Weighted P-Rank Algorithm Based on a Heterogeneous Scholarly Network

Jian Zhou (✉ zhoujian@mail.dlut.edu.cn)

Dalian University of Technology

Lin Feng

Dalian University of Technology

Shenglan Liu

Dalian University of Technology

Jie Yang

Tsinghua University

Ning Cai

Beijing University of Posts and Telecommunications

Research Article

Keywords: Article ranking, Link weighting, Heterogeneous scholarly network, Weighted P-Rank algorithm

Posted Date: November 29th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1080239/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Weighted P-Rank Algorithm Based on a Heterogeneous Scholarly Network

Jian Zhou^{1*}, Lin Feng¹, Sheng-Lan Liu¹
¹School of Computer Science and Technology
Dalian University of Technology
Dalian 116024, China

Jie Yang²
²Research Institute of Information Technology
Tsinghua University
Beijing 100084, China

Ning Cai^{3 †}
³School of Artificial Intelligence
Beijing University of Posts and Telecommunications
Beijing 100876, China

November 24, 2021

Abstract

The evaluation of scientific article has always been a very challenging task because of the dynamic change of citation networks. Over the past decades, plenty of studies have been conducted on this topic. However, most of the current methods do not consider the link weightings between different networks, which might lead to biased article ranking results. To tackle this issue, we develop a weighted P-Rank algorithm based on a heterogeneous scholarly network for article ranking evaluation. In this study, the corresponding link weightings in heterogeneous scholarly network can be updated by calculating citation relevance, authors' contribution, and journals' impact. To further boost the performance, we also employ the time information of each article as a personalized PageRank vector to balance the bias to earlier publications in the dynamic citation network. The experiments are conducted on three public datasets (arXiv, Cora, and MAG). The experimental results demonstrated that weighted P-Rank algorithm significantly outperforms other ranking algorithms on arXiv and MAG datasets, while it achieves competitive performance on Cora dataset. Under different network configuration conditions, it can be found that the best ranking result can be obtained by jointly utilizing all kinds of weighted information.

Keywords: Article ranking · Link weighting · Heterogeneous scholarly network · Weighted P-Rank algorithm

1 Introduction

Scholarly impact assessment and ranking have always been a hot issue, which plays an important role in the process of the dissemination and development of academic research [1] [2] [3]. However, it is difficult to assess the real quality of academic articles due to the dynamic change of citation networks [4]. Furthermore, the evaluation result will be heavily influenced by utilizing different bibliometrics indicators or ranking methods [5]. As early as in 1972, Garfield proposed journal impact factor (JIF) to rank various academic journals [6]. In 1983, Garfield applied the JIF mechanism to rank individual authors [7]. According to the calculation of the H-index [8], Braun *et al.* [9] introduced journal H-index to indicate the quality of a journal and the number of published articles comprehensively.

*Corresponding author: zhoujian@mail.dlut.edu.cn

†Corresponding author: caining91@tsinghua.org.cn

As a traditional ranking method, PageRank [10] algorithm has already been widely and effectively used in various ranking tasks. Liu *et al.* [11], for instance, employed the PageRank algorithm to evaluate the academic influence of scientists in the co-authorship network. In [12], Bollen *et al.* utilized a weighted version of the PageRank to improve the calculation methodology of JIF. It is worth remarking that the vast majority of ranking algorithms such as PageRank and its variants deem the article (node) creation as a static citation network. In the real citation network, however, articles are published and cited in time sequence. Such approaches do not consider the dynamic nature of the network and are always biased to old publications. Therefore, the recent articles tend to be underestimated due to the lack of enough citations. To address this issue, Sayyadi and Getoor proposed a timeaware method, FutureRank [4], which calculates the future PageRank score of each article by jointly employing citation network, authorship network, and time information. In comparison to the other methods without time weight, FutureRank is practical and ranks academic articles more accurately. Furthermore, Walker *et al.* proposed a ranking model called CiteRank [13], which utilizes a simple network traffic model and calculates the future citations of each article by considering the publication time of articles. However, a main problem of the network traffic model is that it does not reveal the mechanism of how the article scores change. Moreover, although PageRank algorithm is advanced at exploring the global structure of the citation network, it neglects certain local factors that may influence the ranking results. To tackle this problem, Yan *et al.* presented an improved ranking method P-Rank [14], which develops a heterogeneous scholarly network containing different entities (publications, authors and journals) and performs a propagation between subnetworks to calculate the prestige of each entity. Also in [15], Kleinberg developed and verified an algorithm of the authority notion, HITS, which exploits the local structure by distinguishing the entities as hubs and authorities, and computing their scores in a mutual effect manner. To further boost the ranking performance of HITS method, Wang *et al.* developed an article ranking framework called PageRank+HITS [16], which exploits graph-based propagation algorithm to compute the prestige scores of articles in the heterogeneous scholarly network. However, although the ranking methods above achieved better performance, they do not consider the link weightings between different networks, which might lead to biased article ranking results. Inspired by PageRank+HITS framework, Zhang *et al.* introduced a link weighting method W-Rank [17] to assign link weighting to the citation network and authorship network by calculating citation relevance and author contribution. From an analysis of their experimental results, it can be found that link weighting scheme is beneficial to improve the performance of the article ranking algorithm.

This paper aims to develop a weighted P-Rank algorithm based on a heterogeneous scholarly network and explore how the changes of the link weightings between different subnetworks influence the ranking result. To further boost the performance of weighted P-Rank algorithm, we utilize the time information of each article as a personalized PageRank vector to balance the bias to earlier publications in the dynamic citation network. The experimental results demonstrated that the proposed algorithm achieved superior performance in comparison to other methods. The key contributions of this work can be summarized as follows:

- A weighted article ranking method based on P-Rank algorithm and heterogeneous graph is developed.
- The weighted P-Rank algorithm considers the influence of citation relevance, authors' contribution, journals' impact, and time information to the article ranking method comprehensively.
- We evaluate the performance of weighted P-Rank method under different conditions by manipulating the corresponding parameters that can be used to structure graph configurations and time settings.
- By introducing the corresponding link weightings in each heterogeneous graph, the performance of the weighted P-Rank algorithm significantly outperforms the original P-Rank algorithm on three public datasets.

