2022 3rd Int. Conf. on Innovations in Science, Engineering and Technology (ICISET)
26-27 February 2022, Chittagong, Bangladesh

2022 International Conference on Innovations in Science, Engineering and Technology (ICISET) | 978-1-6654-8397-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICISET54810.2022.9775821

A Study on Paper and Author Ranking

Palash Ranjan Roy
School of Data and Sciences
Brac University
Dhaka, Bangladesh
palash.ranjan.roy@g.bracu.ac.bd

Md. Noushin Islam
School of Data and Sciences
Brac University
Dhaka, Bangladesh
md.noushin.islam@g.bracu.ac.bd

Labiba Tasfiya Jeba
School of Data and Sciences
Brac University
Dhaka, Bangladesh
labiba.tasfiya.jeba@g.bracu.ac.bd

Md. Adnanul Haq
School of Data and Sciences
Brac University
Dhaka, Bangladesh
md.adnanul.haq@g.bracu.ac.bd

Iffat Afsara Prome
School of Data and Sciences
Brac University
Dhaka, Bangladesh
iffat.afsara.prome@g.bracu.ac.bd

Mohammad Kaykobad
School of Data and Sciences
Brac University
Dhaka, Bangladesh
kaykobad@bracu.ac.bd

Tanvir Kaykobad
School of Computing
Queen's University
ON, Canada
m.kaykobad@queensu.ca

***Abstract*—As the number of journal issues, conferences and the overall scientific literature have been increasing at an exponential rate, it has become challenging for researchers to find appropriate and useful papers from the vast literature available to them. To solve this issue citation count, h-index, i10-index are used to rank authors. In 1998, Brin and Page introduced the algorithm PageRank which is also used in the scientific community for ranking research papers and authors. However, each of these metrics has its own drawbacks. We hypothesize that papers unveiling deeper truth are often not as well cited as those that are more challenging for a wider number of authors to assimilate and appreciate their works. So a simple count of the number of citations may fail to capture the essence of the quality of a paper. With a view to addressing this issue, we have introduced a new algorithm that also takes into account the quality of the researcher citing an article, and considers it in ranking. We have carried out experiments. While the experiments are not as comprehensive, results have been incorporated. They look promising in ranking authors and papers that are not cited too often due to difficulty in understanding them.**

***Index Terms*—Pagerank, Author Rank, Author Paper Rank, Paper Rank, Citation, Citation Network**

I. Introduction

With the advent of information technology there has been an exponential outburst of research papers and web pages in the internet. This has made it challenging to access more relevant information from the internet manually. That is why Page ranking or author ranking has been a subject of many research works now a days. In paper [1], the authors outline Page or author ranking particularly for promotion of products and services in business.

Dorogovtsev and Mendes [2] presented one of the most popular yet simplistic methods of ranking scientists. It only requires two comparable factors: the number of total citations of the page and its rank. Hirsch [3] describes a new bibliometric indicator $\overline{h}$, which discourages honorary authorship by giving more credit to authors who publish alone or in small partnerships, while deducting credit from coauthors in bigger collaborations. There are other papers like those of Dorogovtsev and Mendes [4], Lü et al. [5], Pacheco et al, [6], Zhang er al [7], Oberesch and Groppe [8], Amjad, Daud and Aljohani [9], Zhao et. al. [11], Amzad et al [12], Kosmulski [13], Daud et al [14] that consider special cases, different variations and brought in improvement of these algorithms.

In this paper we are particularly interested in ranking scholarly research papers and researchers to help them locate the right resources for them. We would like to rank both papers and authors. The website Google Scholar maintains citation counts of each author together with the list of coauthors and number of papers having good citations. Number of citations vary significantly with papers in areas like economics and environmental sciences attracting lot of citations whereas theoretical science areas like physics and mathematics may not be as lucky. Some authors are highly meritorious, but they may not be publishing too often. Their contributions are too deep and difficult for common researchers to understand, and therefore, they fail to attract the citation count they actually deserve. The only female Fields medalist Maryam Mirzakhani has been cited only about 1900 times after 4 years of her death and 5 years after receiving Fields medal. This is why simple citation counts are not good enough to properly appreciate works of significant depth.

