A Study on Author Ranking

by

Md. Adnanul Haq
18301049
Md. Noushin Islam
18301044
Labiba Tasfiya Jeba
18101529
Iffat Afsara Prome
18101425
Palash Ranjan Roy
18101530

A thesis submitted to the Department of Computer Science and Engineering
in partial fulfillment of the requirements for the degree of
B.Sc. in Computer Science and Engineering

Department of Computer Science and Engineering
Brac University
September 2021

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Palash Ranjan Roy

Palash Ranjan Roy
18101530

Md. Noushin Islam

Md. Noushin Islam
18301044

Labiba Tasfiya Jeba

Labiba Tasfiya Jeba
18101529

Iffat Afsana Prome

Iffat Afsara Prome
18101425

Md. Adnanul Haq

Md. Adnanul Haq
18301049

Approval

The thesis/project titled “A Study on Author Ranking” submitted by

1. Palash Ranjan Roy(18101530)
2. Md. Noushin Islam(18301044)
3. Labiba Tasfiya Jeba(18101529)
4. Iffat Afsara Prome(18101425)
5. Md. Adnanul Haq(18301049)

Of Summer, 2021 has been accepted as satisfactory in partial fulfillment of the requirements for the degree of B.Sc. in Computer Science on September 26, 2021.

Examining Committee:

Supervisor:
(Member)

Dr. Mohammad Kaykobad
Distinguished Professor
Department of Computer Science and Engineering
Brac University

Co-Supervisor:
(Member)

M Tanvir Kaykobad
PhD Student
School of Computing
Queen’s University

Head of Department:
(Chair)

Sadia Hamid Kazi, PhD
Chairperson and Associate Professor
Department of Computer Science and Engineering
Brac University

Abstract

In this research we have studied the problem of ranking papers and authors based upon their citations by other authors. Simple count of number of authors citing a particular paper or author may not be very representative of the contributions made simply because a paper or authors work may be of too prohibitive complexity for too many authors to cite. It may so happen that only accomplished researchers of grater heights can understand and assimilate their deep insightful contributions. So simple count of number of citations may well fail to capture the essence. So we proposed algorithms avoiding damping factors and assigning the weight $(1 - d)$ as has been assigned in page ranking algorithm irrespective of any citation. We have experimented with two versions of the algorithm. In the first version the score of a paper/author has been completed based on the scores of papers that have cited it. To do some justice to the papers that fail to attract citation from too many authors due to their complexity we have considered the average quality of citing papers. However, to give some weights to number of papers citing it we have multiplied the score with the square root of number of papers citing it. Experimental results of all these versions including paper rank have been presented.The current algorithms fail to properly rank authors whose citation counts are less but whose contributions are deemed important by the leading experts in the field. So in scoring of a researcher, we would like to factor in the average score of researchers who cited their work. To score these authors more fairly, we want to introduce a new Linear Programming formulation based scoring system for researchers. The purpose of Author Rank is to recognize the expertise of a person within certain subjects and what others think about the content they publish. We want to use an algorithm which is based on the System of linear equations. And the closest algorithm of this is Google's Page Rank algorithm. There is a need to consider the credibility of each author in order to examine the relativity of this broad data.

Keywords: Citation; Page Rank; Author Rank

Dedication

This dissertation is dedicated towards all the researchers around the world who are working hard to give some productive paper for the betterment for the society.

Acknowledgement

Firstly, all praise to the Almighty for whom our thesis have been completed without any major interruption.

Secondly, to our supervisor Dr. Mohammad Kaykobad and co-advisor M Tanvir Kaykobad for their kind support and advice throughout this work. Without their direction and motivation completion of this thesis was nearly impossible.

And finally to our parents for their unconditional support.

Table of Contents

List of Figures

List of Tables

Nomenclature

The next list describes several symbols & abbreviation that will be later used within the body of the document

APR	Author Paper Ranking
$IIRL$	Impact Increment Ranking Learning
$MuInfc$	Mutual Influence Based Rank
PR	PageRank

Chapter 1

Introduction

The advancement of modern research tools enables researchers to quickly access a wealth of scholarly data and encourages academic collaboration among scholars from various disciplines. With the massive penetration of computer technology in all walks of our life it has become so easy to access scholarly papers. For this reason it becomes very helpful for the researchers to conduct their research. But there is a problem here. It is a very lengthy process to filter out the most related scholars or references from the big amount of data. To resolve this problem, Google introduced the PageRank algorithm to enable users filter out papers of significant importance and thus save their time in surfing through innumerable papers possibly containing insignificant significance. Penetration of information technologies and other gadgets into our life have not only eased our life in accessing appropriate information but also have also turned our life very busy where time has become so scarce resource. There have been several paper/author ranking algorithms. These algorithms have been studied and we have also proposed some variants that addresses the issues of papers or authors that are too complex for other authors to understand and assimilate and cite although the results may be too deep for us to ignore those papers or authors. Publication of research findings is increasing exponentially. Number of journals and conferences published are increasing day by day. As a result this has become important to index, order and even rank research papers and authors so that knowledge workers can effectively spend their time in identifying more important contributions.

1.1 Problem Statement

The primary issue in the widely used author ranking algorithm is that it is based solely on the number of citations of the papers he or she publishes. However, a researcher may not receive that many citations if research findings are too deep to understand for most researchers. Usually authors with papers easy to understand are expected to be cited by most of the author, and therefore will get more citations although they may not have deep insightful results. We would like to address this issue existing in paper/author ranking by taking into consideration the quality of papers or authors citing their results.