The rest of this paper is organized as follows. In Section 2, the details of the heterogeneous scholarly network and the weighted P-Rank algorithm are introduced, respectively. In Section 3, experiments are conducted to analyze how the configurations of the graphs and time settings influence the performance of the proposed algorithm and to further validate link weighting in heterogeneous scholarly network is beneficial to the article ranking. Finally, Section 4 presents the concluding remarks.

2 Article Ranking Model

In this section, we introduce the proposed article ranking algorithm in detail. Specifically, we first define and describe a heterogeneous scholarly network that is composed of author layer, paper layer and journal layer, and how the different

elements in the three layers are linked and interacted. Furthermore, a link weighting method based on P-Rank algorithm is developed to compute the article score in the heterogeneous scholarly network.

2.1 Heterogeneous Scholarly Network

A complete heterogeneous scholarly network consists of three subnetworks (i.e., author network, paper citation network, and journal network). There exist three types of edges in the network i.e., undirected edge between the authors and the papers, directed citation edge between the original paper and its citing papers, and undirected edge between the papers and the published journals. As stated in [14], the heterogeneous scholarly graph of papers, authors, and journals can be expressed as the following form:

$$G(V, E) = (V_P \cup V_A \cup V_J, E_P \cup E_{PA} \cup E_{PJ}) \quad (1)$$

where V_P , V_A , and V_J are the paper nodes, author nodes, and journal nodes in the three layers respectively. E_P denotes the citation link in the paper layer, E_{PA} denotes the link between paper and author, and E_{PJ} denotes the link between paper and journal.

Fig. 1. Visualization of a heterogeneous scholarly network

As shown in Fig. 1a, the paper-author network and paper-journal network are two undirected graphs which can be represented as $G_{PA} = (V_P \cup V_A, E_{PA})$ and $G_{PJ} = (V_P \cup V_J, E_{PJ})$, respectively. In Fig. 1b, by contrast, the paper citation network is a directed graph $G_P = (V_P, E_P)$, the arrows point in the direction of paper citation: $P_4 \rightarrow P_5$ means P_4 cites P_5 . In this work, we assign link weights to the corresponding subnetworks such that the three unweighted graphs can be updated as $G_P = (V_P, E_P, W_P)$, $G_{PA} = (V_P \cup V_A, E_{PA}, W_{PA})$, and $G_{PJ} = (V_P \cup V_J, E_{PJ}, W_{PJ})$, in which W_P , W_{PA} , and W_{PJ} refer to the link weight in the three graphs, respectively. With link weightings (W_P , W_{PA} , and W_{PJ}) defined in the corresponding G_P , G_{PA} , and G_{PJ} , the unweighted heterogeneous scholarly graph $G(V, E)$ becomes

$$G(V, E, W) = (V_P \cup V_A \cup V_J, E_P \cup E_{PA} \cup E_{PJ}, W_P \cup W_{PA} \cup W_{PJ}) \quad (2)$$

which can be verified by substituting W_P , W_{PA} , and W_{PJ} into Eq. (1) to get back Eq. (2).

To evaluate articles, authors, and journals more objectively, the proposed algorithm should be predicated on the following assumptions:

- Articles are more important if they are cited by many other important publications, and therefore they are more likely to be published on certain top journals [14] [18] [19] [20].
- Articles with high-quality are more likely to be written by the authors with higher-reputation [4] [21].

- Journals would hold a higher impact if the published articles are cited by other important articles (or cited by some scholars with higher reputation) [2] [22].
- The citation relevance between two articles is mainly influenced by text similarity and citation network structure (see Sect. 2.2 for more details) [17].
- Author’s contribution to each article is different [17] [23].
- Journal impact factor and the publication time of articles are helpful to rank scientific articles [3] [5] [16] [17] [24].

In the following, we introduce the link weightings in paper citation graph G_P (Sect. 2.2), paper-author graph G_{PA} (Sect. 2.3), and paper-journal graph G_{PJ} (Sect. 2.4), respectively.

2.2 Link Weighting in Paper Citation Graph (G_P)

As discussed in Sect. 1, almost all the ranking algorithms such as PageRank [10], FutureRank [4], and P-Rank [14] only handle article citation in a binary way, rather than paying attention to the combination of citation relevance and link weighting systems. In reality, however, the citation relevance plays a very important role in the process of assessing article quality, and also should be considered in citation network. In this study, we develop a link weighting to assign weight in the paper citation graph (G_P) based on the citation relevance between two papers, which can be utilized to improve the reasonability of the article ranking. In comparison with initial citations, the weighted citations are more advanced at evaluating the real impact of scientific article because it takes full account of certain latent but significant factors in the process of citations. To be specific, the citation relevance (link weighting) between two different papers is mainly influenced by two factors, namely, text similarity (semantic-based) and citation network structure (structure-based). Supposing that the citation relevance between two papers is higher if the two papers are more likely to be similar in semantic and share mutual links and common nodes in the citation network. The comparison between unweighted citation network and weighted citation network is shown in Fig. 2.

Fig. 2. Comparison between unweighted citation network and weighted citation network.

