

An Efficient Ranking Algorithm for Scientific Research Papers

Yaser Al-Lahham, Fathi Al-Hatta, Mohammad Hassan

Zarqa University, Jordan

yasirlhm@zu.edu.jo, fathihattab@zu.edu.jo, mzaita@zu.edu.jo

Abstract

Large content of scientific research papers of variable quality became available online, that is generally hard for researchers and students to locate their desired information by using the traditional search engines that return large number of documents. This made it necessary to develop an efficient and effective ranking algorithm to solve this problem. One of the most famous ranking algorithms is the PageRank (PR), which is originally used to rank web pages. This algorithm needs to be modified in order to efficiently rank scientific research papers, since they are different than WebPages. This paper proposes adding some factors that can enhance PageRank results when applied on research papers, making it less biased toward old papers, and taking author ranks into account. The results showed that the proposed method improved the results of the PageRank with respect to the recall and precision measures.

Keywords: Scientific Research Papers' Ranking, PageRank, Search Engines

1. Introduction

One of the reliable resources for information on the web is the digital academic libraries. They considered being one of the most important sources that provide Scientific Research Papers (SRP) for students and researchers. Considerable numbers of universities and public libraries have provided access to books, journals and other documents. This collection helps academic researchers to acquaint with new journal articles and conferences proceedings that relate to their areas of research.

In order to make this scientific content to be effectively used, researchers need to rank the results of their search. Traditional search engines are widely used to search and rank the returned answers. These search engines were basically designed to search and rank web pages of different types and sources. The nature of SRPs differs from the usual web search; since ranking scientific research papers depends on other factors than citation analysis for example. However, measuring SRPs quality by citations is surrounded by a number of different viewpoints and opinions. Some researchers measure SPR quality based on how

often it has been cited on other SRPs, and the date of publication. Others argue that citation counts are an indicator that best assesses a publication's impact rather than its quality or importance [11]. Citation count is a partial indicator of impact, and that other factors such as communication practices and author visibility have to be significantly assumed [18].

One of the most famous ranking algorithms is PageRank [1] algorithm which is adopted by Google search engine, and based on using citation counts as the highest weighed factor in Google scholar engine.[6] Citation count could be used for ranking scientific research papers, but not as the only factor, moreover, other factors can be added to better determine the importance of SRP.

The main objective of this paper is to efficiently rank scientific research papers (SRPs) using the PageRank itself assisted by other factors; such as author's score and the inverse of the age of the paper, in order to make it less biased against new papers.

The remainder of the paper is organized as follows: section 2 surveys the related work. Section 3 presents the proposed ranking method of research paper ranking, section 4 explains the empirical evaluation of the proposed method, and finally section 5 concludes the work, and presents future improvement of this work.

2. Related Work

Relevance Algorithms (either similarity based or probabilistic models) rank documents according to relevance to a query. With query q as input, documents are ranked based on the query likelihood, or the probability that the document's language model will generate the terms in the query.

Query-independent ranking algorithms are connectivity-based (or link based), it ranks a list of documents according to their own importance based on link analysis technique ([2] and [3]). They view the web as a graph where the web pages form the nodes and the hyperlinks between the web pages form the edges between these nodes [13]. Examples of such algorithms: HITS [4] where a web page serves as information provider, and gateway to other pages that gives information on a topic. Another important example is the PageRank[1] (PR) a mathematical algorithm based on the web graph, the rank value indicates importance of a particular page [5]. A hyperlink to a page counts as a vote of support. PageRank favors older pages, because a new page, even a very good one, will not have many links unless it is part of an existing site.

Most of proposed SRP ranking algorithms are variants of the PageRank and HITS algorithms. Factor other than citations were included, such as author score as in [7] and [16]. Time (or date of publication) is also considered as a new factor to extend the PageRank algorithms, for example [11], [15], [16], and [17]. Relationship between citation and keywords is also considered by [9], the effect of topic on citation is studied by [12] and [8].

PageRank was extended using other factors; such as citation relativity, for example, relativity measurements instead of the simple direct citation between papers [9], and mutual reinforcement relationships [10], direct citation enforcement factor [25].