To illustrate this fact, let us consider the case of Maryam Mirzakhani [15], an Iranian mathematician and professor of mathematics at Stanford University. The most prestigious award in mathematics called Fields Medal [15] was awarded to her for her outstanding contributions to the dynamics and geometry of Riemann surfaces and their moduli spaces [15]. However, so far, she was only cited 1797 papers. Her lower citation count is attributable to the fact that the articles she published generally dealt with difficult-to-understand issues that are not referenced by many researchers. As a result, her papers could not be well cited. As a result, her ranking in the current system does not at all reflect her high accomplishments as a researcher. We would like to introduce an algorithm for ranking papers o authors that will take this into account and rank them higher than algorithm like PageRank will do.

1.2 Literature Review

Many ranking algorithms have been introduced to rank papers and authors. Most of these methods give importance to the papers which are highly cited by other papers. The author of this paper [1] introduces a prototype of a large-scale search engine which is Google to makes heavy use of hypertext's structure. This paper looks at how to construct a large-scale system that can take advantage of the additional information found in hypertext. They also consider how to cope successfully with uncontrolled hypertext collections, in which anybody may publish whatever they wish. Moving on, The other paper [2] describes PageRank as a mechanism for objectively and automatically evaluating Web sites, effectively measure the human interest and attention paid to them. PageRank is a worldwide rating of all websites purely based on their location in the Web's graph structure and it is independent of their content. They also demonstrate how to calculate PageRank for a large number of pages quickly. The author of another paper [3] proposed the $\overline{h}$ indicator to distinguish across scientists with different co-authorship patterns. $\overline{h}$ will give extra credit to young and mid-career scientists who lead vigorous independent research programs, encourage them to take on younger students and postdocs without penalty, and discourage them from instead collaborating on research projects led by more senior scientists. On the other hand the paper [4] shows the micro level link between the h-index and other bibliometric indicators. The idea that the h-index underestimates the achievement of certain highly visible but low-productive writers is also examined here.The ArnetMiner [5] system is a search and data mining system for academic publications on the web, based on social network analysis to find connections between researchers, conferences, and publications is the subject of this paper, which tackles a number of critical concerns. The system focuses on automatically extracting researcher profiles from the web, integrating publishing data from current digital libraries into the network, modeling the whole academic network, and providing search services for the academic network. They also mention a unified modeling technique for modeling characteristics of articles, authors, and publication sites all at the same time. The authors [10] provide an operator H that acts as a connecting chain between degree, H-index, coreness, and other intermediate indices in this article. Coreness measure to identify tightly interlink groups within a network. The H-index is a measurement of assessing an author's scholarly output and performance over time. These connecting chains have been treated unrelated in the past. Using the operator H to achieve coreness seems to be the opposite of the iterative removal of nodes with degrees less than k, which is commonly used to identify a network's k-core. k-core is the maximum connected subgraph of a graph where all the vertices have at least degree of k. The authors of the paper [11] analyzed the impact of several academic features on the rankings of the most well-known Brazilian academics in various fields of knowledge. They concentrate on a list of the country's most popular scholars, as well as an analysis of how various areas are distributed around the country. For scholars in different areas, they want to characterize scholar ranking in terms of how each knowledge area is distributed and how distinct academic characteristics are associated with the scholar's ranking position. They also find some interesting facts that include the dominance of Physicists and Health scholars in the top ranking positions. Here [12], the number of citations is used to measure an author's scientific influence. The authors of this paper have presented a group of young researchers who have recently published their first work. This study addresses the challenge of predicting the top k percent of researchers who earn the largest citation increment in a given year. They also look at several characteristics that might influence an author's ability to rise quickly and develop a unique impact increment ranking learning (IIRL) algorithm that uses those elements to predict academic rising stars. Rather than predict the precise citation value, the authors focus on ranking the citation increments of different writers in this paper. As a result, they define fast-rising researchers as those that achieve relatively high citation increments in a short period.The authors [13] determine the criteria for an advanced index and create a new index, the mf-index. mf-index integrates the benefits of current bibliometric indicators while minimizing their downsides as much as possible. This index makes a significant contribution by taking into account several criteria such as career duration, publication and citation age, citation weights for various types of citations, the field of study, and a number of co-authors.In addition, they compare the mf-index values with the other current bibliometric indicators and find that the mf-index better balances different elements of researchers, resulting in a more fair comparison of their performance. This paper [14] identifies the author ranking approaches, analyzes their benefits and drawbacks, and compares them. The authors also want to find the problems and future directions of academic object ranking. They mentioned that they want to do this for future scholars. This paper [16] introduces a new algorithm which is the Author PageRank algorithm(APR) for ranking authors. APR is different from the other methods because it is the only one that takes authors into account in a heterogeneous network. This APR method calculates the importance of authors and papers recursively.The authors [17] of this paper looked at the relationships between self-citation trends and their impact on scientific impact. To anticipate the impact of self-citations on future H-Index, they employed regression-based prediction models.In this paper [18] the author introduced a new indicator c to represent Nobel laureates. The author has told with a vast number of examples that there is no other bibliometric indicator that places as many recent Nobel laureates in the top 6000 as c. In this article [19], the authors first utilize the traditional author ranking method, then to compute the relevance of authors with specified subjects they use the Vector Space Model (VSM), and then use k-Means clustering to split the authors into topic-based groups.