Over the past decade, Natural Language Processing (NLP) has achieved extremely promising performance for various classification tasks such as text categorization, speech recognition, and automatic summarization [25] [26] [27] [28]. Compared with traditional classification methods, the ability and advantage of NLP to learn examples is better at addressing certain real-world applications. In [17], for example, Zhang *et al.* employed a sense-based semantic similarity measure named ADW (Align, Disambiguate and Walk) [29] to measure the semantic similarity between the titles and abstracts of two papers. Technically, ADW calculates the weighted ranking of the importance of senses in WordNet 3.0 [30] for each lexical item. The relevant experimental results in [17] demonstrated that ADW is not only able to handle the texts with different sizes, but also achieves superior performance on different datasets. In our work,

the “slide” weighted overlap approach improved by ADW is employed, which can be used to compute the semantic similarity between the abstracts T_i and T_j from papers i and j . Let S be the intersection of overlapping senses with non-zero probability in both signatures and r_i^j be the rank of sense $s_i \in S$ in signature j , where rank 1 represents the highest rank. The slide overlap $\text{Similarity}_1(P_i, P_j)$ can be computed using:

$$\text{Similarity}_1(P_i, P_j) = \tanh \left(\frac{\alpha \cdot \sum_{i=1}^{|S|} e^{(r_i^1 + r_i^2)^{-1}}}{\beta \cdot \sum_{i=1}^{|S|} e^{(2i)^{-1}}} \right) \quad (3)$$

where $\tanh(\cdot)$ is hyperbolic tangent function, and $\sum_{i=1}^{|S|} (2i)^{-1}$ is the maximum value to bound the similarity distributed over the interval $[0,1]$. Note that the maxium value would occur when each sense has the same rank in both signatures. Moreover, we normalize parameters α and β such that $\alpha + \beta = 1$. It should be explained that the slide ADW will be very helpful to reduce the interdependence of parameters. This is nicely verified by the the performance on the training speed and precision in comparison to the original ADW.

Cosine similarity and Euclidean disatance are widely used to measure the neighbourhoods nodes in citation network [31] [32] [33]. In this work, we employ cosine similarity to measure the citation relevance of two papers in terms of network structure. Assume that if one paper cites another paper, there exist three types of connections between these two papers, as illustrated in Fig. 3. The cosine similarity between two paper nodes in the citation network can be calculated by:

$$\text{Similarity}_2(P_i, P_j) = \cos(P_i, P_j) = \frac{|N_{P_i} \cap N_{P_j}|}{\sqrt{|N_{P_i}| \times |N_{P_j}|}} \quad (4)$$

where N_{P_i} denotes the neighborhood of node P_i , and $|N_{P_i} \cap N_{P_j}|$ denotes the number of nodes that link to both P_i and P_j .

Fig. 3. Three types of connections between P_1 and P_5 . Type-1 ($P_1 \rightarrow P_2, P_2 \rightarrow P_5$): P_1 cites P_2 and P_2 cites P_5 ; Type-2 ($P_1 \rightarrow P_3, P_5 \rightarrow P_3$): P_1 and P_5 cite a common paper P_3 ; Type-3 ($P_4 \rightarrow P_1, P_4 \rightarrow P_5$): P_1 and P_5 are jointly cited by a paper P_4 . (The arrows point in the direction of paper citation).

Based on the Similarity_1 (semantic-based) and Similarity_2 (structure-based), the link weight between two paper nodes in the paper citation graph (G_P) can be represented as follows:

$$W_{i,j} = \lambda_1 \cdot \text{Similarity}_1(P_i, P_j) + \lambda_2 \cdot \text{Similarity}_2(P_i, P_j) \quad (5)$$

where $W_{i,j}$ is the weight from paper i to paper j in G_P , Similarity_1 and Similarity_2 are the semantic-based and structure-based similarities between two papers respectively. Parameters λ_1 and λ_2 are two corresponding coefficients, which can be defined as the following form:

$$\lambda_1 = e^{\mu[\text{Similarity}_1(P_i, P_j) - \varepsilon_1]} \quad (6)$$

$$\lambda_2 = e^{\mu[\text{Similarity}_2(P_i, P_j) - \varepsilon_2]} \quad (7)$$

with μ being a parameter shaping the exponential function, and ε_1 and ε_2 being the media values of Similarity_1 and Similarity_2 respectively. Here let $\mu = 6$ so that those similarity values that exceed the threshold can be constrained by the exponential curve. Parameters λ_1 and λ_2 are normalized as $\lambda_1 + \lambda_2 = 1$.

For a G_P with n papers, the adjacency matrix of the citation network can be denoted as an $n \times n$ matrix (see Fig. 2), where the link weight between two paper nodes can be calculated by:

$$M_{i,j} = \begin{cases} W_{i,j} & \text{if paper } i \text{ cites paper } j \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

Let $\bar{M}$ be the fractionalized citation matrix where $\bar{M}_{i,j} = \frac{M_{i,j}}{\sum_{i=1}^n M_{i,j}}$. Let e be the n -dimensional vector whose elements are all 1 and v be an n -dimensional vector which can be viewed as a personalized vector [34]. Next let $x(v)_{\text{paper}}$ denote the PageRank vector corresponding to the vector $x(v)_{\text{paper}}$, and $x(v)$ can be calculated from $x = \bar{M}x$ where $\bar{M} = d\bar{M} + (1-d)ve^T$. Thus, PageRank vector x can be computed using:

$$x(v)_{\text{paper}} = (1-d)(I - d\bar{M})^{-1}v \quad (9)$$

where d (set at 0.85) is a damping factor. Let $Q = (1-d)(I - d\bar{M})^{-1}$, then $x = Qv$. For any given v , PageRank vector $x(v)$ can be obtained from Qv .

2.3 Link Weighting in Paper-Author Graph (G_{PA})

An academic paper with higher quality implicates the innovativeness and the contribution of the authors behind it [2] [3]. In the paper-author graph (G_{PA}), let $P = \{p_1, p_2, \dots, p_n\}$ denote the set of n papers and $A = \{a_1, a_2, \dots, a_m\}$ denote the set of m authors, then G_{PA} can be represented as an $n \times m$ adjacency matrix, where the link weight $A_{\text{author } i, j}$ from author j to paper i is:

$$A_{\text{author } (i,j)} = \begin{cases} 1 & \text{if author } j \text{ writes paper } i \\ 0 & \text{otherwise} \end{cases} \quad (10)$$