Other proposals focused on estimating the rank by utilize a random walk model to predict the number of future citations for each article, such as [15] and [16], the main idea of these proposals is that If an author publishing many prestigious papers previously then the new publications of him/her can be expected to have good quality.

On the other hand, some research efforts focus on solving problems of existing algorithms, such as Krapivin, M. et al. [13] introduced Focused Page Rank (FPR) to reduce the "effect of outbound links",

which means that if paper P is cited many times by papers with high rank but containing a large quantity of outgoing links, it may decrease P's rank. Other problem investigated by Chen, P. et al. [14] where some classical articles have small number of citations and very high PageRank, these articles were named "*scientific gems*", which are given rank not only according the total citation count, but also the rank of each citing paper.

Related work investigation makes it is clear that PageRank algorithm was the favorite ranking model, and it was the basis of most of SRP ranking algorithms that other factors were added. Many studies used time as an important factor to give the newer papers a higher score.

This paper proposes an algorithm that ranks SRPs based on PageRank score, date of publication, and author score.

3. Scientific Research Paper Algorithm

The proposed algorithm of scientific research papers' ranking is an extension of the PageRank algorithm. The total score of a paper is directly proportional to the PageRank score and the author score, but it is conversely proportional to paper age. This section gives a justification of each selected factor:

Author Score: number of papers published by an author, and citation to his/her papers may reflect the productivity and popularity of the author. Different metrics are used by research community to measure author rank; such as the h-index. In this paper author score is calculated using the following equation:

$$AU = \frac{[(AW_1 + \dots + AW_m) + (AC_1 + \dots + AC_m)]}{N.H} \quad (3.1)$$

Where AW_i is the number of papers published by the i^{th} author, AC_i is the citation count for the i^{th} author, N is the number of all authors for the current paper, and H is a constant ≥ 1 .

The constant H is used as a balancing factor in order to reduce the impact of the author score on the final ranking score in case it seems to dominate its value. H was empirically tested, and found that the optimal value is 10, since it gave the most balanced result compared to other tested values such as 0.5, 20 and 50.

This factor gives higher weight to papers authored by experienced authors than papers published by students, for example. Total number of citations is given more importance since experienced authors will produce better papers so will have more citations. Other researchers include other factors such as the impact factor of the publication; the journal or the conference the paper is published in [11].

Date of Publication: One of the PageRank shortcomings is that it favors old papers over newer ones [13], even if the newer one is better, that is because as longer the paper as more citations it has. Recent papers have little citations so that they should be given some promotion in the ranking process ([11] and [20]). To overcome this biasness, and making the ranking more reliable, the proposed method used the inverse of the paper age as a factor; in fact the log of the age is used instead in order make little degradation of the score of a paper, so as not to affect popular papers unless its age gets as twice as its current age. Age of a paper is the number of years since it has been published.

$$A = Y - Y_i \quad (3.2)$$

Where Y is the current year, and Y_i is the year of publication of paper i .

Combining equations (3.1) and (3.2), the final ranking score of a research paper is calculated as given in equation (3.3).

$$SRP_Score = \frac{PR}{(1 + \log_2(A))} + AU \quad (3.3)$$

Where PR is the PageRank score, A is the Age of the SRP, and AU is the author score. The pseudo code of proposed ranking algorithms is as follows:

Algorithm: Scientific Research Ranking R

Required:

T_i = paper Title. , Ac = Authors Citation Count.
 Aw = Authors Number of Works. , N = Number of Authors for each Paper.
 Di = Date of Publication. , PR = PageRank Score. , D = Current Year.

```

1: For each paper in dataset do
2: Initialize PR, Ac, Aw, Di, PR to 0.0;
3: Get : N[current paper], Ac[current paper], Aw[current paper]
      Di [current paper], PR[current paper]
4: compute AU= (Ac+Aw)/(N.H) // H is 10 as explained in sec 3
5: compute A=Current_Year-Di
6: compute Scientific Research Ranking R= (PR/(1+log (A))+AU)
7: end

```

In fact the age of publication is used to *advocate* newest papers, and to give high quality papers the opportunity to gain high ranks once they published. The age of the paper is highlighted by many researchers; such as in ([11], [19] and [20]).