1.3 Research Objective

To better rank researchers according to their contribution, we aim to develop a new algorithm based on a system of linear equations that give more weights to citations from authors/papers of high ranks. We use linear algebra techniques to find solution to systems having a very large number of variables and equations as is true in case of authors and papers available in the literature. Iterative schemes, convergence of which is established theoretically, are used to approximate solutions

For our problem we have a set of authors $A = (a_1, a_2, \ldots, a_n)$. Here $a_1, a_2, \cdots$ are identities of authors. Papers are denoted by $P_1, \cdots, P_n$. Here i represents the number of papers written by a particular author. Then we have a set of papers $P = (p_1, p_2, \ldots, p_n)$. Here p_1, p_2 are individual papers. Lastly we also have a set of references for each paper p_i in P where $p_i = R_{p_i}$. Here R_{p_i} represents how many papers are cited by a singular paper.

Our goal is to formulate a system of linear equations based on these variables, using which we can assign a score to each author that would represent their research contribution in the scientific community in a better way.

Chapter 2

Related Work

In this field of author ranking or expert finding, previously many methods were approached. In the survey paper [1] an extensive rundown of a large-scale web search engine is given, which is the first such in-depth public description we know to this date. There are new technical challenges included in using the additional knowledge in hypertext to deliver better search results, apart from scaling traditional search methods to data of this volume. In this case, they also look at the question of how to process unregulated hypertext (meaning which contains texts of other pages) collections effectively, where everyone can publish what they want.

Furthermore, in another paper [2], the authors outline PageRank that is approached to designate web pages for Google's search engine, which is a system for objectively and mechanically rating web pages. And also effectively measuring human interest. This PageRank is near to our goal we are searching for. This is why there was devoted attention to them. In addition, PageRank has a better description standing for the model of a random web surfer surfing the web. But the difference is, we are giving attention to authors rather than giving attention to websites.

Consider a regular website surfer's browsing usage where she may use a search engine to find a website on the topic of her preference. If the website meets her needs, she may keep browsing to pages via one of the hyperlinks (a link which is a reference to other links or websites) given on her current page.

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.

However, In paper [3] the index $\overline{h}$ is proposed which is kind of an adaptation of the previous h-index method. The merit of scientists should not be determined by the number of publications (which happens in the h-index). Rather, it should be evaluated by the quality of their paper. In this paper, authors came up with a bibliometric indicator $\overline{h}$ that gives extra credit to those authors who publish alone or in a small collaboration. Also, this method enables coauthors to achieve credits from their collaborated publications.

Dorogovtsev and Mendes [4] represented one of the most popular yet simplistic methods of ranking scientists, which had become widely used methods of all time. h-index is adequately reliable as this method does not take into account too many complex computations. Instead, it only requires two comparable factors: the number of total citations of the page and its rank. But still, this method comes with its drawbacks. Since the increase of the h-index means the higher rank of the page so if someone cites their own page many times, then the h-index won't be able to determine the difference; hence it will increase the rank of the page.

Hirsch [7] 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. As a result, he proposes $\overline{h}$ as a helpful metric for distinguishing across scientists with various co-authorship patterns. The $\overline{h}$ indicator, on the other hand, should not discourage partnerships, which are essential for scientific development. The $\overline{h}$ indicator was created solely to measure individual contribution of independent researchers. The publication that counts for a scientist's h-index also counts for his or her $\overline{h}$ when evaluating his or her independent research output.

The paper [8] represents a study where the authors Amjad and Daud proposed that the possibility of authors getting consideration or notice depends on how many authors they have been collaborated with. This chance of the influence even gets higher if they have collaborated with renowned authors. The authors have introduced a ranking method named MuInfc (Mutual Influence Based Rank) method which determines the influence of authors, co-authors and presence of prestigious authors and they have received better results using this methods than implementing the existing baseline formula. The authors have analyzed the position of an authors through observing his or her own work and the influence of works of their collaborating authors.

Dorogovtsev and Mendes [9] claimed in their paper that the merit of a researcher is rather evaluated by his or her strongest results than by the number of publications. This theory is continually rejected by the commonly used h-index-based ranking of scientists. Essentially in this paper, they perceived another index-based ranking named o-index which is favorable for proper ranking. It escalates with an increasing average number of citations per paper. The authors stated that the o-index categorizes good researchers fairly and delivers the scientists with a natural, easily applicable ranking standard.

Lü et al. [10] introduce an operator H that acts as a connecting chain between degree, H-index, coreness, and other intermediate indices. Coreness measure to identify tightly interlink groups within a network. The H-index is a measurement of assessing an author's scholarly output and performance over time. These connecting chains have been treated unrelated in the past. Using the operator H to achieve coreness seems to be the opposite of the iterative removal of nodes with degrees less than k, which is commonly used to identify a network's k-core. k-core is the maximum connected subgraph of a graph where all the vertices have at least degree of k.

In paper [11], the authors analyzed the impact of several academic features on the rankings of the most well-known Brazilian academics in various fields of knowledge. They concentrate on a list of the country's most popular scholars, as well as an analysis of how various areas are distributed around the country. For scholars in different areas, they want to characterize scholar ranking in terms of how each knowledge area is distributed and how distinct academic characteristics are associated with the scholar's ranking position. They also find some interesting facts that include the dominance of Physicists and Health scholars in the top ranking positions.

Ranking authors and papers have been widely research area in the literature. The number of citations is used to measure an author's scientific influence. The authors of the paper [12] have studied a group of young researchers who have recently published their first work. This study also addresses the challenge of predicting the top k percent of researchers who earn the largest citation increment in a given year. They also look at several characteristics that might influence an author's ability to rise quickly and develop a unique impact increment ranking learning (IIRL) algorithm that uses those elements to predict academic rising stars. Rather than predicting the precise citation value, the authors focus on ranking the citation increments of different writers in this paper. As a result, they define fast-rising researchers as those that achieve relatively high citation increments in a short period.

