-relevant	Future-Rank	Cite-Rank	P-Rank
No-time	0.4115	0.4472	0.4773	0.5593	0.5010	0.5653	0.5937	0.5966	0.6445	0.6451	0.4635
Time	0.6358	0.6445	0.6591	0.6687	0.6611	0.6736	0.6896	0.6917			
Time-weighted	—	0.6516	0.6774	0.6778	0.6839	0.6860	0.7052	**0.7093**			
Cora dataset (1996-01-01)											
	Page-Rank	Author-relevant	Journal-relevant	Hub-relevant	AJ-relevant	AH-relevant	JH-relevant	All-relevant	Future-Rank	Cite-Rank	P-Rank
No-time	0.2815	0.2812	0.2890	0.2956	0.2980	0.2993	0.3067	0.3154	0.3649	0.3682	0.2952
Time	0.3593	0.3649	0.3687	0.3796	0.3730	0.3825	0.3900	0.3916			
Time-weighted	—	0.3676	0.3680	0.3917	0.3748	0.3954	0.3967	**0.3994**			

Therefore, we have 24 (8×3) kinds of settings for evaluation. The probability of random jump is set to be 0.15 experimentally. Thus, we have $\alpha + \beta + \gamma + \delta + \theta = 0.85$. To further reduce the degrees of freedom, the parameters are restricted to $\beta = \gamma$ experimentally when we use the author and journal information simultaneously. Therefore, there are at most 3 variations for all the settings. In each setting, the parameters are set to be optimal.

We also evaluate three state-of-art methods for comparison: (1) FutureRank (consistent with our approach of "Author + Time" setting); (2) CiteRank; and (3) P-Rank (refer to the related work section for their introductions). We use the optimal parameters for all the methods above. The algorithms are evaluated on two datasets. For each dataset, we set an appropriate historical time point for evaluation according to the distribution of papers over time (see Table 1). For the arXiv dataset, the historical time point is set to be 2000-01-01; and for the Cora dataset, it is 1996-01-01.

Results

The results of various settings (reported by the Spearman's rank correlation coefficient), as well as those of three state-of-art methods are shown in Table 2. Firstly, we can see that the approach can benefit from PageRank and HITS algorithm simultaneously. As shown in the table, performing HITS on each of the three networks has a mutual reinforcing effect with PageRank, which helps to generate a more reasonable ranking list. Secondly, capturing the dynamic nature of the evolving network is useful for rank calculation. Comparing the results of "No-time" configuration with the corresponding "Time" configuration, it verifies that it is beneficial to promote recent articles to higher scores. Moreover, using time-aware weights can give a further improvement to the result (see the comparison between "Time" and "Time-Weighted" settings).

The best result can be achieved by using all kinds of information jointly. As shown in Table 2, our approach outperforms other state-of-art methods on both datasets. We can also notice that the ranking result of arXiv dataset is much more accurate than that of Cora dataset. The reasons are: (1) The citation network in arXiv dataset is more compact. It has about 30,000 articles with 350,000 citations; while Cora dataset contains 16,456 papers with only 35,463 citations. (2) The paper-author and paper-journal network in the arXiv dataset are also more compact. This indicates that we should use compact graphs in order to get more accurate results. In the web environment, we can first collect a compact graph for each topic before applying the ranking algorithm.

Table 3: Results on two datasets

Dataset	Fixed Setting	Best Setting	Future-Rank	CiteRank	P-Rank
Arxiv	0.7065	0.7093	0.6445	0.6451	0.4635
Cora	0.3931	0.3994	0.3649	0.3682	0.2952

Sensitivity of Parameters

In this section, we carry out another experiments to test the sensitivity of parameters. First, we use the historical data in the Arxiv dataset to learn the parameters, that is, finding the best setting for predicting article citations at 1997-01-01. Then, the setting is fixed and applied to the future time point on both Arxiv and Cora dataset. For Arxiv dataset, we evaluate the ranking at 2000-01-01; for Cora dataset, we evaluate at 1996-01-01. In Table 3, we report the results of both fixed setting and best setting for our approach. The results of other state-of-art methods are reported on the best settings. As shown in the table, the usage of fixed setting does not loss much of the accuracy achieved by the best setting. The setting is fairly stable even across different datasets (learned from Arxiv and applied to Cora).

Conclusion & Future Work

In this paper, we introduce a new approach for ranking scientific articles. We provide a framework of exploiting different kinds of information (including but not restricted to citations, authors, and journals/conferences) in a heterogeneous network, which benefits from the PageRank and HITS algorithm collaboratively to estimate the article's future prestige. Moreover, we study how to capture the dynamic nature of the evolving publication network, and propose two time-relevant strategies: (1) promoting recent articles to higher scores; (2) using time-aware weights of edges between hubs and authorities. The experiments were carried out on two public datasets of scientific articles and promising results were achieved. In the next step, we will test the proposed method on more datasets and examine its usability in ranking articles from different topics.

References

Bollen, J.; Rodriquez, M.; and de Sompel, H. V. 2006. Journal status. *Scientometrics* 69(3):669–687.

Chen, P.; Xie, H.; Maslov, S.; and Redner, S. 2007. Finding scientific gems with google. *Journal of Informetrics 1* 8–15.

Das, S.; Mitra, P.; and Giles, C. L. 2011. Ranking authors in digital libraries. In *Proceedings of the 11th annual international ACM/IEEE Joint Conference on Digital Libraries (JCDL'11)*, 251–254.

Garfield, E. 1972. Citation analysis as a tool in journal evaluation. *Science* 178:471–479.

Garfield, E. 1983. How to use citatiuon analysis for faculty evaluations and when is it relevant? (part 1). *Current Contents* 5–13.

Hwang, W.-S.; chae, S.-M.; and Kim, S.-W. 2010. Yet another paper ranking algorithm advocating recent publications. In *Proceedings of the 19th international conference on World Wide Web (WWW'10)*, 1117–1118.

Kleinberg, J. M. 1999. Authoritative sources in a hyperlinked environment. *Journal of the ACM* 46(5):604–632.

Liu, X.; Bollen, J.; Nelson, M. L.; and de Sompel, H. V. 2005. Co-authorship networks in the digital library research community. *Information Processing and Management* 41(6):1462–1480.

Ma, N.; Guan, J.; and Zhao, Y. 2008. Bringing pagerank to the citation analysis. *Information Processing and Management* 44(2):800–810.

Maslov, S., and Redner, S. 2008. Promise and pitfalls of extending googles pagerank algorithm to citation networks. *The Journal of Neuroscience* 28(44):11103–11105.

Myers, J. L., and Well, A. D. 2003. *Research Design and Statistical Analysis*. Lawrence Erlbaum. pp.508.

Page, L.; Brin, S.; Motwani, R.; and Winograd, T. 1998. The pagerank citation ranking: Bringing order to the web. *Technical Report, Stanford University Database Group*.

Pinski, G., and Narin, F. 1976. Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics. *Information Processing and Management* 297–312.

Sayyadi, H., and Getoor, L. 2009. Futurerank: Ranking scientific articles by predicting their future pagerank. In *Proceedings of the Ninth SIAM International Conference on Data Mining (SDM'09)*, 533–544.

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

Yan, E.; Ding, Y.; and Sugimoto, C. R. 2011. An indicator measuring prestige in heterogeneous scholarly networks. *Journal of the American Society for Information Science and Technology* 62(3):467–477.

Zhou, D., and Orshanskiy, S. A. 2007. Co-ranking authors and documents in a heterogeneous network. In *Proceedings of the Seventh IEEE International Conference on Data Mining (ICDM'07)*.

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