4. Implementation and Evaluation

This section explains how good is the proposed SRP-Rank algorithm to give better papers (authored by experienced authors) higher ranks. The implementation procedure involves the following steps:

- Dataset preparation and processing to extract important information.
- Constructing the citations' graphs.
- Presenting the proposed ranking method.
- Compare the results to that of the PageRank algorithm applied on the same data set.

The comparison criteria includes: distribution of Ranked SRPs among the age of the papers, and among number of citations recall and precision.

Table 4.1: Distribution of paper publication dates

Year	2013	2012	2011	2010	2009	2008	2007	2006	2005	2004
# of Papers	19	31	32	34	28	26	28	19	14	12
Year	2003	2002	2001	2000	1999	1998	1997	1996	1995	1994
# of Papers	9	8	8	9	29	15	7	10	14	17
Year	1993	1992	1991	1990	1989	1988	1987	1986	1985	1984
# of Papers	13	13	17	15	13	17	13	20	19	26

Year	1983	1982	1981	1980	1979	1978	1977	1976	1975	1974
# of Papers	17	15	18	16	23	23	28	28	25	19
Year	1973	1972	1971	1970	1969	1968	1967	1966	1965	1964
# of Papers	25	26	27	28	20	21	27	25	30	26
Year	1963	1962	1961	1960	1959	1958	1957	1956		
# of Papers	19	33	23	28	25	24	17	21		

4.1 Dataset

To evaluate the proposed algorithm, dataset containing scientific research papers metadata such as title and authors is required. The data set that used in this study is obtained from the Web of Science, it contains for each paper: the abstract, and basic metadata. It also includes 9583 citations for 1,189 SRPs in the history and physiology of the Science field, and covers publications from 1956 to 2013. The dataset contains additional information such as page number and authors contact address, the distribution of paper publication over publication year is shown in table 4.1.

```

52469 "Vallicrosa M, 1942, Traducciones Orienta, P152" localcitationcount 1
52470 "Bjornbo Aa, 1912, Bibliotheca Math, V12, P220" localcitationcount 1
52471 "Liber Centiloqui Pth" localcitationcount 1
52472 "Dinegro A, Hist Magic Expt Sci, V3, P404" localcitationcount 1
52473 "Dinegro A, Infusione Spermatis" localcitationcount 1
52474 "Valentinelli, Bibliotheca Manuscri, V4, P280" localcitationcount 1
52475 "1899, Pubblicazione R Osser, V40" localcitationcount 1
52476 "1907, Bibliofilia, V8, P372" localcitationcount 1
52477 "Thorndike, Catalogue Incipits" localcitationcount 1
52478 "Pruckner H, 1933, Studien Astrologisch, P8" localcitationcount 1
52479 "Defundis J, Hist Magic, V4, P232" localcitationcount 1
*Arcs
127 135
21 135
74 135
38 135
23 135
65 135

```

Figure 3.4: Sample of the Paper Citation Network

4.2 Data Preparation and Extracting

To extract the data that will be used in ranking score calculation that data need to be pre-processed. It is first arranged using Sci2 tool, which automatically saves the dataset as excel sheet.

The unwanted and missing fields were removed, and only the data needed by the proposed algorithm is kept. The final database contains information about authors, title, year of publication, citation count, and bibliography.

Extracting Citations' Networks: the data that used by the proposed ranking algorithm is stores as:

- Paper citation network (graph): Paper citation network that required for PR calculation, is obtained . Sample of paper citation network is shown in figure 3.4.

	A	B	C	D
1	label	timesCited	numberOfWorks	uniqueIndex
2	Abbud, F	7	1	1857
3	Abraham, Ga	13	1	1020
4	Acker, Cj	6	1	666
5	Adkins, Gm	1	1	288
6	Africa, Tw	6	1	1915
7	Agassi, J	2	1	1892
8	Aiton, Ej	17	1	1561
9	Alberti, Fb	1	1	241
10	Albury, Wr	3	1	1490
11	Alder, K	5	3	27
12	Alexander, A	0	1	124

Figure 3.5: Sample of the information extracted from author paper graph.