Oberesch and Groppe [13] determine the criteria for an advanced index and create a new index, the mf-index. mf-index integrates the benefits of current bibliometric indicators while minimizing their downsides as much as possible. This index makes a significant contribution by taking into account several criteria such as career duration, publication and citation age, citation weights for various types of citations, the field of study, and a number of co-authors. In addition, they compare the mf-index values with the other current bibliometric indicators and find that the mf-index better balances different elements of researchers, resulting in a more fair comparison of their performance.

In the survey paper by Amjad, Daud and Aljohani [14] we found some classifications for ranking the authors. In this following paper, the authors had categorized the existing algorithms using three main categories: *Link Analysis Methods, Text Similarity Methods and Learning Based Methods.* Link analysis methods are often reliable when we are interested in analyzing the node relationships. In addition to that, it is also conducive in evaluating, fetching information and discovering knowledge of all forms of networks. However, it is quite possible to carry out these techniques in an unsupervised way. This method includes the technique that calculates the rank of an academic object, mainly authors, by reckoning with the linkage structure of a relevant graph. These Link analysis methods are further divided into two main subclasses. They are Iterative and Bibliometric methods. In the Iterative Method, it uses a number of iterations to calculate the ranks of authors. On the other hand,

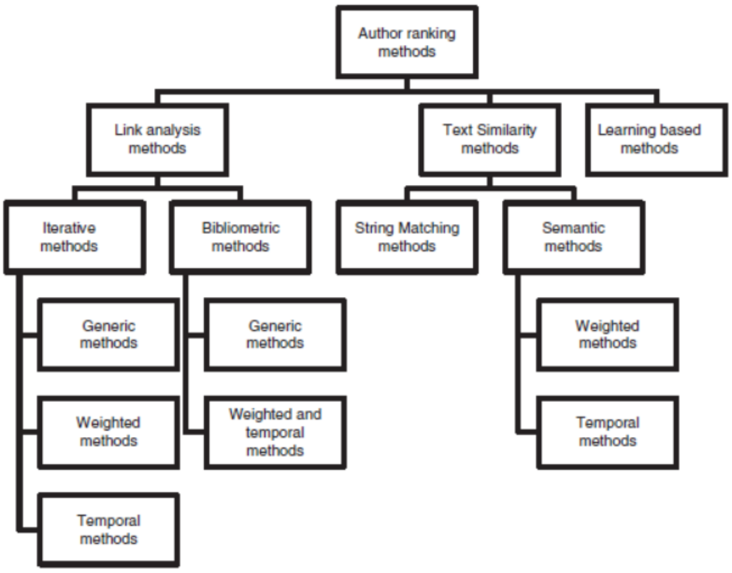


Figure 2.1: Classification of Ranking Methods (Source: Amjad, Daud, Aljohani [14])

the Bibliometric Methods are dependent on some sort of calculation involving the bibliometric citations. Text Similarity Methods are the second main category mentioned in that paper. This method takes in consideration from a specific text data and finds some related text and uses the data for rank calculation. Moving on to the third main category mentioned in this paper is Learning Based Methods that apply the approach of machine learning and classification rules to estimate the ranks of academic objects.

This paper [16] proposes Author PageRank (APR) as a tool for determining the academic impact of authors in a heterogeneous network (interconnected nodes and links of different types) to represent the data. Zhao et al. used this heterogeneous network to evaluate the authors quantitatively. The network has two types of nodes. The nodes are authors and papers. In between those nodes there are two types of link. One is a citation link between papers and the other one is an authorship link between a paper and an author. They show that, on two very large data sets, it outperforms 10 other approaches. They also demonstrate that for all the other strategies, their ranking outcomes vary greatly. The APR method by Zhao et al. compared with other ranking methods. This method successfully differentiated award winner authors compared to other ranking methods. In this method we can rank both papers and authors at the same time. The APR method was experimented on two large academic networks, one in the Health domain and other is in the Computer science domain.

The research paper [17] indicates self-citation which is used to view the progress of one's previous work in a new publication. The author claimed it is necessary because scientists can show that they are building upon their own previous work which avoids redundant expansion in the length of manuscript by re-introducing the already published work. This method is stated to have a considerable impact on science by contributing to artificial manipulation.

Kosmulski [18] studied a vast number of examples to reach conclusion that there is no other bibliometric indicator that places as many recent Nobel laureates in the top 6000 as C. C is a composite scoring that was proposed by Loannides et al in 2016 which is not field-normalized. In field-normalized indicators, the rankings of recent Nobel laureates are not much higher. Field normalization may play an important role, especially for economists if it is compared with chemists, physicists, or doctors. Field normalization means to create fields manually. The number of publications and citation rates in Chemistry, Medicine, and Physics are comparable. That is why field normalization is not so important when only these three disciplines are compared. Kosmulski in his paper asserts that 97 Nobel Prize winning scientists, in Chemistry, Economics, Medicine, or Physics, are compared to the achievements of top non-Nobel scientists in terms of many standard bibliometric indicators such as citations, Hirsch index, highly cited papers, number of publications, and hot papers of C. It may be noted that t according to the current criteria, *hot paper* is a paper which is published in the last two years and got enough citations to be ranked in the top 0.1 percent of papers in their academic disciplines. On the other hand, *highly cited* articles are only relevant for the last ten years. In this paper, some examples are given to give us a clear idea about recent Nobel laureates. The idea is that recent Nobel laureates do better in terms of C than in any other bibliometric indicators. For example, only 32 recent Nobel laureates were among the top 6000 scientists in terms of ch where ch is the Hirsch-type index that excludes self-citations and is not field normalized. Only 17 recent Nobel laureates were among the 6000 Highly Cited Researchers, 4 recent Nobel laureates were among the top 6000 scientists by several hot publications, and 2 recent Nobel laureates were among the top 6000 scientists by highly cited papers. When ranked using ch, the author shows that the number of Nobel laureates were among the top 6000 scientists is very low. So this ch indicator does not represent the Nobel laureates very well. But in terms of C, 45 recent Nobel laureates were among the top 6000 scientists, which is high, compared to the other indicators. That is why we can say that C is a better indicator to represent Nobel laureates.

