Intelligent Information Management, 2022, 14, 53-70
https://www.scirp.org/journal/iim
ISSN Online: 2160-5920
ISSN Print: 2160-5912

A Bias-Free Time-Aware PageRank Algorithm for Paper Ranking in Dynamic Citation Networks

Moath Abu Dayeh, Badie Sartawi, Saeed Salah

Department of Computer Science, Al-Quds University, Jerusalem, Palestine
Email: moathabudayeh94@gmail.com, sartawi@staff.alquds.edu, sasalah@staff.alquds.edu

How to cite this paper: Dayeh, M.A., Sartawi, B. and Salah, S. (2022) A Bias-Free Time-Aware PageRank Algorithm for Paper Ranking in Dynamic Citation Networks. *Intelligent Information Management*, **14**, 53-70.
https://doi.org/10.4236/iim.2022.142004

Received: December 29, 2021
Accepted: February 19, 2022
Published: February 22, 2022

Abstract

The process of ranking scientific publications in dynamic citation networks plays a crucial rule in a variety of applications. Despite the availability of a number of ranking algorithms, most of them use common popularity metrics such as the citation count, h-index, and Impact Factor (IF). These adopted metrics cause a problem of bias in favor of older publications that took enough time to collect as many citations as possible. This paper focuses on solving the problem of bias by proposing a new ranking algorithm based on the PageRank (PR) algorithm; it is one of the main page ranking algorithms being widely used. The developed algorithm considers a newly suggested metric called the Citation Average rate of Change (CAC). Time information such as publication date and the citation occurrence's time are used along with citation data to calculate the new metric. The proposed ranking algorithm was tested on a dataset of scientific papers in the field of medical physics published in the Dimensions database from years 2005 to 2017. The experimental results have shown that the proposed ranking algorithm outperforms the PageRank algorithm in ranking scientific publications where 26 papers instead of only 14 were ranked among the top 100 papers of this dataset. In addition, there were no radical changes or unreasonable jump in the ranking process, *i.e.*, the correlation rate between the results of the proposed ranking method and the original PageRank algorithm was 92% based on the Spearman correlation coefficient.

Keywords

Bibliometric, Citation Analysis, Pagerank Algorithm, Scientific Publications, Metrics, Time-Aware

1. Introduction

Ranking scientific publications is an important task in many aspects, starting

with the research itself in order to improve the quality of research, as it gives an overview of the research output's quality and institutions' world ranking. Researchers also need to prove the impact of their research for several reasons such as satisfying or persuading the funding agencies and improving the scholarly search to get the most relevant publications to specific topics when the research community refers to research databases or search engines.

These issues have prompted to produce the rich history of studies and research in bibliometrics, which is a term commonly given by the scientific community to sets of indicators and measures that are used to refer to the popularity and quality of scientific publications [1]. Citation data is an important source for providing bibliometric metrics and the most used approach in citation analysis is the link-based analysis such as the PageRank (PR) algorithm [2] and the Hyperlink-Induced Topic Search (HITS) algorithm [3].

From a network perspective, publications are represented by nodes, while the directed edges are represented by citations. This network is called a citation network where graphs represent the relationship between documents. These graphs are dynamic in nature and change with time as new publications and citations appear. Furthermore, a citation network is one of the network theory applications which depend on link analysis. It is represented by the adjacency matrix, if we assume a citation network contains N nodes, the presence or absence of an edge between two nodes is represented by 1 or 0, respectively.

The PageRank algorithm is the most widely used algorithm for ranking scientific publications based on citation analysis. However, this algorithm was primarily designed to deal with webpages rather than scientific publications. Compared to webpage networks, publication networks differ in their characteristics in such a way that the PageRank algorithm treats them as static networks rather than dynamic. The PageRank algorithm is highly dependent on the citation count; it implicitly considers the importance of the citing paper in order to assign weights to citations instead of treating them equally. However, this algorithm still depends on the number of citations, *i.e.*, the old papers that took enough time to collect a large number of citations, even if these citations are of little importance to the scientific community will get high scores. This causes the problem of bias in favor of the old publications which have been accumulated over the years without considering the publication date or the citations occurrence time.

The main objective of this research work is to improve the ranking accuracy by solving the problem of bias in favor of old publications compared to new ones. This will support research databases with an appropriate ranking mechanism that ensures queries' results are ranked in a fair manner. Thus, in this paper, we proposed and tested a new ranking algorithm; named a Bias-free Time-aware PageRank algorithm (BTPR)—based on the PageRank (PR) algorithm; it is one of the main page ranking algorithms being widely used. The developed algorithm considers a newly suggested metric called the Citation Average rate of Change (CAC). Time information such as publication date and the citation occurrence's time

are used along with citation data to calculate the new metric. The new algorithm was tested on a well-known dataset of scientific publications in the field of medical physics published in the Dimensions database from years 2005 to 2017.

The remainder of the paper is structured as follows: Section 2 overviews the most recent contributions in ranking scientific publications with a main focus on those based on the PageRank algorithm. Section 3 details the proposed Bias-free Time-aware PageRank algorithm (BTPR) and its new calculated metric. The dataset, experimental results, comparisons, and discussions are detailed in Section 4. In Section 5, we present some statistical methods to validate the results. Finally, in Section 6, we conclude the paper and shed light on some future research lines.