However, the vast majority of scientific articles are multi-authored, and author's contributions to each article is also different in most cases [35]. Actually, the concept of authorship has become a hot topic in recent years, which can be attributed to the proliferation of authors over the years [36] [37]. In this study, the link weights in G_{PA} can be deemed as the level of authors' contributions to their articles. Modified Raw Weight ($W_{R,j}$) [23] is adopted to assess the authors' contributions according to the relative rankings of authors in co-authored publications. The great advantage of $W_{R,j}$ is that the individual contribution to each paper can be evaluated in a relatively accurate manner, regardless of the total number of authors, and how many publications they have written jointly. The experimental results demonstrated that it achieved many practical successes in different scientific fields such as economic management and bioscience [38] [39]. For the author of rank j the Modified Raw Weight is:

$$W_{R,j} = \frac{n - \frac{j}{2} + 1}{\sum_{j=1}^n n_j} = \frac{2n - j + 2}{n \cdot (n + 1)} \cdot \frac{2}{3} \quad (11)$$

where $W_{R,j}$ is the Modified Raw Weight of author j , j is the position of author j in the author list, n is the total number of authors in the paper, and $\sum_{j=1}^n n_j$ is the sum of author positions. Hence, the unweighted G_{PA} can be updated by:

$$A_{\text{author } (i,j)} = \begin{cases} W_{R,j} & \text{if author } j \text{ writes paper } i \\ 0 & \text{otherwise} \end{cases} \quad (12)$$

2.4 Link Weighting in Paper-Journal Graph (G_{PJ})

In the initial P-Rank algorithm, the paper-journal graph (G_{PJ}) can be represented as an $n \times q$ adjacency matrix, where n and q are the number of papers and journals, respectively:

$$A_{\text{journal } (i,j)} = \begin{cases} 1 & \text{if paper } i \text{ is published on journal } j \\ 0 & \text{otherwise} \end{cases} \quad (13)$$

So far as our knowledge is concerned, the vast majority of article ranking algorithms only consider the link weighting in paper citation graph (G_P) and paper-author graph (G_{PA}), while the link weighting in the paper-journal graph (G_{PJ}) is not yet taken into account. From an analysis of experimental results in literature [17], it can be found that there exists a low correlation between paper layer and journal layer since the link weighting is not adopted in G_{PJ} . Clearly, scientific publication and citation relevance need to be studied comprehensively, particularly as a complex network system [40]

[41]. To address this issue, we develop a weighted G_{PJ} in which the corresponding link weight can be updated by the journal impact factors [6] [42]. Similar to G_{PA} , the link weights in G_{PJ} can be regarded as the level of journals' impact to the published articles. As one of the most significant scientific evaluation indicators in SCI, journal impact factor (JIF) is computed by the scientific division of Clarivate Analytics, and can be usually utilized for ranking and assessing the grades of various scientific journals in the Journal Citation Report (JCR) database [7]. Notwithstanding JIF first aims at assessing scientific journals, it is now increasingly used to evaluate research and orient publishing strategies of researchers [2] [43]. In [24], Larivière *et al.* suggested that JIF not only reflects the "quality" of that paper but also that of the journal in which it is published because the possibility that citations to a paper are also influenced by the impact factor of the journal. Here, the "mapminmax" function defined in MATLAB R2018b version is used to normalize the JIF list, the range distributed over the interval [0.1,1]. The formula 13 can thus be rewritten as below:

$$A_{\text{journal}(i,j)} = \begin{cases} \text{Normalize}[\text{JIF}_j] & \text{if paper } i \text{ is published on } j \\ 0 & \text{otherwise} \end{cases} \quad (14)$$

2.5 Weighted P-Rank Algorithm

The weighted P-Rank score of papers can be expressed as $x(v)_{\text{paper}}$ in Eq. 9, where the personalized vector is

$$v = (\phi_1((\frac{x(v)_{\text{author}}}{n_{\text{p_author}}})^T \times A_{\text{author}}^T) + \phi_2((\frac{x(v)_{\text{journal}}}{n_{\text{p_journal}}})^T \times A_{\text{journal}}^T))^T \quad (15)$$

where $n_{\text{p_author}}$ represents a vector with the number of publications for each author, and $n_{\text{p_journal}}$ represents a vector with the number of publications for each journal. The mutual dependence (intra-class and inter-class walks) of papers, authors, and journals is coupled by the parameters ϕ_1 and ϕ_2 , which are set at 0.5 as default. The weighted P-Rank scores of author and journal can be expressed as:

$$x(v)_{\text{author}} = A_{\text{author}}^T \times x(v)_{\text{paper}} \quad (16)$$

$$x(v)_{\text{journal}} = A_{\text{journal}}^T \times x(v)_{\text{paper}} \quad (17)$$

As discussed in Sect. 1, in the real citation network, articles are published and cited in time sequence. Such approaches do not consider the dynamic nature of the network and are always biased to old publications. Therefore, the recent articles tend to be underestimated due to the lack of enough citations. To tackle this problem, Sayyadi and Getoor [4] proposed a time-aware method to achieve a better ranking result in the citation network with time information. In [16] [17], Wang *et al.* and Zhang *et al.* employed the time information of article to boost the prestige of recent published articles and thus improve the ranking accuracy. In this study, we also adopt a time weight T_i to eliminate the bias to earlier publications, which can be regarded as a personalized PageRank vector. Here according to the time-aware method proposed in FutureRank [4], the function T_i is defined as:

$$T_i = e^{-\rho \times (T_{\text{current}} - T_{\text{publish}})} \quad (18)$$

where T_{publish} denotes the publication time of paper i , and $T_{\text{current}} - T_{\text{publish}}$ denotes the number of years since the paper i was published. ρ is a constant value set to be 0.62 based on FutureRank [4]. The sum of T_i for all the articles is normalized to 1.

Taken together, the weighted P-Rank score of a paper can be calculated by:

$$x(v)_{\text{paper}} = \gamma \cdot \text{Pagerank}(\overline{M}, v) + \delta \cdot T + (1 - \gamma - \delta) \cdot \frac{1}{n_p} \quad (19)$$

with parameters γ and δ being constants of the algorithm. $(1 - \gamma - \delta) \cdot \frac{1}{n_p}$ represents the probability of random jump, where n_p is the number of paper samples.