The research paper [19] focuses on ranking the authors based on their expertise-related topics. Unlike the traditional ranking method, this technique is different because the calculation does not only depend on the generic ranking of authors. So we can find authors based on their expert area. also the author introduced some new algorithms to evaluate this ranking system. First, they rank the authors traditionally by ranking algorithms and after that they apply some more set of algorithms to identify the relevance of authors in some particular topics

2.1 PageRank

Among various ranking methods proposed by the authors to rank resources available on the web, Brin and Page [1] introduced the page rank algorithm which enjoys the widest acceptability. PageRank(PR) is an algorithm that attempts to estimate the importance of a page based upon the number of pages that cite it.

2.1.1 Important definitions

Definition 2.1 (Page rank) *The PageRank algorithm generally assumes that if many other pages cite a particular page or website then that page would probably bear significance proportional to the number of pages citing it.*

Definition 2.2 (Random surfer model) *This model is referred as chronological directed graphs, $M_t, t = 1, 2, 3, \ldots$ where the graph M_t has t vertices and t edges.*

Definition 2.3 (Damping factor) *The damping factor is denoted with d. Normally it is set between the range* $(0, 1)$. *It works as a probability and is used for the prevention of pages which have no out going links.*

Definition 2.4 (Eigenvalue) *if for a linear operator L, an eigenvalue of L is where x is a non-zero vector such that $Lx = \lambda x$*

Definition 2.5 (Eigenvector) *for a linear operation L, eigenvector is a non-zero vector where $Lx = \lambda x$, for a scalar value of λ*

The PageRank algorithm assumes a paper that is cited by more pages would be considered more important than a web page cited by a lesser number of web pages. This assumption appears to underestimate scientific works that cannot be cited by too many authors at least for the reason that these works are too complex to be comprehended by authors not able to fathom enough.

However, there is a complication to this approach. The problem is that it is certainly easy for someone to artificially inflate their own web pages' importance. For example, A wants to increase the importance of its web page. Essentially it creates lots of other pages and links to its main page. So using this strategy A can make its web page as dominant as it wanted it to be. So it is really prominent to define what it indicates for a page to be important. Another way it can be implicated in that a page would be considered important if other important pages are linked to that particular page. But this idea seems quite circular as it is tough to calculate the importance of the page if doing so requires knowing the importance of other pages' importance.

One way to calculate this, authors introduced Random Surfer Model. The concept is this suppose surfer browsing the web. The surfer begins on a web page that has been chosen at random, then it randomly picks a link from that page to another page that it visits. Here the random surfer keeps repeating this process. Now the idea is that the Random Surfer Model method will keep a score that maintains a count of the number of times our random surfer visits each page. Every time it lands on a new page, it will update the score of the page. The pages that have more links to them are more likely to be visited. So these pages will eventually have higher scores. So the surfer is linked from a more important page will matter more than a link from a less important page. After continuing the process multiple times, the resulting scores calculate what percent of the total score of each page is. This indicates the measures for the relative importance of these pages represented as what percent of the time a random surfer on the internet is expected to be on that page. Nevertheless, this approach has a setback. It is a fact that pages on the internet might not be connected to each other. Let's assume a separate network where a random surfer starts on a page and it keeps following links that only ever visit one set of pages on the web. It completely ignores the rest of the internet since none of the other pages are reachable by any of the links from the pages that were currently visited.

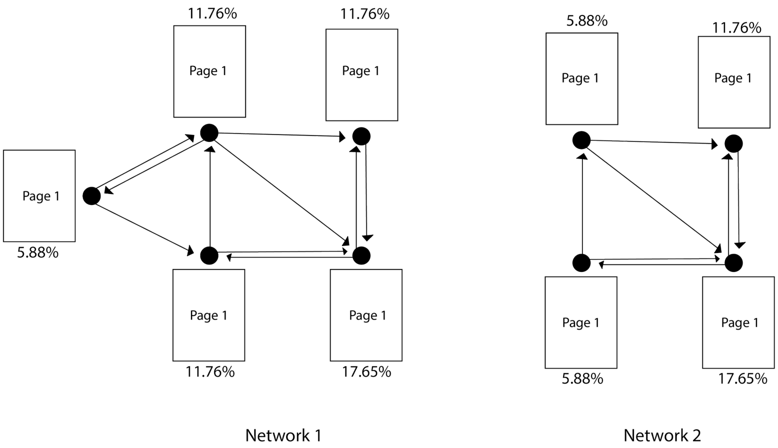


Figure 2.2: The Random Surfer model

The solution to this problem, a new concept is introduced which is the damping factor. For example, the damping factor is set to 0.85 then it stands for 85% of the time our random surfer will follow a link from the page that is currently on as it was doing before. Whereas 15% of the time though the random surfer will be switched to a page on the internet chosen completely at random. Now, this model takes any network of the web pages and calculates the relative importance of those pages. Eventually, a stable PageRank value will form for each page which will determine what order of search results should appear. So the more significant pages will appear first.