2. Related Work

In recent years, a large number of algorithms for ranking scientific publications have been proposed, especially those that depend on citation network analysis. Kanellos *et al.* [4] conducted a theoretical and experimental study of impact-based ranking. They classified the ranking methods based on two main categories; the awareness of time and the use of side information. Of the many existing ranking algorithms, the PageRank is the most widely used with many versions and extensions, which is the main focus of this study. The original PageRank algorithm was developed by Sergey Brin and Lawrence Page [2] for the purpose of ranking Web pages and query results on search engines such as Google. It relies on links between webpages that refer to the citations or linkages. The links are divided into two types, backlinks, and forward links. A scientific paper is highly rated if it has a large number of backlinks, and it increases whenever these links come from papers with high rating. It is an iterative algorithm, and its values are calculated using the following Equation (1):

$$R(u) = c\sum_{v\in B_u}\frac{R(v)}{N_v} \tag{1}$$

where B_u is the set of backlinks for a specific paper u, and $R(v)$ is the citing paper's value, equal initial values are assigned to all papers at the beginning, N_v is the number of forward links of a specific paper v, and c is the normalization factor. The original version of this algorithm was extensively used for evaluation purposes in various applications. For example, it was applied on authors' network to investigate the influence of authors by Liu *et al.* [5]. It was also applied by Bollen *et al.* [6] and Chen *et al.* [7] on citation networks to evaluate scientific papers. Yao, L., *et al.* [8] proposed a modified version of the PageRank algorithm by introducing the nonlinearity principle in order to improve the algorithm against malicious citations. In general, the PageRank score refers to the possibility of choosing a scientific publication by a random user through simulating the random search process. This is to enable the researcher to begin reading a random paper and then

moving to another one listed in the references section.

To consider the dynamic nature of the citation network and aging characteristics, and to alleviate the problem of bias in favor of old publications, a number of studies was conducted to produce time-aware ranking algorithms. Most of these algorithms are modified versions of the original PageRank algorithm. Some of these versions applied modifications on the adjacency matrix, the assumption here is that the more recent information is often preferred by the random researcher who avoids references to old publications. Therefore, these algorithms do not treat citations equally and effect on citation count variant by using time quantities such as citation age, or citation gap and add it to the adjacency matrix. Weighted Citation (WC) algorithm [9] is an example of this approach. It uses a weighted citation matrix by the time quantity called citation gab which is the elapsed time from the publication date of the cited paper until the citation occurs, this time quantity also used by [10] for the same purposes.

Ghosh *et al.* [11] proposed another algorithm called Retained Adjacency Matrix (RAM). In this algorithm, the cited paper's age affects the citation value. It gives a higher value to the link coming from a recent paper and the paper's associated value decreases with age. This algorithm assumes that more recent information is often preferred by researchers. The parameter ($\gamma < 1$) is used to give a higher weight to a recent paper, and this weight decreases with the age of that paper. Given v, the correlated value with the citation link for a specific paper published in year t_n, a scaled down value $\gamma^{n_i} v$ is the correlated value with a citation link paper published in year t_{n-n_i}. Therefore, any paper published in year t_n will be given a higher weight than those published earlier. The retained adjacency matrix is constructed using Equation (2):

$$R_{n,\gamma}(i,j) = \begin{cases} \gamma^{N-n_i}, & \text{if } p_i \text{ cites } p_j \text{ and } t(p_i) = t_{n_i} \le t_n \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where γ is the retention probability, N is the current date, and n_i is the publication date of a paper i.

In the same line, Dunaiski and Visser [12] proposed a new algorithm called NewRank (NR). It assigns weights to citations depending on the cited paper's age based on landing probabilities. Given p, a vector that includes the probabilities of choosing a paper i where $p_i = \mathrm{e}^{-t_i/\tau}$. t_i represents the paper's age and τ is the characteristic of a decay time.

Let $D(p_j)$ is the probability of reaching a reference from paper j, it can be calculated using the following Equation (3):

$$D(p_j) = \frac{p_j}{\sum_{pk \in N+(p_i)} p_{pk}} \tag{3}$$

Equation (3) normalizes the paper's initial value through the initial values of other papers in their references' list. The main objective is to direct the random researcher to recent research to get better citation counts compared to old research.

The NewRank algorithm adopts the iterative approach used in the PageRank algorithm. As a result, recent research has the potential to obtain more citations than the old work.

Some studies have produced other types of weights. For example, Giuffrida *et al.* [13] suggested a model for evaluating citations by the impact of the citing papers. In [14], a weighting mechanism is used based on the citations count in the cited paper divided by its age. Also in [15], a citation age was used as a time quantity. Wei *et al.* [16] integrated the Text Similarity Approach (TSA) to bypass restrictions of the traditional PageRank algorithm in the context of standard citation networks. Compared to the original PageRank algorithm which gives equal values to the downstream nodes, their proposed modification gives different importance weights for the downstream nodes using the cosine similarity algorithm which calculates the text similarity score between each pair of nodes (publications) with a citation relationship.

Some other researchers suggested modifications to the original PageRank algorithm by utilizing the landing probabilities, where the researcher prefers choosing new papers during the random jumps. The papers landing probabilities decay exponentially with their ages, instead of assigning all papers equal landing probability. This means that recent papers have a higher probability of appearing to the random researcher [4].

Walker *et al.* [17] introduced the CiteRank algorithm. It uses the time-aware landing probabilities approach. In this work, a random walk model was developed to predict future citations by relying on time information by assuming that the researcher always starts his/her research from a recent publication and then moves to an older publication and so on until he/she is satisfied. In this algorithm, the probability of initially choosing of a specific paper i is calculated using Equation (4):

$$p_i = \mathrm{e}^{-ege_i/\tau_{dir}} \tag{4}$$

and the CiteRank traffic is calculated using Equation (5):

$$S = 1\cdot \overline{p} + (1-a)W\cdot \overline{p} + (1-a)^2 W^2 \cdot \overline{p} + \cdots \tag{5}$$

where p_i is the probability of choosing a paper i, w is the adjacency matrix that represents the citation network, and a and τ are constant values.