In the proposed algorithm, the initial score of each paper is set to be $\frac{1}{n_p}$. For articles which do not cite any other papers, we suppose that they hold links to all the other papers. Hence, the sum of $x(v)_{\text{paper}}$ for all the papers will keep to be 1 in each iteration. The steps above are recursively conducted until convergence (threshold is set at 0.0001). The pseudocode of the weighted P-Rank algorithm is given in Algorithm 1.

Algorithm 1: Weighted P-Rank Algorithm Based on Heterogeneous Network**Input** : G_P, G_{PA}, G_{PJ} , JIF list of all journals, and time list of all papers**Output** : Weighted P-Rank score of paper $x(v)_{\text{paper}}$ **Parameters** : $\alpha, \beta, \gamma, \delta, \rho, \lambda_1, \lambda_2, \mu, \varepsilon_1, \varepsilon_2, d, \varphi_1, \varphi_2$ **Steps** :

- 1 Initialize all the scores of papers: $x(v)_{\text{paper}} = \frac{\text{ones}(n_p, 1)}{n_p}$, where n_p is the number of paper samples
- 2 Normalize JIF of each journal in dataset: $\text{mapminmax}[J] \leftarrow \text{JIF list } (J)$
- 3 Compute and normalize time score of each paper based on Eq. 18: $T_i = \text{Normalize}[e^{-\rho \times (T_{\text{current}} - T_{\text{publish}})}]$
- 4 Update G_P by Eqs. 5 and 8: $C_w \leftarrow C$
- 5 Update G_{PA} by Eqs. 11 and 12: $A_w \leftarrow A$
- 6 Update G_{PJ} with Eq. 14: $J_w \leftarrow J$
- 7 **while not converging do**
- 8 Eq. 17: $x(v)_{\text{journal}} = A_{\text{journal}}^T \times x(v)_{\text{paper}}$
- 9 Eq. 16: $x(v)_{\text{author}} = A_{\text{author}}^T \times x(v)_{\text{paper}}$
- 10 Eq. 15: $v = (\varphi_1((\frac{x(v)_{\text{author}}}{n_{p_author}})^T \times A_{\text{author}}^T) + \varphi_2((\frac{x(v)_{\text{journal}}}{n_{p_journal}})^T \times A_{\text{journal}}^T))^T$
- 11 Calculate $\text{Pagerank}(\bar{M}, v)$
- 12 Update the score of each paper based on time information (Eq. 19):

$$x(v)_{\text{paper}} = \gamma \cdot \text{Pagerank}(\bar{M}, v) + \delta \cdot T + (1 - \gamma - \delta) \cdot \frac{1}{n_p}$$
- 13 **end**
- 14 **return** $x(v)_{\text{paper}}, x(v)_{\text{author}},$ and $x(v)_{\text{journal}}$

3 Experiments

In this section, the weighted P-Rank algorithm is comprehensively evaluated under different conditions and graph configurations. Moreover, extensive experiments are also conducted to analyze the robustness and ROC performance of the proposed method on three different datasets.

3.1 Datasets and Experimental Settings

Three public datasets are used in this study, i.e. arXiv (hep-th), Cora, and MAG. The three datasets were chosen since they should be better suited to showing general results. Moreover, the convergence rate and the robustness of the proposed algorithm could be tested on three different datasets. In this study, each article can be featured by seven items, namely article ID, article title, author list, published journal, JIF, publication time, and article score. An example of article list information is illustrated in Fig. 4. The summary statistics of three datasets are listed in Tab. 1. It is worth remarking that the A_{journal} values of all conference articles were sampled from the average JIF of all journals calculated in the corresponding dataset.

Fig. 4. An example of article list information

Table 1. The Datasets Utilized in Experiments

Dataset	Articles	Citations	Authors	Journals
arXiv	28,500	350,000	14,500	410
Cora	16,252	43,850	12,348	8156
MAG	15,640	200,483	26,430	9575

All experiments are conducted on a computer with 3.30GHz Intel i9-7900X processor and 64GB RAM under Linux 4.15.0 operating system. The program codes of data preprocessing and graphs modeling are written by Python 3.6.5, which is available on <https://github.com/pjzj/Weighted-P-Rank>.

3.2 Evaluation Metrics

Spearman's Rank Correlation

The evaluation of article ranking has always been a very challenging task because the real "quality" of article is difficult to be quantified in practice. Moreover, the ranking results will be strongly influenced by using different ranking indicators [44] [45]. In 2009, Sayyadi and Getoor suggested that the future PageRank score can be regarded as ground truth [4]. However, certain old articles achieved higher PageRank scores since the PageRank algorithm itself is biased to old publications. Wang *et al.* [16] and Zhang *et al.* [17] adopted future citation number as the ground truth, and then they evaluated the similarity between the estimated rank and ground truth rank by calculating the Spearman's rank correlation coefficient. In this paper, Spearman's rank correlation coefficient is also used to assess the performance of proposed algorithm under different conditions. For a dataset $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N] \in \mathbb{R}^{D \times N}$ with N samples, N original data are converted into grade data, and the correlation coefficient ρ can be calculated by:

$$\rho = \frac{\sum_{i=1}^n (R_1(P_i) - \bar{R}_1)(R_2(P_i) - \bar{R}_2)}{\sqrt{\sum_{i=1}^n (R_1(P_i) - \bar{R}_1)^2 \sum_{i=1}^n (R_2(P_i) - \bar{R}_2)^2}} \quad (20)$$

where $R_1(P_i)$ denotes the position of paper P_i in the first rank list, $R_2(P_i)$ denotes the position of paper P_i in the second rank list, and $\bar{R}_1$ and $\bar{R}_2$ denote the average rank positions of all papers in the two rank lists respectively.

Robustness

Robustness is an evaluation indicator that can be used to reflect the anti-interference (consistency) ability of system or algorithm. Here according to the corresponding historical time point on three datasets, the whole time on each dataset can be divided into two periods. The time period before the historical time point can be denoted as T_1 , while the whole period can be denoted as T_2 . The robustness of algorithm can thus be measured by calculating the correlation of ranking scores in T_1 and T_2 . This shows that the higher the correlation between two rank lists, the higher the robustness of the algorithm would tend to be.