2.1.2 Page Rank Calculation

The formula of the PageRank as follows:

$$PR(A) = (1 - d) + d(PR(T1)/C(T1) + \cdots + PR(Tn)/C(Tn))$$

Here, $PR(A)$ is the PageRank of page A, $PR(Ti)$ denotes PageRank of pages Ti which link to page A. $C(A)$ is defined as the number of links going out of page A. d is the damping factor which is set between 0 and 1. Usually, we set d to 0.85. And assume page A has $T1 \dots Tn$ number of pages linked to it. Initially, PageRank of all the web pages is 1 as it is a form of a probability distribution.

Since PageRank's equation follows the $Lx = \lambda x$ structure, the score vector of pagerank can be obtained by computing the dominant eigenvector of the paper citation

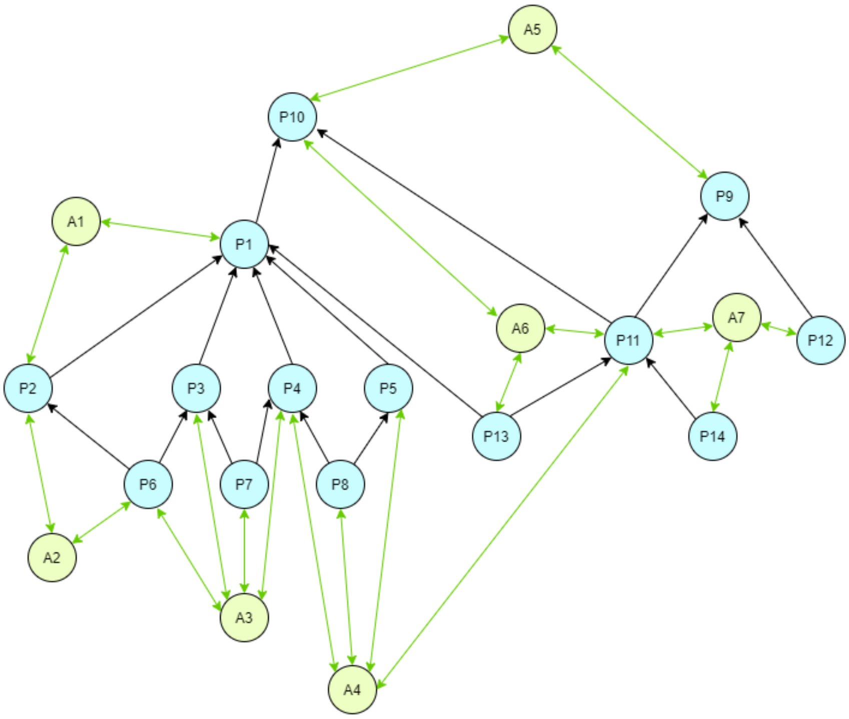


Figure 2.3: Paper-Author Citation Network

matrix. We follow an iterative method to compute this eigenvector. Here, 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 the above Figure 2.3 we have represented a small data set consist of both author and paper. Here the blue circle represent a paper and the green circle means a author. From the Figure 2.3 we can also see that the link between two papers is visualize with one directional arrow and the link between a paper and author is represent as bidirectional arrow. Here we deliberately avoid relation between two authors. From the Figure 2.3 we only took the paper-paper link and used google page rank algorithm to rank the papers.

Rank	Papers	Ranking Values
1	P10	0.2431
2	P1	0.2148
3	P9	0.0866
4	P11	0.0699
5	P4	0.0568
5	P3	0.0568
6	P5	0.0438
6	P2	0.0438
7	P8	0.0307
7	P7	0.0307
7	P6	0.0307
7	P14	0.0307
7	P13	0.0307
7	P12	0.0307

Table 2.1: Paper Ranking Value

2.2 Author Paper Rank

Zhao et al. [16] provide a new model named Author PageRank (APR) for the academic influence of the author with the PageRank algorithm. This is the first time where authors and papers are integrated into an academic network. The authors of this paper, decorate the paper based on heterogeneous network, author ranking, page rank, and scholarly data. PageRank-based algorithms give priority to citations from influential papers.

2.2.1 Important definitions

Definition 2.6 (Author PageRank algorithm) *In this algorithm paper and authors both are integrated into the academic networks. This method successfully distinguished award-winning writers.*

Definition 2.7 (H-index) *H-index excludes papers that have a low number of citations. It makes citation count easier. Authors, who publish one or more than one papers that are being highly cited, after they exceed a certain threshold value, this indicator penalize that authors.*

Definition 2.8 (G-index) *This indicator give credit for each paper's citation counts that are above a certain threshold (the G-index)*

Definition 2.9 (Heterogeneous network) *A network that contains different types of interconnected nodes and links is known as a heterogeneous network*

Traditionally, Academic influence is calculated by citation counts and by the use of H-index and G-index. Citation count plays a more dominant role in computing the score of authors of the page rank algorithms rather than focusing on more influential papers. The authors of this paper [16] introduced a new algorithm for ranking the authors and they tested the algorithm on two large networks, one in health and another in Computer Science. They [16] pointed out that their method can identify 8 Turing award winners among the top 20 authors. In their [16] model, they took into account the authors' network(co-authorship). So, authors and papers are ranked at the same time.

Some issues have been addressed in this paper [16]. There are some differences between academic networks and the Web. The first difference is, citation networks are acyclic in nature. Previously published papers are only cited by a paper, but a paper cannot cite the papers that are published later. The second difference found is that academic networks are inherently heterogeneous. Web pages are one type of node and hyperlinks are one type of link in the Web where PageRank is used. On the other hand, author and paper, these two nodes are found in academic networks.