Another algorithm called FutureRank [18] uses the landing probabilities approach designed to capture the dynamic nature of the publication networks. In addition to the citation network, the author's reputation and time information are used in order to generate future citations for recent papers based on several assumptions such as newly published papers are more useful, and a good research paper is written by highly reputable researchers, so that the value of a particular author is distributed on the papers that he/she authored, and the value of the paper is distributed among its authors.

Kanellos, I., *et al.* [19] proposed the AttRank algorithm where they discussed a new mechanism for ranking scientific papers based on their Short-Term Impact

(STI), it is measured by the near future citation count (in some previous works it is called future citations [18], or new citations [17]). The new mechanism depends on determining where new citations stop. The hypothesis that recent citations greatly affect the STI was tested to remain within a certain extent across citation networks.

Some works argued that additional information should be used alongside the citation network to improve ranking algorithms, such as using the paper metadata (e.g., journal information, venues information, and authors' information). Getting scores based on paper metadata can be done in two ways, by conducting statistical calculations on papers' scores (e.g., average paper scores for authors or venues), or by using measures such as author H-index [20] or journal Impact Factor (IF) [21]. Most algorithms in this category exploit papers' metadata in PageRank-like models, in order to modify the citation matrices such as the work in [9]. Also other algorithms aimed to modify both the citation matrices and the random jump probabilities such as the work in [22].

In another research line, some authors incorporated side information by conducting analysis over multiple networks such as paper-author network, paper-journal network, and paper-venue network. In [23], the authors proposed the Hyperlink-Induced Topic Search (HITS) algorithm. The basic idea behind it is to create two-side graphs with different types of nodes (hubs and authorities) where the nodes on both sides of the graph mutually reinforce each other. This approach is followed by the PaperRank algorithm [24] that depends on the indirect relationships between scientific papers, instead of the traditional relationships that are represented by citations. Another algorithm was presented by Wang *et al.* [25] focused on addressing the problem of ranking scientific publications in a heterogeneous network. In addition to that it presented the problem of ignoring time information in the ranking process, it depends on using multiple networks that include citations, journals, authors, and time information. In [26], both the PageRank and HITS algorithms features are combined in order to provide influence model in paper citation networks. This hybrid approach can be also used in an iterative process on a single graph containing a heterogeneous set of nodes, all calculations done based on this graph and the scores are propagated among all nodes' types as detailed in [27] and [28].

In summary, existing solutions that were introduced to bypass the problem of bias created new issues. Some solutions use time information such as publication date to influence the ranking result arbitrarily. In this case, old publications that are still valuable and frequently cited will not be ranked fairly. Others have relied on certain assumptions in order to anticipate future citations and improve the ranking of recent papers such as more recently published papers are more useful, and the author with a good reputation will always have valuable publications, but it is not permissible to generalize these assumptions, despite their validity in most cases. Also, the citation behavior is not fixed and may be affected by many factors that might not be considered at the time of creating these expectations.

To handle the problem of bias in ranking scientific publications, in this paper, a Bias-free Time-aware PageRank algorithm (BTPR) was developed that considers a newly suggested metric called the Citation Average rate of Change (CAC).

3. The Proposed Algorithm

As previously mentioned, the main objective of this research work is to improve the accuracy of the paper ranking process by considering the citation network dynamic nature and its changing over time. To do that, we proposed an extension to the PageRank algorithm, named BTPR, considering a newly suggested ranking metric, we called it Citation Average rate of Change (CAC). This metric measures the change in reliance on a particular paper, whether it is recent or old. The new metric ensures fairness and minimizes bias in favor of old publications, *i.e.*, if the paper was published in the past and obtained a large number of citations during its life, but few of these citations occurred recently and the number of its citations are constantly decreasing, this means that the paper is no longer important, and its ranking value must be reduced. While papers that still receive continuous citations, will receive good scores and their values may not be underestimated only because the date of their publication is old.

The newly developed extension avoids generalizing assumptions to solve the problem of bias, such as assuming that a reputable author always produces valuable literature, or assuming that recent publications are always more valuable than older publications. Also, time information such as publication date cannot be used to influence ranking results arbitrarily by increasing or decreasing the score of a particular paper based on its recency.

3.1. Citation Average Rate of Change (CAC)

CAC is the newly proposed metric that gives a clear perception of the ability of an old scientific paper to keep giving and being important in its field by consistently appearing in the reference list of recent papers. It is not sufficient for the paper to receive a large number of citations to obtain a good ranking, because the citation date also matters. On the other hand, for recent papers that are still in the growth process, we can identify the nature of this growth. If the citation rate is high and is increasing year after year, it indicates that the paper is valuable and will receive many citations in the future. Therefore, it must be given good rankings, instead of solely relying on the citation count as these papers are still new and did not take enough time to collect many citations. Equation (6) [29] calculates the average rate of change:

$$\text{average rate of change} = \frac{f(x_2) - f(x_1)}{x_2 - x_1} \tag{6}$$

where f is a function that depends on x, and the number of citations (C) in t years. Thus, the modified formula becomes as shown in Equation (7):

$$\mathrm{CAC} = \frac{C(t_2) - C(t_1)}{t_2 - t_1} = \frac{\Delta C}{\Delta t} \tag{7}$$

3.2. The Additional Information in the Citation Network

The citation network should contain the publication date for each paper. This is in order to calculate the cited paper's age. Also, to identify the time when each citation occurred using the publication date of the citing paper. For more illustrations, Figure 1 shows a simple example of a citation network containing 15 papers published over 5 years. By making a simple comparison between nodes (1) and (2), which are the oldest in the network, they both have 3 citations. But the CAC for Node (1) is higher because it gets citations continuously from recent papers. As for Node (2), the citation on it stopped three years ago, which means that the reliance on it is declining; hence paper 2 should receive a lower score.