3.3 Experimental Results

Graph configurations:

Two parameters can be set in graph configurations: φ_1 and φ_2 . If $\varphi_1 = 0$ and $\varphi_2 = 0$, there exists no coupling, which can be regarded as the situation of ranking the papers using PageRank algorithm only. Of course, if G_{PA} or G_{PJ} is introduced into the network, the parameters can be updated as $\varphi_1 = 1, \varphi_2 = 0$ or $\varphi_1 = 0, \varphi_2 = 1$. The final heterogeneous graph is composed of one intra-walk (G_P) and two inter-walks (G_{PA} and G_{PJ}). By using various combinations of graphs, we compare and assess four different cases of P-Rank algorithm with previous works. The cases and the associated parameters are listed below:

- G_P ($\varphi_1 = 0, \varphi_2 = 0$): which is the traditional PageRank algorithm for rank calculation.
- $G_P + G_{PA}$ ($\varphi_1 = 1, \varphi_2 = 0$): A new graph (G_{PA}) is introduced into the heterogeneous network which only utilizes citation and authorship.
- $G_P + G_{PJ}$ ($\varphi_1 = 0, \varphi_2 = 1$): A new graph (G_{PJ}) is introduced into the heterogeneous network which only utilizes citation and journal information.

- $G_P + G_{PA} + G_{PJ}$ ($\phi_1 = 0.5, \phi_2 = 0.5$): Two new graphs (G_{PA} and G_{PJ}) are introduced into the heterogeneous network which uses citation, authorship, and journal information simultaneously.

Time settings:

Based on whether to use time information, there exist two kinds of settings:

- No-Time ($\delta = 0$): which does not utilize article time information to enhance the effect of the recent published articles.
- Time-Weighted (see Eq. 19): which can be used to balance the bias to earlier published articles in the citation network.

With these assumptions, we are now ready to verify Spearman’s ranking correlation of different cases on three datasets, as shown in Tabs. 2-5. From an analysis of Tab. 2, it can be found that the best performance (arXiv: 0.5449; Cora: 0.3352; MAG: 0.4994) of proposed algorithm is all achieved from the weighted graph configurations as follows: $G_P + G_{PA} + G_{PJ}$. In addition, we note that under the four graph configuration conditions (G_P ; $G_P + G_{PA}$; $G_P + G_{PJ}$; $G_P + G_{PA} + G_{PJ}$), an important observation from the experimental results is that weighted graphs significantly outperform unweighted graphs. This result seems to show that article ranking might benefit from link weighting in heterogeneous scholarly network.

Table 2. Spearman’s ranking correlation of different graph configurations on three datasets.

Graph Configurations	arXiv		Cora		MAG	
	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted
G_P	0.4153	0.4339	0.2607	0.2793	0.3521	0.3764
$G_P + G_{PA}$	0.4133	0.4490	0.2879	0.3096	0.4125	0.4530
$G_P + G_{PJ}$	0.4082	0.4273	0.2730	0.2894	0.4049	0.4254
$G_P + G_{PA} + G_{PJ}$	0.4915	0.5449	0.3135	0.3352	0.4748	0.4994

Table 3. Spearman’s ranking correlation of two time settings on arXiv dataset.

Time Settings	G_P		$G_P + G_{PA}$		$G_P + G_{PJ}$		$G_P + G_{PA} + G_{PJ}$	
	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted
No-Time	0.4153	0.4339	0.4133	0.4490	0.4082	0.4273	0.4915	0.5449
Time-Weighted	0.5880	0.6228	0.5616	0.6496	0.5800	0.6574	0.6753	0.7115

Table 4. Spearman’s ranking correlation of two time settings on Cora dataset.

Time Settings	G_P		$G_P + G_{PA}$		$G_P + G_{PJ}$		$G_P + G_{PA} + G_{PJ}$	
	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted
No-Time	0.2607	0.2793	0.2879	0.3096	0.2730	0.2894	0.3135	0.3352
Time-Weighted	0.3120	0.3490	0.3593	0.3848	0.3116	0.3729	0.3772	0.3962

Table 5. Spearman’s ranking correlation of two time settings on MAG dataset.

Time Settings	G_P		$G_P + G_{PA}$		$G_P + G_{PJ}$		$G_P + G_{PA} + G_{PJ}$	
	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted	Unweighted	Weighted
No-Time	0.3521	0.3764	0.4125	0.4530	0.4049	0.4254	0.4778	0.4994
Time-Weighted	0.4245	0.5051	0.4693	0.5474	0.4500	0.5139	0.5548	0.5933

The best performance is highlighted in bold.

By comparing and analyzing the data from Tabs. 3-5, under the conditions of two time settings (No-Time and Time-Weighted), it can be seen that the performance of Time-Weighted configurations always outperform the results of corresponding No-Time configurations, and the best performance (arXiv: 0.7115; Cora: 0.3962; MAG: 0.5933) is obtained by jointly utilizing all kinds of configurations as follows: $G_P + G_{PA} + G_{PJ}$ +Time-Weighted. The experimental

results demonstrate that Time-Weighting is beneficial to improve the reasonability of ranking algorithm, which is completely consistent with the conclusion in Ref. [16].

For better comparison, we also measure the performance of the weighted P-Rank and five famous algorithms (PageRank, FutureRank, HITS, CiteRank, and P-Rank) on three datasets by using Spearman's rank correlation and robustness. Note that all the algorithms above are implement with the optimal parameters. In addition, for each dataset, a suitable historical time point is set based on the distribution of papers over a period of time. We see from Fig. 5 that weighted P-Rank achieved superior rank correlation (arXiv: 0.707; Cora: 0.388; MAG: 0.599) and robustness performance (arXiv: 0.918; Cora: 0.484; MAG: 0.732), in particular compared to the initial P-Rank algorithm.

Fig. 5. Spearman's ranking correlation and robustness of six algorithms on three datasets.