In this paper, we present a new algorithm to better rank both

papers of deep consequences and the researchers generating them, whose body of work may be small in quantity but are of great significance in terms of quality and impact.

Brin and Page [15] give an extensive rundown of a large-scale web search engine, where they look at the question of how to process unregulated hypertext (containing texts of other pages) collections effectively, where everyone can publish what they want. Brin and Page [15] developed the pagerank algorithm for ranking webpages based upon the number of webpages having link to this particular webpage. pagerank algorithm is based on the philosophy that a random web surfer browses different websites with the links available in the web page it is browsing. At some point of time she stops browsing. pagerank represents this behavior by a model that determines the nature of the user who maintains randomly clicking on successive connections. Nevertheless, sometimes the surfer gets bored and hops to a chosen random page based on the categorization. So, the surfer is not going to stay in an infinite loop. Primarily pagerank works by measuring the number and quality of links to a page to calculate a rough estimation of how resonating the website is. The authors have also presented some demonstrations of how pagerank can be computed effectively for the high magnitude of pages. They have also introduced a dampening factor so that a highly ranked webpage, distant apart in terms of links, does not induce excessive credits to a paper only because there is a path of links between these two papers. Brin and Page considered random surfer model that ultimately keeps provision of a surfer stopping surfing after being sufficiently tired. They have also taken into account that certain web pages may not have any inlinks at all. The inlink shows the connection if a paper cites another paper.

Brin and Page introduced the following equations for computing page ranks.

$$\mathrm{PR}(i) = \sum_{j \in B_i} \frac{\mathrm{PR}(j)}{L(j)}, \quad \forall i \in U \tag{1}$$

$\mathrm{PR}(i)$ is the page rank of page i, B_i is the set of papers having outlink to paper i and $L(j)$ is the set of links coming out of paper j, and U is the set of all papers. Page rank theory holds that a fictitious surfer who randomly clicks on the links, will eventually stop clicking. The probability that at any step the surfer will continue to click is the damping factor. There are studies for determining damping factors, but it is generally assumed to be near 0.85. The equations, incorporating damping factor, looks like

$$\mathrm{PR}(i) = (1-d) + d \sum_{j \in B_i} \frac{\mathrm{PR}(j)}{L(j)} \tag{2}$$

Note that the formula used in the above paper does not necessarily represent probability distribution as claimed in the paper since probability of i is bounded below by $(1-d)$, and when added will well exceed 1. However, the following change will result in page ranks being probability.

$$\mathrm{PR}(i) = \frac{1-d}{N} + d \sum_{j \in B_i} \frac{\mathrm{PR}(j)}{L(j)} \tag{3}$$

In the above N is the total number of papers available.

Keeping in mind researchers pursuing more complex problems that requires techniques not adequately accessible/comprehensible to researchers of average calibre, we have decided to rank both papers and authors according to the quality of average papers citing them. Citations received from papers or authors of significant quality will contribute more than the ordinary papers/authors.

One approach is to give an individual paper a score equal to the weighted average of the papers' citing it; weight being the number of citations the paper has attracted. The higher the average score of citing a particular paper has, the higher its score.

The score of the vector of pagerank can be obtained by computing the dominant eigenvector of the paper citation matrix. We follow an iterative method to compute this eigenvector. The higher the value of pagerank, the more significant the page will be considered. To evaluate the value of pagerank, we need to consider a matrix of eigenvalues so that we can work with the eigenvector. After each iteration we check the difference between the previous score vector of the papers with the corresponding new vector. This iterative process stops once the difference between the corresponding components of the two vectors become less than a prespecified ϵ value.

In Section II, we present our proposed algorithm with an elaborate explanation derived from an example along with initial evaluation and comparison. Section III-A covers our work on dataset, following that we present our result analysis in Section IV and provide our conclusion in Section V.