3.3. Bias-Free Time-Aware PageRank

The PageRank algorithm does not depend on the citation count directly, but rather it considers the citing papers' values that mainly depend on the number of citations leading to a bias in the ranking process. To get better ranking results, in the new proposed algorithm (BTPR), the PageRank score is modified according to the following equations (Equations (8) and (9)):

$$\mathrm{BTPR} = c \sum_{v \in B_u} \frac{R(v)}{N_v} + \left(\frac{\Delta C}{\Delta t} \right) \tag{8}$$

$$\mathrm{BTPR} = \mathrm{PR} + \mathrm{CAC} * S \tag{9}$$

where PR is the original PageRank score, CAC is the suggested citation average rate of change, and S is a scale value.

To illustrate how the new metric affects the ranking process, Table 1 shows a sample of papers taken from the Dimension dataset. As previously mentioning, in this table, the first two papers are recent, but they have a high CAC value which indicate that regardless of the number of citations, they are constantly increasing, therefore they have promoted in their rank. In contrast, papers 3 and 4

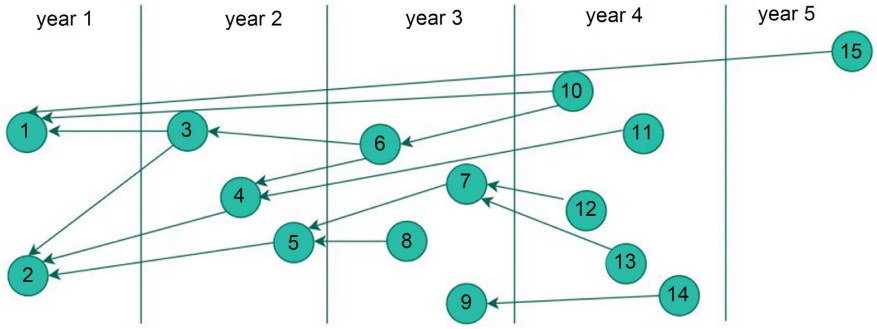

Figure 1. An example of a time-aware citation network.

Table 1. A sample of papers with their PageRank and BTPR based ranks.

Paper	Pub. date	PageRank	CAC	BTPR
1	2017	280	14	34
2	2016	48	7.75	22
3	2008	22	1.6	24
4	2005	19	1.8	16
5	2009	54	−0.3	72
6	2006	76	−0.357	101

are old and still received ranking scores close to the PageRank based score, the reason is that these papers are keep receiving citations at the same rate. Therefore, their CAC values will not significantly affect their ranking score whether it is high or low. Hence, these papers will not be underestimated just because they have become old. However, in the case that the papers obtained a negative CAC value as the case of papers 5 and 6, this mean that these papers are no longer receiving citations as in the past, thus the ranking score will be less than the PageRank-based score.

4. Experimental Results and Discussion

4.1. Dimension Dataset

We conducted the experiments on a real dataset of scientific papers taken from Dimensions database [30]. It provides an analytics API that supports the extraction of Dimensions data for use in complex analyses and visualizations. The API uses a query language called Dimensions Search Language (DSL) specifically developed for Dimensions data. So, data can be retrieved, aggregated, and sorted from highly specific requests in a single API call. Using Dimensions API, we got the required data based on the following set of conditions to ease the process of conducting the experiments: 1) the scientific papers must be related to one field; 2) the papers must be also published in a number of years (a long time period); and 3) the papers must have a considerable number of citations distributed among several years.

Based on these conditions, we extracted the relevant papers published in medical physics between 2005 and 2017, and the number of citations is tuned between 20 and 200 citations/paper. To facilitate the process of interpreting the results, all dataset's records (papers) are relevant in the sense that they are published in the same field, because the characteristics of citation differ from one field to another. The number of researchers and the number of research varies between fields. Also, these papers must be published over a long time period to be able to test whether the new method reduces bias or not. As for citations, it is necessary that the data set does not contain papers without citations or having very few numbers, because in this case they will take the same rank whether the original or

modified algorithm is used. Figure 2 shows the required query that meets these conditions.

Table 2 shows a sample of the returned data using the above query, for each

```
%dsldf search publications
in title_abstract_only for "Medical physics"
where year in [2005:2017]
and times_cited in [20:200]
return publications[id+doi+title+year+times_cited]
```

Figure 2. The DSL query to get a set of papers that meet the conditions.

Table 2. A sample of the collected data using dimensions API.

Index	Num. of citations	DOI	Year	ID	Title
0	23	10.1109/tkde.2017.2785824	2017	pub.1099918061	MCS-GPM: Multi-constrained simulation based graph pattern matching in contextual social graphs
1	55	10.1088/1361-6633/aa8b1d	2017	pub.1091615833	Review of medical radiography and tomography with proton beams
2	30	10.3762/bjoc.13.219	2017	pub.1092367839	Phosphonic acid: preparation and applications
3	30	10.1088/1361-6595/aa8d4c	2017	pub.1091850388	Foundations of low-temperature plasma physics—an introduction
4	102	10.3390/polym9100494	2017	pub.1092148337	Block copolymers: synthesis, self-assembly, and applications
5	23	10.1016/j.nima.2017.06.017	2017	pub.1090670633	Proton beam characterization in the experimental room of the Trento Proton Therapy facility
6	28	10.1098/rsfs.2016.0159	2017	pub.1091274054	Evolution viewed from physics, physiology and medicine
7	30	10.1002/acm2.12146	2017	pub.1091085605	AAPM-RSS medical physics practice guideline 9.a. for SRS-SBRT
8	31	10.1097/hp.0000000000000674	2017	pub.1090306251	Appropriate use of effective dose in radiation protection and risk assessment
9	59	10.1088/1742-6596/874/1/012029	2017	pub.1090837242	Horizon 2020 EuPRAXIA design study
10	67	10.1002/mp.12371	2017	pub.1085591560	Future of medical physics: real-time MRI-guided proton therapy
11	25	10.1186/s41747-017-0006-5	2017	pub.1085475539	Trends in radiology and experimental research
12	31	10.1115/1.4037671	2017	pub.1091274530	Applicability analysis of validation evidence for biomedical computational models
13	28	10.1002/acm2.12080	2017	pub.1085591656	AAPM medical physics practice guideline 8.a.: linear accelerator performance tests
14	43	10.4324/9781315268897	2017	pub.1090323410	A history of technoscience
15	26	10.1142/s0217732317400090	2017	pub.1085292639	Overview of the future upgrade of the INFN-LNS superconducting cyclotron