Chen et al.(2007) presented a way to overcome the first obstacle named random walk by giving a lower damper factor in the PageRank algorithm [16]. A lower damping factor will make calculations easier. On the other hand, there are two solutions to the second problem. From the heterogeneous network, the author network is come from and then applying the PageRank algorithm on that author network. Another approach is to build the algorithms straight away on the heterogeneous academic network.

Normally, only older papers can be cited by the newer papers. APR(Author-PageRank) solves the acyclic network problem [16]. It gives links between papers and authors. While the same author writes older and newer, random walks can begin with the older paper and end with the newer paper. As a result, random walks will be able to see newer publications. By using the Author-PageRank method the heterogeneous problem gets solved by combining paper and author.

Zhao et al. used a heterogeneous network to evaluate the authors quantitatively [16]. The network has two types of nodes. The nodes are authors and papers [16]. There are two types of links between these nodes. Citation link is linking between papers and there is an authorship link between a paper and an author.
$a = (a_1, a_2, \ldots, a_n)$, a set of authors.
$p = (p_1, p_2, \ldots, p_n)$, a set of papers.

Furthermore, E_{PP}, is the citation link between papers and E_{PA}, denotes the relationship between a paper and an author. The author-citation network represents a Graph:

$$G = (a \cup p, E_{PP} \cup E_{PA})$$

.
In the paper [16] the authors have created $a[(m+n)x(m+n)]$ square matrix A where m is the number of papers and n is the number of authors in a network [16].

$$A = \begin{pmatrix} A_{PP} & A_{PA} \\ A_{AP} & 0 \end{pmatrix}$$

Here, A_{PP} is the paper paper relation and A_{PA} and A_{AP} are paper-author relationship where, $A_{PA} = A_{AP}^T$. Here, A is (nearly) symmetric along its major diagonal, excluding the submatrix A_{PP}. Another point to note in this paper [16] is that there is no relation between authors.

The main goal of this paper [16] is to generate a vector r for the given heterogeneous network $G = (a \cup p, E_{PP} \cup E_{PA})$. The author denotes r as the importance/influence of authors a and papers p.

$$M = \alpha B + (1 - \alpha)\frac{1}{n}ee^T$$

Here, B is a column scholastic matrix which the authors get from the adjacency matrix A [16]. Scholastic matrix is a square matrix used to describe the transitions of a Markov chain. Each entity of the scholastic matrix is a non-negative real number that represents probability. Markov chain is a system that experiences transition between one state to another according to certain probabilistic rules. A Markov process is one for which predictions about future events can be made purely based on its current state, and these forecasts are just as good as those that could be made knowing the process's whole history. In this paper [16], the authors interpreted an author's influence or importance as the random surfer visiting the node. They [16] have stated that every Markov chain does not have stationary distribution (stationary distribution is (stationary distribution is an entity that is unaffected by the effects of a matrix or operator). To guarantee stationary distribution, they [16] modify the network. A new scholastic matrix M is introduced by the authors in the paper [16]. Here, e is a vector of 1's, and α is the damping factor whose value is around 0.85 [1]. Here, n is the length of the matrix. The ranking we are getting from matrix M is the principal Eigenvector r of the matrix which is computed by $r = Mr$.

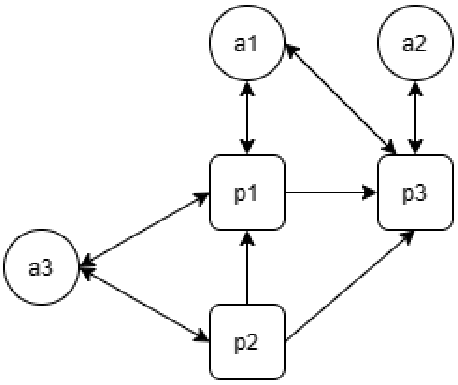


Figure 2.4: An Example of APR Framework

$$A = \begin{pmatrix} 0 & 1 & 0 & 1 & 0 & 1 \\ 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$B = \begin{pmatrix} 0 & 1/3 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 0 & 1/2 \\ 1/3 & 1/3 & 0 & 1/2 & 1 & 0 \\ 1/3 & 0 & 1/2 & 0 & 0 & 0 \\ 0 & 0 & 1/2 & 0 & 0 & 0 \\ 1/3 & 1/3 & 0 & 0 & 0 & 0 \end{pmatrix}$$

$$ee^T = \begin{pmatrix} 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \end{pmatrix}$$

$$r = \begin{pmatrix} r_{p_1} \\ r_{p_2} \\ r_{p_3} \\ r_{a_1} \\ r_{a_2} \\ r_{a_3} \end{pmatrix} = \begin{pmatrix} 0.17 \\ 0.06 \\ 0.31 \\ 0.21 \\ 0.09 \\ 0.16 \end{pmatrix}$$

The APR method is compared to various ranking algorithms. In comparison to previous ranking systems, this method successfully distinguished award-winning writers. We may rank both papers and authors simultaneously using this technique. The APR technique was tested on two big academic networks, one in the field of health and the other in the field of computer science [16].

Chapter 3

Data Collection and Pre-processing

To full fill our research we required some to data work on, therefore down below we have described how we got the data and tried to suit the data with our work in progress.