II. Proposed Algorithm

In our experimentation phase, we created a small dataset consisting of a number of papers. In Fig. 1 we have drawn a graph representing paper to paper citation of a number of papers and evaluated it by Google page rank algorithm [15] and our proposed algorithm. We present our findings and discuss a few fundamental differences in Google page rank algorithm [15] and our proposed algorithm.

$$\mathrm{PR}(i) = \sum_{j \epsilon P_i} \frac{\mathrm{PR}(j)}{O(p_j)} \tag{4}$$

In our work we introduce a variation of page rank algorithm and the formula we used to calculate paper rank from a dataset. In Equation(4) we simply used the summation of ranking value of all paper $\mathrm{PR}(j)$ that are citing another paper P_i divided by the total number of paper, paper P_j is citing. $\mathrm{PR}(i)$ is ranking value of P_i and $j \epsilon P$ means a set of papers that are citing paper P_i. $O(P_j)$ is the total number of paper cited by P_j.

$$\mathrm{PR}(i) = \frac{\mathrm{PR}(i)}{\sqrt{I(p_i)}} \tag{5}$$

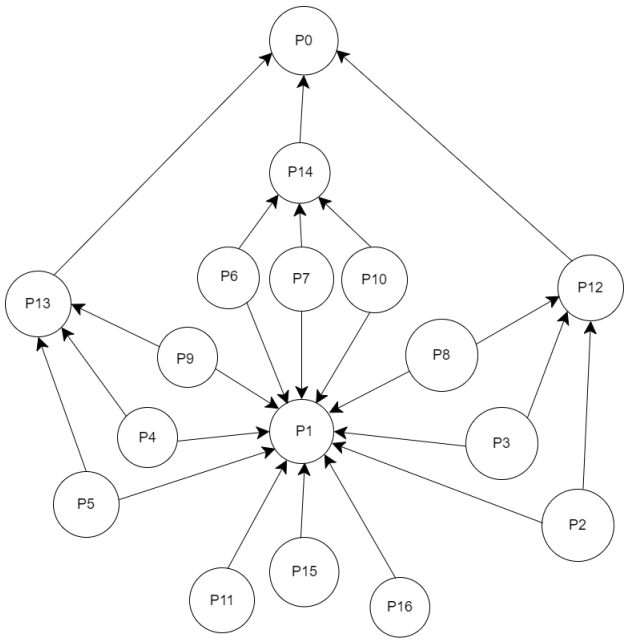

Fig. 1. Paper-Paper Citation Network

Brin and Page introduced Equation(4) in their page rank algorithm [1]. For our algorithm, when we divided the value $\text{PR}(i)$ by the number of paper that is cited paper P_i. We used the Equation(5) to get an average and better value of $\text{PR}(i)$. Here, by $I(p_i)$ we denote the number of papers that is cited paper P_i. As a result of adding the square root in Equation(5), we are dampening the impact of citation count of a paper. Furthermore, in Equation(5), the ranking score of a paper is decided by the ranking score of the papers that cited it. As a consequence, the equation would give more weight to papers that have been cited by well scored papers rather than its citation count. We want to have such rank value of papers that summation of all rank values add up to one ($\sum \text{PR}(i) = 1$). Since this is not the case for the ranking values obtained by Equation(5), by using Equation(6), we normalized the ranking value of the papers after they are calculated. In Equation(6), the denominator in the right hand side is the summation of the rank values of all papers obtained from Equation(5).

$$\text{PR}(i) = \frac{\text{PR}(i)}{\sum_{\forall j \in P} \text{PR}(j)} \tag{6}$$

In Fig. 1, we made a paper-paper network. In this network, $P1, P2, \ldots P16$ are papers written by different authors. In the Fig. 1 the arrow represents citation link from a paper to another paper. In the Fig. 1 we can see that $P1$ has a total of twelve edges directed to it. This means that the paper $P1$ is cited by twelve different papers. We used Google page rank algorithm to rank the papers firstly and then from the ranking value of papers we have calculated the rank of the authors.