published paper it returns the paper's title, ID, DOI, publication date (year of publication), and the number of citations. The resulting dataset contains the first part of the necessary meta data to build a complete citation network.

To build the citation network, it is also necessary to collect all citing papers. This is because the citing papers must be a part of the citation network so that we can make links between it and its cited papers (the papers returned from the previous query). So, we used another query that takes the paper ID, and searches for it in the references of all papers published in the Dimensions database. If it finds the ID in the reference list of one of the papers, the paper is returned. This query was applied on all relevant papers that resulted from the first query to get all citing papers.

4.2. Calculating the PageRank and BTPR Scores

To calculate the original and modified PageRank scores, we first used an open-source network visualization and analysis software called Gephi [31] to create the citation network. As the PageRank algorithm is based on the linking structure of the papers. Next, we applied the original PageRank algorithm on the citation network to get the scores for the purpose of comparing the results. Referring to Equation (8), the CAC was calculated for each paper, as the required parameters were collected in the previous stage, including the paper publication date and the citation occurrence time. Then we calculated the BTPR scores using Equation (9) where we built python scripts to conduct these calculations.

5. Results Validation

The evaluation process for ranking algorithms faces many challenges that make it difficult and non-standardized such as the absence of a ground truth of the actual ranking [25], and the lack of recognition by the research community of comprehensive evaluation standards [12]. Moreover, each ranking algorithm is designed to achieve specific goals and satisfy the desires and requirements of specific users. Nevertheless, to evaluate the results and ensure the achievement of the study objectives, a set of measures was used to evaluate the performance of the modified algorithm. In addition to that a comparative analysis was conducted between the results of the original PageRank algorithm and the BTPR. To analyze the results and make comparisons, a list of top 100 papers ranked according to each algorithm was used.

5.1. Distribution of the Top 100 Papers by Publication Date

In order to identify the nature of the change in the results, and to ensure that the objective of the study is achieved by improving the ranking of some recent publications. The papers were sorted in descending order by using Microsoft Excel power query editor from highest rated to lowest rated based on the results of both the original PageRank algorithm and the modified one. A sample of 100 papers

in each year was considered to conduct further analysis and discussions. Figure 3 shows the number of papers published in each year that ranked among the top 100 using the PageRank algorithm. All these papers were published between years 2005 and 2017. The results show that among all papers published during the last three years, only 14 ranked among the top 100, and only 3 of them were published during the last two years (2016 and 2017). It is evident that the original PageRank is biased against recent publications and gives higher scores to the oldest ones.

Figure 4 shows the number of papers published each year that ranked among the top 100 using the BTBR algorithm, the results show an improvement in the scores of recent published papers, 26 paper instead of only 14 were ranked among the top 100, and 12 of them were published during the last two years (2016 and 2017). So, the bias against recent publications has diminished, and the rapidly growing papers are taking better scores. On the other hand, the old publications that have become less reliable, even though they have a large number of citations obtained in the past, will be taking fewer scores.

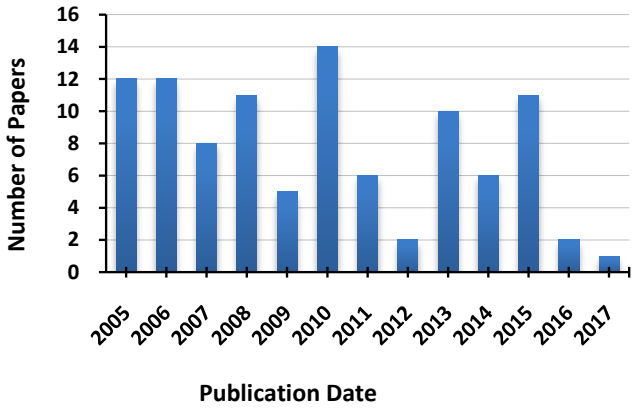

Figure 3. Distribution of the top 100 ranked papers using the PageRank algorithm by publication date.

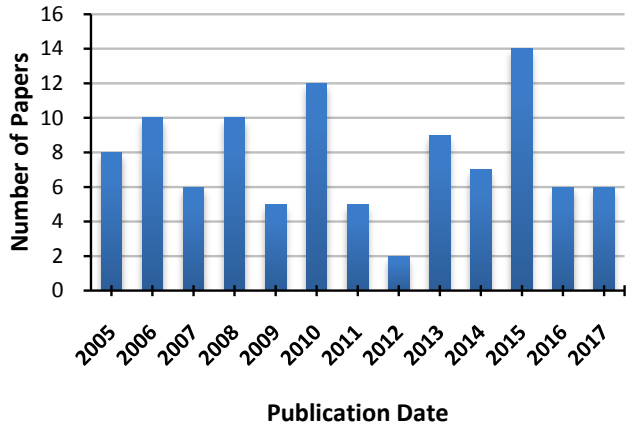

Figure 4. Distribution of the top 100 Ranked papers using the BTPR algorithm by publication date.