- Authors' Paper Network (graph): Author citation graph provides information about number of works and citation for each author. Sample of the information extracted from author paper graph is shown in figure 3.5.

4.3 Evaluation

To illustrate the results, a query is used to narrow the list of papers, since it contains articles in different subjects that are not all related to each other, the retrieved list contained 41 articles. Both of the proposed ranking algorithm and PR were used to rank the list to evaluate the results. The results are presented using distributions as follows:

The distribution of the Ranked SRPs over the Age of the Paper: To evaluate whether the proposed ranking method is less biased against new papers, and to examine the effect of the age on the distribution of the top ranked papers, charts for all the 41 ranked list were used. Figures 4.7 and 4.8 show the distribution of the proposed rank method and distribution of PR among the age of the paper.

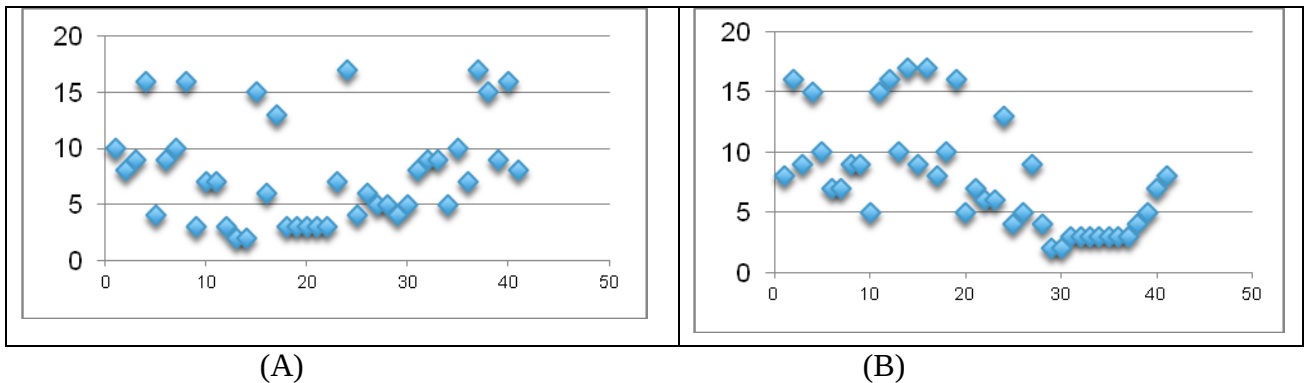


Figure 4.7: Distribution of ranks over the Age of the Paper, y-axis presents paper age, and x-axis presents ranks in both: (A) the Proposed Rank method (B) the PageRank

In figure 4.7 (B), the distribution of ranked SRPs using PR among the age of the paper showed that the PageRank algorithm is biased against new papers, since it gives papers that has ages less than 5 a rank more than 25, as shown in figure 4.7 (B), while the distribution of the ranked SRPs using the proposed

method shows that the these papers got a rank starting from 5 and more, consequently; the proposed method is less biased against new papers, as shown in figure 4.7 (A).

The Distribution of Ranked SRPs over the Citation number: To evaluate the citation effect on the proposed method, charts for ranked list are used. Figures 4.8 (A) and (B), show the distribution of the proposed rank method and distribution of PR among the citation count.