Our purpose is to rank the authors in a manner such that the ranking can show the importance of the authors' work and their influence in the scientific world. The newly introduced variation are giving more importance to authors with deeper insights. Our goal is to establish a system where an author's paper have a value for the citation number as well as which paper is citing that paper. For example, as illustrated in Fig. 1, $P0$ and $P1$ are two papers written by two different authors. Now, paper $P1$ is cited by twelve lowly ranked papers as these papers have not been cited. On the other hand, although paper $P0$ has been cited by only three papers but these three papers have a significant ranking value as they are highly cited (three times each). Therefore, we can see that although $P0$ has been cited less than $P1$, when compared to $P1$, on the average $P0$ is cited by papers of much higher importance. As a result, we conclude that $P0$ deserves a better ranking than $P1$ despite having a lower citation count. But, if we evaluate this graph 1 by the Google pagerank algorithm [15] we will see that $P1$ has higher rank value than $P0$ just because of the highly citation number of $P1$.

We used Google's page rank algorithm [15] to rank all papers in the network presented in Fig. 1. From the above diagram, we can see that the score of $P1$ is much higher than the other papers. This is because $P1$ is cited by twelve different papers whereas other papers are cited by lesser number of papers. But from the Fig. 1 we can also see that $P0$ is ranked higher than all the other papers, though it is cited by only three papers. From the Table I we can see that $P1$ has the highest value and $P0$ has a lesser value than $P1$. We can also see that the $P12, P13$ and $P14$ has similar scores and ranked just after $P0$. All other papers also have similar scores and ranked consecutively.

TABLE I
RANKING VALUE FROM PAGERANK ALGORITHM

Rank	Papers	Ranking Values
1	P1	0.2235
2	P0	0.2061
3	P12	0.0689
3	P13	0.0689
3	P14	0.0689
4	P2	0.0689
4	P3	0.0303
4	P4	0.0303
4	P5	0.0303
4	P6	0.0303
4	P7	0.0303
4	P8	0.0303
4	P9	0.0303
4	P10	0.0303
4	P11	0.0303
4	P15	0.0303
4	P16	0.0303

From Table I we can see the ranking value of the papers. Our objective is to rank authors as well. Initially we have used scores of papers to calculate rank of authors. We distributed the score of a paper equally to all its co-authors. For example, for a paper of score 6 with three co-authors, each of its authors receive a score of 2.

$$Ar(i) = \sum_{j \epsilon P} \frac{\text{PR}(j)}{N(a_j)} \tag{7}$$

$Ar(i)$ is the score of author i. P is a set of papers written by author i. Score of paper j is defined as PR(j). $N(a_j)$ is the number of authors who have written paper j.

A. Comparison

If we see our two tables which are Table I, and Table II we can see that in page rank algorithm $P1$ is the first ranked with the value of 0.2235; but in our algorithm paper $P0$ is the first with 0.2838. So, the reason behind $P0$ being the top in our ranking system is, in Table 1 we can see that $P1$ is cited by many papers but those papers do not have any citation which means that the value of the other papers which cited $P1$ are low. On the other hand, $P0$ is cited by those papers which are cited by many other papers which means the papers which cited $P0$ carries more value. So, in our algorithm we are getting the value which we think should be the actual rank of the papers.