5.2. Fairness in Paper Ranking of Different Ages

To ensure that the modified version produces fair ranking scores for papers of all ages and does not imparting bias in favor of recent publications, we divided the changes into three cases, recent publications that received a better score, old publications that received same or better scores, and old publications that received lower scores. Then the citation behavior of these papers was tracked over time to compare it with the changes in ranking.

Figure 5 shows a sample of recent publications that received better scores. By looking at the citation behavior of these publications, we note that all of them share an ascending pattern of citation over the publication age. This indicates that they are in continuous growth, and dependence on them is also increasing. This explains the positive change in the ranking of these publications. Therefore, the modified algorithm is considered successful in ranking this group of publications.

Figure 6 shows a sample of old publications that received same or better scores. These papers are still valuable, and their citation average rate of change has not decreased. They are still maintaining the same growth rate; therefore, it is not fair to underestimate their value only because their publication date is old. So, the results of the new algorithm are very close to the results of the original algorithm regarding this case of papers. It had obtained good scores using the original algorithm, and the goal here is not to use time information in a way that underestimates their value unlike the other proposed solutions.

Figure 7 shows a sample of old publications that received lower scores. We can note that the reliance on them is constantly decreasing and has disappeared in some cases. Therefore, the new algorithm gives lower scores for these papers compared to the original algorithm scores.

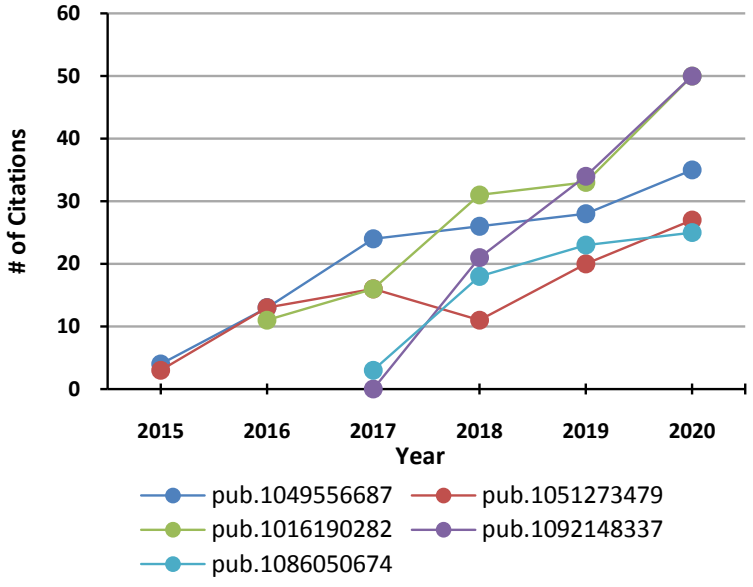

Figure 5. Citations by year for a sample of recent papers that received better scores.

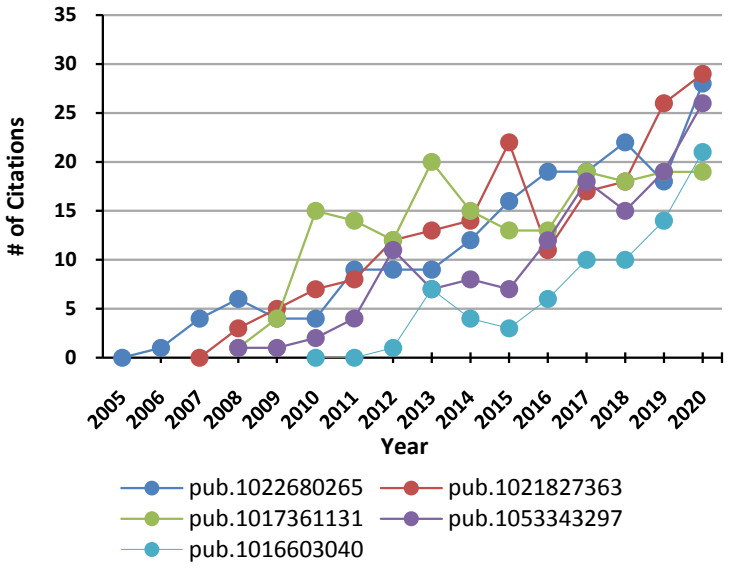

Figure 6. Citations by year for a sample of old publications that received same or better scores.

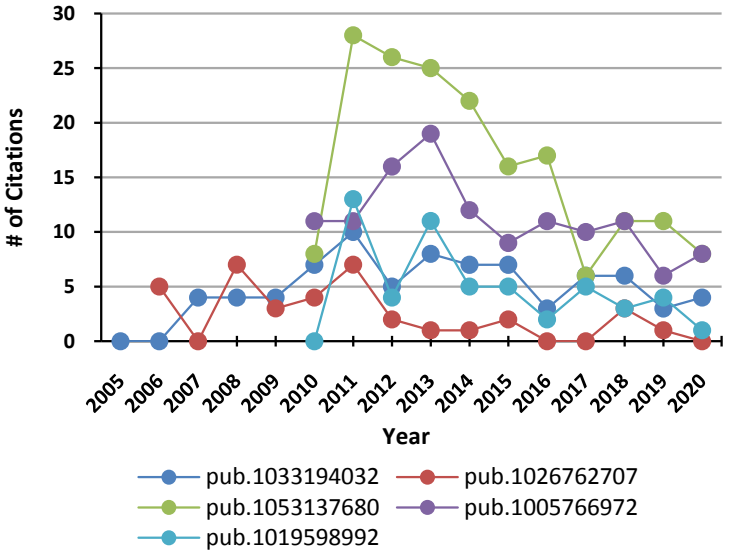

Figure 7. Citations by year for a sample of old publications that received lower scores.