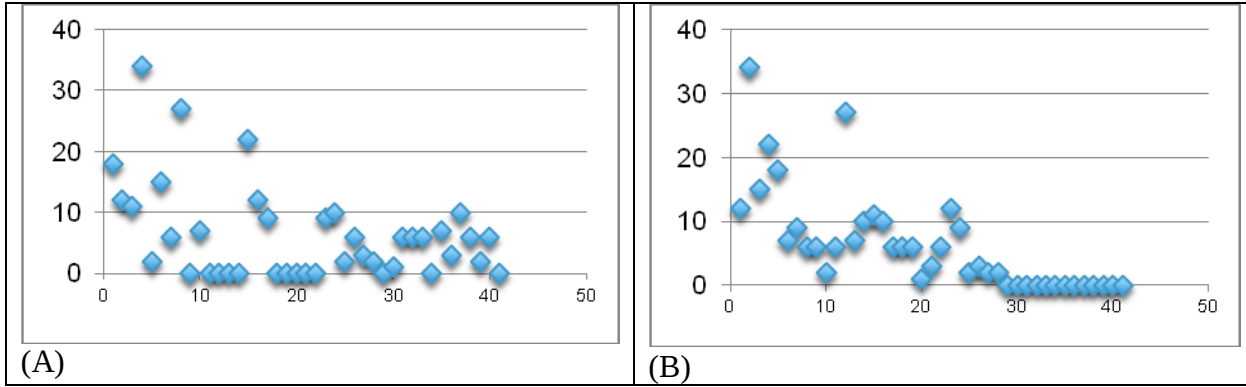


Figure 4.8: Distribution of the citation count, x-axis the rank, and y-axis represents the number of citations, of both: (A) the proposed ranking method, (B) The PageRank.

Figure 4.8 (B) shows that the PageRank algorithm is heavily based on citations count, since most of papers that have high citation given ranks less than 10. In figure 4.8 (A), the distribution of the ranked SRPs using proposed method among the citation count shows that the proposed method is less depending on citation count unlike PR, since many papers with low citations gain ranks less than 10; these papers could be written by expert authors for example.

Recall and Precision: These measures are used in IR with binary classification (e.g. similar/non-similar) to measure the relevance of a set of retrieved items and evaluate the performance of information retrieval systems. A modified version of these measures was presented by [21] to evaluate web service ranking methods. According to [21] recall is calculated by dividing the highest rank score by total rank score of all paper in, as shown in equation (4.6).

$$Precision = \frac{\text{Highest rank score}}{\text{Total rank score of all papers}} \quad (4.6)$$

Recall is found by dividing the highest rank score by the score of 2nd highest algorithm, as in equation (4.7)

$$Recall = \frac{\text{Highest rank score}}{\text{score of 2nd highest paper}} \quad (4.7)$$

The analysis and comparison of the proposed algorithm and PR based on this evaluation metric is shown in table 3.2. This metric also gives more evidence that the proposed method offers a better solution to ranking SRPs when compared with the basic PageRank algorithm.

The recall value of the proposed SRP ranking method shows that this method gave a higher rank for the top ranked papers. This result argues that the proposed method gives more distinction to the top ranked paper.

Table 3.2: Results of Precision and Recall in both page rank and proposed SRP rank algorithms

	Precision	Recall
The PageRank	0.025813562	1.018441539
The proposed SRP-Rank	0.082599556	1.266729551

The precision value achieved by the proposed method is higher than the PR, means that the distribution of ranks is biased towards higher ranked paper. Consequently, this result makes it possible to distinguish interesting papers in certain topics to the ranked paper faster than PR.

5. Conclusion and Future Work

In this paper, scientific research papers ranking algorithm was proposed to balance the impacts of the PR score on old papers and new papers, and to involve the authors' impact on the papers they wrote. The results show that the proposed ranking algorithm doesn't rely heavily on citation count as PR do. The proposed method got higher precision and recall than PageRank, which makes it clearly distinct highly ranked papers,

In terms of future work, there are several directions that can be explored, another interesting direction might be examined to conduct more types of evaluation of the results including query evaluation.