TABLE II
ALGORITHM RANKING VALUE WITH SQUARE ROOT

Rank	Papers	Ranking Values
1	P0	0.2838
2	P1	0.1029
3	P12	0.0693
3	P13	0.0693
3	P14	0.0693
4	P2	0.0338
4	P3	0.0338
4	P4	0.0338
4	P5	0.0338
4	P6	0.0338
4	P7	0.0338
4	P8	0.0338
4	P9	0.0338
4	P10	0.0338
4	P11	0.0338
4	P15	0.0338
4	P16	0.0338

III. METHODOLOGY

A. Dataset

Although there are many large public datasets listing papers and its authors, it was challenging for us to find a large dataset that also contains all the outlinks of the papers. We finally settled on using the dataset of ArnetMiner website [17]. We utilized their version 10 of the accessible datasets. This dataset incorporates all the papers from DBLP, the citation relationship between these papers in the form of references, citation count, abstract, publishing year and venue. For ease of code implementation, we removed some special characters from the the author and paper names. We also removed the abstract, publishing year and venue from the dataset. In our sanitized dataset, the total number connections of paper to author is 282525 and paper to paper connection is 634395 from (2017-10-27).

However, when we crosschecked the number of citation count from Google scholar, we noticed a discrepancy between the total citation count of all papers and the total number of references present in the papers in the dataset. This is because the dataset contains an incomplete history of the papers and as a result not all the papers cited by the papers in the dataset are present in the dataset. To avert this issue of mismatched count of the papers, we counted the outlink count of a paper as the number of valid outlinks of that paper present in our dataset.

B. Algorithm Implementation

To implement the algorithm we used to import pandas, csv and networkx libraries as we used python programming language. After taking inputs from the sanitized dataset we implemented the algorithm below to calculate the ranking of papers:

Algorithm 1 Paper rank Algorithm

$Error \leftarrow 0$
for $\forall i \in n$ **do**
    $PR(i) \leftarrow \frac{1}{N}$, LPR$(i) \leftarrow 0$
    $Error \leftarrow Error + |LPR(i) - \text{PR}(i)|$
end for
$eps \leftarrow 0.000001$
while $Error > eps$ **do**
    $sum \leftarrow 0$
    $LPR(i) \leftarrow PR(i)$, for all $i \in n$
    for $\forall i \in n$ **do**
        for $\forall j \in i$ **do**
            $PR(i) \leftarrow PR(i) + \frac{PR(j)}{O(j)}$
        end for
        $PR(i) \leftarrow \frac{PR(i)}{\sqrt{I(j)}}$
        $sum \leftarrow sum + PR(i)$
    end for
    $PR(i) \leftarrow \frac{PR(i)}{sum} * N$
    $Error \leftarrow Error + |LPR(i) - \text{PR}(i)|$, for all $i \in n$
end while
$PR(i) \leftarrow \frac{PR(i)}{N}$, for all $i \in n$

Here, PR(i) is paper rank value of paper i. LPR(i) is last paper rank value of paper i. Initially, we gave all papers rank value as $\frac{1}{N}$ and last rank value of paper as 0. The total number of the papers is denoted by N. $Error$ denotes the sum total of the difference between PR(i) and LPR(i). eps is very little value which decides the accuracy of our result. Once $Error$ becomes less than eps, we come out of the loop. In the while loop we are calculating the paper rank value as state in chapter II.

We validate the code by checking the result of a small paper network created by us. The network we used to validate and all the necessary codes and data can be found in this link [https://github.com/Roy101/Author-Ranking-IEEE-ICIEST].

IV. RESULT ANALYSIS

To calculate the result, we collected a list of recent Nobel laureates, Touring award winners and Fields medalists and searched their names in our dataset. When we compared their ranking position obtained using our algorithm against the page rank algorithm, our algorithm gave them slightly better ranking

value. We are giving a comparison table highlighting these award winning authors' ranking in both algorithm.

TABLE III
AUTHOR RANKING COMPARISON

Author Name	pagerank Ranking	Our Ranking
Yoshua Bengio	92	92
Leslie Lamport	1156	1152
Shafi Goldwasser	11625	11622
Elon Lindenstrauss	41590	41563
Vladimir Voevodsky	88808	88783
Shuji Nakamura	161976	161970

After going through the Table III, one might say the improvement is not significant. Currently, the pagerank algorithm gets to rank an author according to their citation count hence this algorithm does not account for all the connections to rank the authors. Despite not having all the connections within our dataset that we needed, due to not having a full dataset; we can see that our algorithm produced a better result. Having a proper complete dataset can allow us to have more remarkable results where the difference in ranking within these award winning authors will be more noticeable.