5.3. Assessing the Similarity between the Original PageRank and BTPR

The change in results should be logical in which they don't differ radically, and don't cause large and illogical jumps in the publications' ranking. To achieve this, the similarity between the original PageRank algorithm and the BTPR is assessed by the Spearman's correlation coefficient (Spearman's ρ). It is calculated by Equation (10) [32].

$$R = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \tag{10}$$

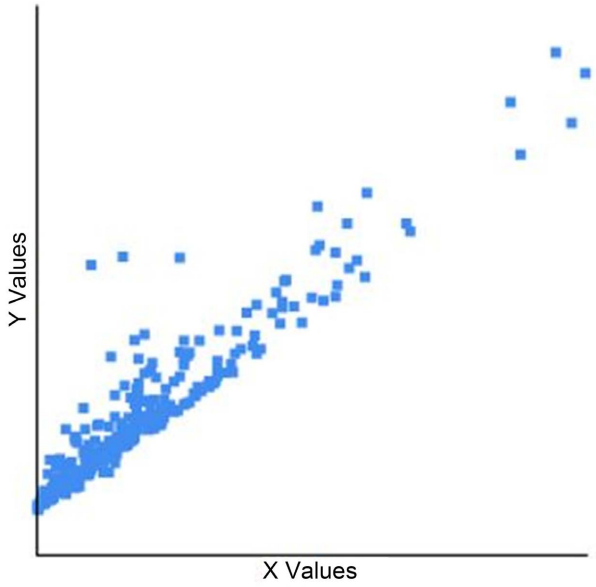

Figure 8. Positive correlation between the PageRank and the BTPR.

For this purpose, we constructed two lists of papers' rankings, one for the original PageRank algorithm, and another one for the new BTPR algorithm, where x_i is the rank position of paper i in the first list, y_i is the rank position of paper i in the second list, and x bar and y bar are the average ranking positions of all papers.

The value of R was calculated, and the result was 0.92. This is a strong positive correlation. It indicates that the changes are logical, and the results are reliable. The PageRank algorithm gives good results. Therefore, what is required is improvement on a certain part, without radical changes in the results. **Figure 8** shows this positive correlation.

6. Conclusions and Future Work

In this paper, we proposed a new modified version of the PageRank algorithm for ranking scientific publications by adding a new metric called the Citation Average rate of Change (CAC), where time information and citation data were used to calculate it. The aim was to reduce the bias in favor of old publications, which resulted from relying heavily on the citation count in the PageRank algorithm. The results showed that the proposed ranking method was time-aware considering the citation occurrence's time. As a result, recent publications that were still in the growth process but were continuously getting citations received better scores, even if they do not get enough time to collect a large number of citations. On the other hand, the results also showed that the old publications got fair scores when their ranking scores compared with their citation behaviors.

For future work, we will test the proposed algorithm on more datasets from other journals and databases, such as Scopus and Web of Science. Also, we plan to extend the new method to be able to rank a set of publications from different fields.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Joshi, M.A. (2014) Bibliometric Indicators for Evaluating the Quality of Scientific Publications. *The Journal of Contemporary Dental Practice*, **15**, 258-262. https://doi.org/10.5005/jp-journals-10024-1525

[2] Page, L., Brin, S., Motwani, R. and Winograd, T. (1998) The PageRank Citation Ranking: Bringing Order to the Web. Stanford InfoLab, California.

[3] Kleinberg, J.M. (1999) Hubs, Authorities, and Communities. *ACM Computing Surveys*, **31**, 5-es. https://doi.org/10.1145/345966.345982

[4] Kanellos, I., Vergoulis, T., Sacharidis, D., Dalamagas, T. and Vassiliou, Y. (2019) Impact-Based Ranking of Scientific Publications: A Survey and Experimental Evaluation. *IEEE Transactions on Knowledge and Data Engineering*, **33**, 1567-1584. https://doi.org/10.1109/TKDE.2019.2941206

[5] Liu, X., Bollen, J., Nelson, M.L. and Van de Sompel, H. (2005) Co-Authorship Networks in the Digital Library Research Community. *Information Processing & Management*, **41**, 1462-1480. https://doi.org/10.1016/j.ipm.2005.03.012

[6] Bollen, J., Rodriquez, M.A. and Van de Sompel, H. (2006) Journal Status. *Scientometrics*, **69**, 669-687. https://doi.org/10.1007/s11192-006-0176-z

[7] Chen, P., Xie, H., Maslov, S. and Redner, S. (2007) Finding Scientific Gems with Google's PageRank Algorithm. *Journal of Informetrics*, **1**, 8-15. https://doi.org/10.1016/j.joi.2006.06.001

[8] Yao, L., Wei, T., Zeng, A., Fan, Y. and Di, Z. (2014) Ranking Scientific Publications: The Effect of Nonlinearity. *Scientific Reports*, **4**, Article No. 6663. https://doi.org/10.1038/srep06663

[9] Yan, E. and Ding, Y. (2010) Weighted Citation: An Indicator of an Article's Prestige. *Journal of the Association for Information Science and Technology*, **61**, 1635-1643. https://doi.org/10.1002/asi.21349

[10] Zhang, F. and Wu, S. (2018) Ranking Scientific Papers and Venues in Heterogeneous Academic Networks by Mutual Reinforcement. *Proceedings of the* 18*th ACM*/*IEEE on Joint Conference on Digital Libraries*, Fort Worth, 3-7 June 2018, 127-130. https://doi.org/10.1145/3197026.3197070

[11] Ghosh, R., Kuo, T.-T., Hsu, C.-N., Lin, S.-D. and Lerman, K. (2011) Time-Aware Ranking in Dynamic Citation Networks. 2011 *IEEE* 11*th International Conference on Data Mining Workshops*, Vancouver, 11 December 2011, 373-380. https://doi.org/10.1109/ICDMW.2011.183