To further verify the performance of each ranking algorithm, we also conduct experiments on three datasets by using ROC curve and AUC value. Specifically, we first rank all the articles in the dataset based on PageRank algorithm. We then set a median threshold to hold different numbers of articles. According to the threshold, all articles can be divided into two classes i.e., higher ranked articles (positive objects) and lower ranked articles (negative objects). To make the results more reliable, 10 times of independent experiments are conducted on each test dataset and the average results are obtained (see Fig. 6).

Fig. 6. ROC curves obtained by 6 ranking algorithms (Weighted P-Rank, P-Rank, PageRank, FutureRank, HITS, and CiteRank) on three different datasets.

It can be seen from Fig. 6 that weighted P-Rank algorithm (as plotted by red curve) significantly outperforms other ranking algorithms on arXiv and MAG datasets, while it achieves competitive performance on Cora dataset. It shows

better performance and generalization ability of weighted P-Rank algorithm on three datasets. The AUC vales obtained by weighted P-Rank on arXiv, Cora, and MAG datasets are 0.6733, 0.5586, and 0.6593 respectively. By a sharp contrast, the AUC values achieved by initial P-Rank algorithm are unsatisfactory, especially on arXiv dataset (only 0.3461). This result indicates that link weighting plays an important role in heterogeneous graphs, which will be very helpful to improve the performance of the article ranking algorithm.

4 Conclusion

Assessing scientific papers is an important but challenging task, mainly due to the dynamic nature and complexity of the heterogeneous scholarly network. This paper developed a weighted P-Rank algorithm based on a heterogeneous scholarly network for article ranking evaluation. The study is dedicated to assigning weight to the corresponding links in G_P , G_{PA} , and G_{PJ} by calculating citation relevance (G_P), authors' contribution (G_{PA}), and journals' contribution (G_{PJ}). The experiments are conducted on three public datasets (arXiv, Cora, and MAG). The performance of the proposed weighted P-Rank method is measured by using three evaluation metrics (Spearman's ranking correlation, robustness, and ROC curves). Moreover, under conditions of two weighting combinations (Unweighted and Weighted) and four graph configurations (G_P , $G_P + G_{PA}$, $G_P + G_{PJ}$, and $G_P + G_{PA} + G_{PJ}$), the performance of weighted P-Rank algorithm is further evaluated and analyzed. The experimental results showed that the weighted P-Rank method achieved promising performance on three different datasets, and the best ranking result can be achieved by jointly employing all kinds of weighting information. Additionally, we note that the article ranking result can be further improved by utilizing time-weighting information, which is nicely demonstrated by the comparsion results between "Time-Weighted" and the corresponding "No-Time" settings. In summary, it can be found that link weighting in heterogeneous graphs will be very helpful to improve the performance and reasonability of the article ranking method.

In the future, a series of meaningful studies can be conducted subsequently, combining network topology and link weighting. In addition to article ranking algorithms such as P-Rank and normal PageRank, the advantages of various approaches and certain factors that can be weighted should be taken into account. For instance, we would test the effect of link weighting on more ranking methods and verify how the parameters influence the performance of the algorithms. Furthermore, one may consider employing some machine learning methods such as principal component analysis (PCA), linear boundary discriminant analysis (LBDA) and Laplacian score (LS) for further analyzing the correlation between weighted P-Rank score and article citation counts.

Data Availability

The program code will be made available on <https://github.com/pjzj/Weighted-P-Rank>.

Acknowledgements

This work was supported by Fundamental Research Funds for the Central Universities (Grants 2019RC29 and DUT19RC(3)012), by National Natural Science Foundation (NNSF) of China (Grants 61867005, 61972064 , and 61672130), by the Gansu Provincial First-Class Discipline Program of Northwest Minzu University (Grant 11080305), and by LiaoNing Revitalization Talents Program (Grant XLYC1806006).

Conflict of Interest

The authors declare that they have no competing interests regarding the publication of this paper.

References

- [1] L. Cai, J. Tian, J. Liu, X. Bai, I. Lee, X. Kong, and F. Xia, "Scholarly impact assessment: A survey of citation weighting solutions," *Scientometrics*, vol. 118, no. 2, pp. 453–478, 2019.