V. CONCLUSION

Our work focused on ranking authors and their scientific papers in a fair manner. The concept of our approach is to construct an algorithm to rank scientific papers and researchers whereas the papers and authors of great significance do not get penalized for their work not being easily understandable by their peers. We have compared our results with many well known rank lists including famous scientists. We have considered lists of recent Nobel laureates, Turing Prize winners and Fields medallists. Experimental results obtained are favourable. While this system could be implemented within the software system for taxonomic search, it may be possible to further improve this result by having a different exponent instead of square root in Equation(5).

REFERENCES

[1] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation rank-ing: Bringing order to the web.," Stanford InfoLab, Tech. Rep., 1999

[2] R. Costas and M. Bordons, "The h-index: Advantages, limitations and its relation with other bibliometric indicators at the micro level," Journal of informetrics, vol. 1, no. 3, pp. 193–203, 2007.

[3] J. E. Hirsch, "An index to quantify an individual's scientific research output that takes into account the effect of multiple coauthorship," Scientometrics, vol. 85, no. 3, pp. 741–754, 2010.

[4] S. N. Dorogovtsev and J. F. Mendes, "Ranking scientists," Nature Physics, vol. 11, no. 11, pp. 882–883, 2015.

[5] L. Lü, T. Zhou, Q.-M. Zhang, and H. E. Stanley, "The h-index of a network node and its relation to degree and coreness," Nature communications, vol. 7, no. 1, pp. 1–7, 2016.

[6] G. Pacheco, P. Figueira, J. M. Almeida, and M. A. Gonçalves, "Dissecting a scholar popularity ranking into different knowledge areas," in International Conference on Theory and Practice of Digital Libraries, Springer, 2016, pp. 253–265.

[7] C. Zhang, C. Liu, L. Yu, Z.-K. Zhang, and T. Zhou, "Identifying the academic rising stars," arXiv preprint arXiv:1606.05752, 2016.

[8] E. Oberesch and S. Groppe, "The mf-index: A citation-based multiple factor index to evaluate and compare the output of scientists," Open Journal of Web Technologies (OJWT), vol. 4, no. 1, pp. 1–32, 2017.

[9] T. Amjad, A. Daud, and N. R. Aljohani, "Ranking authors in academic social networks: A survey," Library Hi Tech, 2018.

[10] A. Gibbons, "The life of maryam mirzakhani," Journal of Mathematics Education at Teachers College, vol. 10, no. 1, pp. 11–16, 2019.

[11] F. Zhao, Y. Zhang, J. Lu, and O. Shai, "Measuring academic influence using heterogeneous author-citation networks," Scientometrics, vol. 118, no. 3, pp. 1119–1140, 2019.

[12] T. Amjad, Y. Rehmat, A. Daud, and R. A. Abbasi, "Scientific impact of an author and role of self-citations," Scientometrics, vol. 122, no. 2, pp. 915–932, 2020.

[13] M. Kosmulski, "Nobel laureates are not hot," Scientometrics, vol. 123, no. 1, pp. 487–495, 2020.

[14] A. Daud, S. Arabia, T. Amjad, H. Dawood, and S. H. Chauhdary, "Topic sensitive ranking of authors,"

[15] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web searchengine,"Computer networks and ISDN systems, vol. 30, no. 1-7, pp. 107–117,1998.

[16] 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. 16 569– 16 572, 2005.

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

[18] M. Ley, "Dblp: Some lessons learned,"Proceedings of the VLDB Endowment,vol. 2, no. 2, pp. 1493–1500, 2009.

[19] T. Amjad, A. Daud, and A. Akram, "Mutual influence based ranking of authors,"Mehran University Research Journal of Engineering 'I&' Technology, vol. 34, no. S1, pp. 103–112, 2015.