[12] Dunaiski, M. and Visser, W. (2012) Comparing Paper Ranking Algorithms. *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference*, Pretoria, 1-3 October 2012, 21-30. https://doi.org/10.1145/2389836.2389840

[13] Giuffrida, C., Abramo, G. and D'Angelo, C.A. (2019) Are All Citations Worth the Same? Valuing Citations by the Value of the Citing Items. *Journal of Informetrics*, **13**, 500-514. https://doi.org/10.1016/j.joi.2019.02.008

[14] Hsu, C.-C., Chan, K.-H., Feng, M.-H., Wu, Y.-H., Chen, H.-Y., Yu, S.-H., *et al.* (2016) Time-Aware Weighted Page Rank for Paper Ranking in Academic Graphs. *Proceedings of the* 9*th ACM International Conference on Web Search and Data Mining*,

WSDM'16, San Francisco, 22-25 February 2016, 1-4.

[15] Ma, S., Gong, C., Hu, R., Luo, D., Hu, C. and Huai, J. (2018) Query Independent Scholarly Article Ranking. 2018 *IEEE* 34*th International Conference on Data Engineering* (*ICDE*), Paris, 16-19 April 2018, 953-964. https://doi.org/10.1109/ICDE.2018.00090

[16] Wei, Y., Yi, F., Cui, X. and Chen, F. (2021) An Improved Page Rank Algorithm Based on Text Similarity Approach for Critical Standards Identification in Complex Standard Citation Networks. *Complexity*, **2021**, Article ID: 8825947. https://doi.org/10.1155/2021/8825947

[17] Walker, D., Xie, H., Yan, K.-K. and Maslov, S. (2007) Ranking Scientific Publications Using a Model of Network Traffic. *Journal of Statistical Mechanics*: *Theory and Experiment*, **2007**, Article ID: P06010. https://doi.org/10.1088/1742-5468/2007/06/P06010

[18] Sayyadi, H. and Getoor, L. (2009) Futurerank: Ranking Scientific Articles by Predicting Their Future Page Rank. *Proceedings of the* 2009 *SIAM International Conference on Data Mining*, Sparks, 30 April-2 May 2009, 533-544. https://doi.org/10.1137/1.9781611972795.46

[19] Kanellos, I., Vergoulis, T., Sacharidis, D., Dalamagas, T. and Vassiliou, Y. (2021) Ranking Papers by Their Short-Term Scientific Impact. 2021 *IEEE* 37*th International Conference on Data Engineering* (*ICDE*), Chania, 19-22 April 2021, 1997-2002. https://doi.org/10.1109/ICDE51399.2021.00190

[20] Hirsch, J.E. (2005) An Index to Quantify an Individual's Scientific Research Output. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 16569-16572. https://doi.org/10.1073/pnas.0507655102

[21] Garfield, E. (2006) The History and Meaning of the Journal Impact Factor. *The Journal of the American Medical Association*, **295**, 90-93. https://doi.org/10.1001/jama.295.1.90

[22] Hwang, W.-S., Chae, S.-M., Kim, S.-W. and Woo, G. (2010) Yet Another Paper Ranking Algorithm Advocating Recent Publications. *Proceedings of the* 19*th International Conference on World Wide Web*, Raleigh, 26-30 April 2010, 1117-1118. https://doi.org/10.1145/1772690.1772832

[23] Kleinberg, J.M. (1999) Authoritative Sources in a Hyperlinked Environment. *Journal of ACM*, **46**, 604-632. https://doi.org/10.1145/324133.324140

[24] Du, M., Bai, F. and Liu, Y. (2009) PaperRank: A Ranking Model for Scientific Publication. 2009 *WRI World Congress on Computer Science and Information Engineering*, Los Angeles, 31 March-2 April 2009, 277-281. https://doi.org/10.1109/CSIE.2009.479

[25] Wang, Y., Tong, Y. and Zeng. M. (2013) Ranking Scientific Articles by Exploiting Citations, Authors, Journals, and Time Information. *Proceedings of the* 27*th AAAI Conference on Artificial Intelligence*, Bellevue, 14-18 July 2013, 933-939.

[26] Lu, Y., Ma, K. and Duan, J. (2021) Influence Model of Paper Citation Networks with Integrated PageRank and HITS. 2021 *IEEE* 24*th International Conference on Computer Supported Cooperative Work in Design (CSCWD)*, Dalian, 5-7 May 2021, 1081-1086. https://doi.org/10.1109/CSCWD49262.2021.9437678

[27] Nie, Z., Zhang, Y., Wen, J.-R. and Ma, W.-Y. (2005) Object-Level Ranking: Bringing Order to Web Objects. *Proceedings of the 14th International Conference on World Wide Web*, Chiba, 10-14 May 2005, 567-574. https://doi.org/10.1145/1060745.1060828

[28] Bai, X., Zhang, F., Ni, J., Shi, L. and Lee, I. (2020) Measure the Impact of Institution

and Paper via Institution-Citation Network. *IEEE Access*, **8**, 17548-17555. https://doi.org/10.1109/ACCESS.2020.2968459

[29] Adams, R.A. and Essex, C. (1995) Calculus: A Complete Cours. 3rd Edition, Addison-Wesley, Boston.

[30] Digital Science (2018) Dimensions [Software]. https://app.dimensions.ai

[31] Bastian, M., Heymann, S. and Jacomy, M. (2009) Gephi: An Open Source Software for Exploring and Manipulating Networks. *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 3, San Jose, 17-20 May 2009, 361-362.

[32] Myers, J.L., Well, A.D. and Lorch, R.F.J. (2013) Research Design and Statistical Analysis. Routledge, New York. https://doi.org/10.4324/9780203726631