- [2] J. Zhou, N. Cai, Z.-Y. Tan, and M. J. Khan, "Analysis of effects to journal impact factors based on citation networks generated via social computing," *IEEE Access*, vol. 7, pp. 19775–19781, 2019.
- [3] J. Zhou, L. Feng, N. Cai, and J. Yang, "Modeling and simulation analysis of journal impact factor dynamics based on submission and citation rules," *Complexity*, no. 3154619, 2020.
- [4] H. Sayyadi and L. Getoor, "Futurerank: Ranking scientific articles by predicting their future pagerank," in *Proceedings of the 2009 SIAM International Conference on Data Mining*, pp. 533–544, SIAM, 2009.
- [5] L. Feng, J. Zhou, S.-L. Liu, N. Cai, and J. Yang, "Analysis of journal evaluation indicators: An experimental study based on unsupervised Laplacian Score," *Scientometrics*, vol. 124, pp. 233–254, 2020.
- [6] E. Garfield, "Citation analysis as a tool in journal evaluation," *Science*, vol. 178, no. 4060, pp. 471–479, 1972.
- [7] M. Bordons, M. Fernández, and I. Gómez, "Advantages and limitations in the use of impact factor measures for the assessment of research performance," *Scientometrics*, vol. 53, no. 2, pp. 195–206, 2002.
- [8] J. E. Hirsch, "An index to quantify an individual's scientific research output," *Proceedings of the National academy of Sciences*, vol. 102, no. 46, pp. 16569–16572, 2005.
- [9] T. Braun, W. Glänzel, and A. Schubert, "A hirsch-type index for journals," *Scientometrics*, vol. 69, no. 1, pp. 169–173, 2006.
- [10] L. Page, "The pagerank citation ranking: Bringing order to the web. technical report," *Stanford Digital Library Technologies Project*, 1998, 1998.
- [11] X. Liu, J. Bollen, M. L. Nelson, and H. Van de Sompel, "Co-authorship networks in the digital library research community," *Information processing & management*, vol. 41, no. 6, pp. 1462–1480, 2005.
- [12] J. Bollen, M. A. Rodriguez, and H. Van de Sompel, "Journal status," *Scientometrics*, vol. 69, no. 3, pp. 669–687, 2006.
- [13] 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*, vol. 2007, no. 6, pp. 1–5, 2007.
- [14] E. Yan, Y. Ding, and C. R. Sugimoto, "P-rank: An indicator measuring prestige in heterogeneous scholarly networks," *Journal of the American Society for Information Science and Technology*, vol. 62, no. 3, pp. 467–477, 2011.
- [15] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM (JACM)*, vol. 46, no. 5, pp. 604–632, 1999.
- [16] Y. Wang, Y. Tong, and M. Zeng, "Ranking scientific articles by exploiting citations, authors, journals, and time information," in *Twenty-seventh AAAI conference on artificial intelligence*, 2013.
- [17] Y. Zhang, M. Wang, F. Gottwalt, M. Saberi, and E. Chang, "Ranking scientific articles based on bibliometric networks with a weighting scheme," *Journal of Informetrics*, vol. 13, no. 2, pp. 616–634, 2019.
- [18] P. Chen, H. Xie, S. Maslov, and S. Redner, "Finding scientific gems with googles pagerank algorithm," *Journal of Informetrics*, vol. 1, no. 1, pp. 8–15, 2007.
- [19] N. Ma, J. Guan, and Y. Zhao, "Bringing pagerank to the citation analysis," *Information Processing & Management*, vol. 44, no. 2, pp. 800–810, 2008.
- [20] S. Maslov and S. Redner, "Promise and pitfalls of extending google's pagerank algorithm to citation networks," *Journal of Neuroscience*, vol. 28, no. 44, pp. 11103–11105, 2008.
- [21] D. Zhou, S. A. Orshanskiy, H. Zha, and C. L. Giles, "Co-ranking authors and documents in a heterogeneous network," in *Seventh IEEE international conference on data mining (ICDM 2007)*, pp. 739–744, IEEE, 2007.
- [22] P. M. Davis, "Eigenfactor: Does the principle of repeated improvement result in better estimates than raw citation counts?," *Journal of the American Society for Information Science and Technology*, vol. 59, no. 13, pp. 2186–2188, 2008.
- [23] F. J. Trueba and H. Guerrero, "A robust formula to credit authors for their publications," *Scientometrics*, vol. 60, no. 2, pp. 181–204, 2004.
- [24] V. Larivière and Y. Gingras, "The impact factor's matthew effect: A natural experiment in bibliometrics," *Journal of the American Society for Information Science and Technology*, vol. 61, no. 2, pp. 424–427, 2010.

- [25] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.
- [26] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, “Natural language processing (almost) from scratch,” *Journal of Machine Learning Research*, vol. 12, no. Aug, pp. 2493–2537, 2011.
- [27] T. Mikolov, W.-t. Yih, and G. Zweig, “Linguistic regularities in continuous space word representations,” in *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 746–751, 2013.
- [28] R. Navigli and S. P. Ponzetto, “Babelnet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network,” *Artificial Intelligence*, vol. 193, pp. 217–250, 2012.
- [29] M. T. Pilehvar, D. Jurgens, and R. Navigli, “Align, disambiguate and walk: A unified approach for measuring semantic similarity,” in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, vol. 1, pp. 1341–1351, 2013.
- [30] G. A. Miller, *WordNet: An electronic lexical database*. MIT press, 1998.
- [31] J. Ye, “Cosine similarity measures for intuitionistic fuzzy sets and their applications,” *Mathematical and Computer Modelling*, vol. 53, no. 1-2, pp. 91–97, 2011.
- [32] S. Liu, L. Feng, and H. Qiao, “Scatter balance: An angle-based supervised dimensionality reduction,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 2, pp. 277–289, 2014.
- [33] X. Jiang and H. Zhuge, “Forward search path count as an alternative indirect citation impact indicator,” *Journal of Informetrics*, vol. 13, no. 4, p. 100977, 2019.
- [34] T. Haveliwala, S. Kamvar, and G. Jeh, “An analytical comparison of approaches to personalizing pagerank,” tech. rep., Stanford, 2003.
- [35] E. A. Correa Jr., F. N. Silva, D. F. C. Luciano, and D. R. Amancio, “Patterns of authors contribution in scientific manuscripts,” *Journal of Informetrics*, vol. 11, no. 2, pp. 498–510, 2016.
- [36] J. P. Drenth, “Multiple authorship: the contribution of senior authors,” *JAMA*, vol. 280, no. 3, pp. 219–221, 1998.
- [37] S. R. Sahu and K. C. Panda, “Does the multi-authorship trend influence the quality of an article?,” *Scientometrics*, vol. 98, no. 3, pp. 2161–2168, 2014.
- [38] Y. Berker, “Golden-ratio as a substitute to geometric and harmonic counting to determine multi-author publication credit,” *Scientometrics*, vol. 114, no. 3, pp. 839–857, 2018.
- [39] J. Xu, Y. Ding, M. Song, and T. Chambers, “Author credit-assignment schemas: A comparison and analysis,” *Journal of the Association for Information Science and Technology*, vol. 67, no. 8, pp. 1973–1989, 2016.
- [40] P. P. Maglio and P. L. Mabry, “Agent-based models and systems science approaches to public health,” *American Journal of Preventive Medicine*, vol. 40, no. 3, pp. 392–394, 2011.
- [41] J. M. Epstein, *Generative social science: Studies in agent-based computational modeling*. Princeton University Press, 2006.
- [42] E. Garfield, “The history and meaning of the journal impact factor,” *JAMA*, vol. 295, no. 1, pp. 90–93, 2006.
- [43] L. Bornmann and W. Marx, “The journal impact factor and alternative metrics,” *EMBO Reports*, vol. 17, no. 8, pp. 1094–1097, 2016.
- [44] J. H. Zar, “Significance testing of the spearman rank correlation coefficient,” *Journal of the American Statistical Association*, vol. 67, no. 339, pp. 578–580, 1972.
- [45] J. Myers, A. Well, and R. Lorch, “Research design and statistical analysis, routledge,” New York, 2010.