REFERENCES

- [1] C. Ridings and M. Shishigin (2002), **Pagerank Uncovered**, Technical report.
- [2] Liu T.Y. (2011), **Learning to Rank for Information Retrieval**. Berlin. Heidelberg: Springer.
- [3] Devi P, Gupta A, and Dixit A, (2014). **Comparative Study of HITS and PageRank Link Based Ranking Algorithms**. International Journal of Advanced Research in Computer and Communication Engineering, Vol 3, Issue 2, PP. 5749 - 5754.
- [4] Kleinberg J, (1999). **Authoritative Sources in a Hyperlinked environment**. Journal of the ACM, Vol. 46, No. 5, pp. 604-632.
- [5] Brin S, and Page L, (1998) **The Anatomy of a Large-Scale Hypertextual Web Search Engine**. Computer Networks, Vol. 30, pp. 107-117.
- [6] Beel, J. and Gipp, B. (2009) **Google Scholar's Ranking Algorithm: An Introductory Overview**, 12th International Conference on Scientometrics and Informetrics, Vol. 1, PP. 230-241.
- [7] Anh, V.L. Hoang, H.V. Trung, H.L. Trung, K.L. and Jung, J.J (2014) **Evaluating Scientific Publications By N-Linear Ranking Model**, ANNALES Universitatis Scientiarum, Sectio Computatorica, Vol. 43, PP. 123-147.
- [8] Sohn, B.S. and Jung, J. (2015) **A Novel Ranking Model for a Large-Scale Scientific Publication**. Mobile Networks and Applications, Vol. 20, Issue 4, PP 508-520.
- [9] Due M, Bai F, and Liu Y, (2009). **PaperRank: A Ranking Model for Scientific Publication**. IEEE World Congress on Computer Science and Information Engineering, Vol 4, PP. 277- 281.

- [10] Jiang X, Sun X, Zhuge H (2012) **Towards an Effective and Unbiased Ranking of Scientific Literature through Mutual Reinforcements**. 21st ACM Conference on Information and Knowledge Management, Hawaii, USA, pp 714–723.
- [11] Wang, Y. Tong, Y. and Zeng, M (2013) **Ranking Scientific Articles by Exploiting Citations, Authors, Journals, and Time Information**. Proceedings of the Twenty-Seventh AAAI Conference on Artificial Intelligence, Bellevue, Washington, USA, PP. 933 - 939.
- [12] Shubhankar, K. Singh, A. and Pude, V. (2011), **An Efficient Algorithm for Topic Ranking and Modeling Topic Evolution**. Database and Expert Systems Applications, Vol 6860, PP 320-330.
- [13] Krapivin, M and Marchese, M. (2008), **Focused Page Rank in Scientific Papers Ranking**. Digital Libraries: Universal and Ubiquitous Access to Information, Vol 5362, PP 144-153.
- [14] Chen, P. Xie, H. Maslov, S. and Redner, S. (2007) **Finding Scientific Gems with Google's PageRank Algorithm**. Elsevier, Journal of Informatics, Vol. 1, PP. 8 -15.
- [15] Walker D, Xie H, Yan KK, and Maslov S, (2007) **Ranking Scientific Publications Using a Model of Network Traffic**. Journal of Statistical Mechanics: Theory and Experiment.
- [16] Sayyadi, H., and Getoor, L. (2009) **Futurerank: Ranking Scientific Articles by Predicting Their Future PageRank**. Ninth SIAM International Conference on Data Mining, PP. 533–544.
- [17] Singh, A. Shubhankar, K. and Pudi, V. (2011) **An Efficient Algorithm for Ranking Research Papers Based on Citation Network**. 3rd Conference on Data Mining and Optimization, Putrajaya, PP. 88 -95.
- [18] Dunaiski D, and Visser W, (2012). **Comparing Paper Ranking Algorithms**. Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference, Pretoria, South Africa, PP. 21-30.
- [19] Hwang, W.S, Chae, S.M, and Kim, W.K. (2010), **Yet Another Paper Ranking Algorithm Advocating Recent Publications**. 19th International Conference on World Wide Web, Raleigh, North Carolina, USA.
- [20] Mushtaq, H. (2014) **Scientific Research Paper Ranking Algorithm PTR: A Tradeoff between Time and Citation Network**. Applied Mechanics and Materials. Vol 551, PP. 603-611.
- [21] Manoharan, R. Archana, A. and Cowlagi, S. N. (2011) **Hybrid Web Services Ranking Algorithm**. IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 3, No. 2, PP 452 – 460.