3.1 Data Collection

For our programming test purpose we needed a data set to run. Here, we have tried to use the information in a XML file named “dblp.xml” which we downloaded from https://dblp.org/xml/. The XML file mainly contains the information of different types of papers and their authors and also other necessary information about the papers. PhD Thesis papers, conference papers published in an LNCS volume, LNCS journals are mainly mentioned on that file. We also have different elements on that file, for an example:

$< inproceedingsmdate = ”2016 - 03 - 11”key = ”conf/sigmetrics/Dubois82” >$
$< author > DonaldF.Dubois < /author >$
$< title > AHierarchicalModelingSystemforComputerNetworks. < /title >$
$< pages > 147 - 155 < /pages >$
$< year > 1982 < /year >$
$< crossref > conf/sigmetrics/1982 < /crossref >$
$< url > db/conf/sigmetrics/sigmetrics82.html/Dubois82 < /url >$
$< /inproceedings >$

- “Inproceedings” used to describe a record of a conference paper published in an LNCS volume [6]. In the “inproceedings” element there is mdate which is the last modification of the record and the key [6] is the unique key of the record. DBLP keys resemble Unix file names with a slash separator. Conf/* for conference or workshop papers, and journals/* for articles published in journals, transactions, magazines, or newsletters are the two most significant sub-trees in the key namespace. The second portion of a DBLP [6] key usually specifies the conference series or journal where the papers were published. The last component of the key can be any sequence of alphanumeric characters; in most situations, these IDs are formed from the authors' names and the year of publication; occasionally, a letter is added to make this key portion unique.

- “author” contains the author name of the paper. There can be multiple “author” elements if there are more than one authors. The author elements' sequence within a record is essential; it should be the same as on the paper's title page.
- The paper's title is mentioned in the “title” element.
- The element ”pages” contains the paper's page number in the ”from-to” format. If the final page's number is uncertain, it is written from-. If the paper is only one page long, the page number is simply given without the hyphen.
- In which year a paper is published, stored in the “year” element.
- The “crossref” [6] field in the inproceedings record has the key of the proceedings record.
- The “url” section [6] contains the location of the table of contents.

3.2 Data Pre-processing

In the data processing section of our research, we have faced most of the difficulties. First of all, the data we used has some traditional Icelandic name as “Ingibjörg Sólrún Gísladóttir” which contains many non-ASCII characters. This makes it difficult to traverse the data or find matches. Secondly, as this is a huge data base the file size is extremely huge to compile as we always end up with out of RAM. Furthermore, there was a lot data such as ”school”, ”pages” and ”volume” were non necessary in our research. Although,we added some essential data as we needed that for our program as citation numbers. Furthermore we needed a smaller data set for our debugging purposes to cut down on running time of our algorithm during its development.Therefore, we decided to prune the data to create a smaller data set containing relevant information with regards to our research.

Chapter 4

The Modified Algorithm

In our experimentation phase, we created a small data set ourselves consist of a number of papers. Here in Figure 4.1 we made a graph representing paper to paper citation of a number of papers and evaluated it by google page rank algorithm [1] and our proposed algorithm. We present our findings and discuss a few fundamental differences in google page rank algorithm [1] and our proposed algorithm.

4.1 Experimental Setup

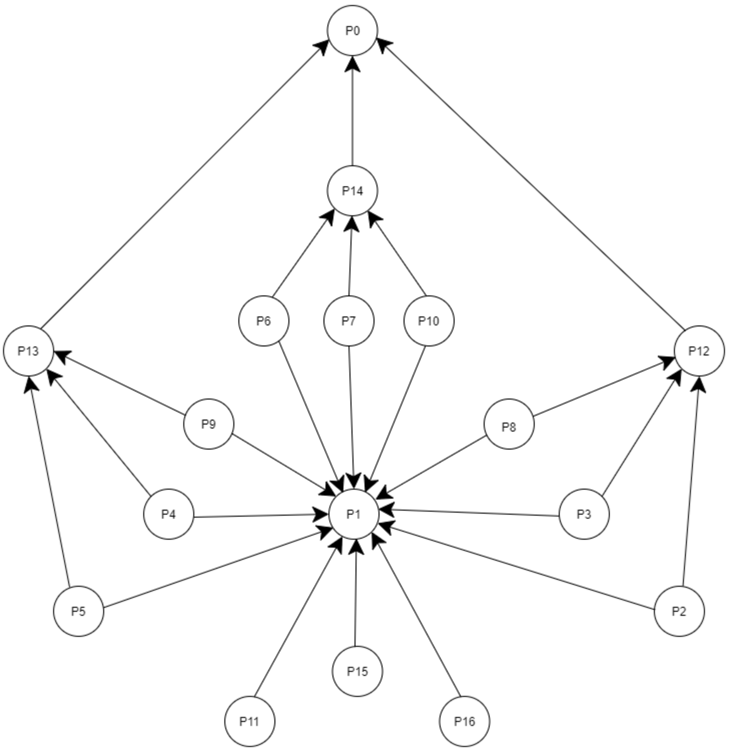


Figure 4.1: Paper-Paper Citation Network

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 our world. The ranking system we are using now is giving more importance to the citation number of an author.

This system is not giving sufficient attention to the fact of which an author's paper is cited by whom. Our goal is to establish a system where an author's paper will have a value for the citation number as well as which paper is citing that paper. For example, as illustrated in Figure 4.1, let us say $P0$ and $P1$ are two papers written by two different authors. Now, let $P1$ be cited by twelve papers but the ranking value of these twelve papers is not significant. Let $P0$ be cited by only three papers but these three papers have a significant ranking value. This is why $P0$ should have much more ranking value than $P1$. But, if we evaluate this graph 4.1 by the google pagerank algorithm [1] we will see that $P1$ has greater rank value than $P10$ just because of the highly citation number of $P1$.

Here in Figure 4.1, we made a paper-paper network. In this network, $P1, P2, \ldots P16$ are papers written by different authors. In the Figure 4.1 the arrow represent citation link from a paper to another paper. In the Figure 4.1 we can see that $P1$ has total twelve number of edge directed to it, this means that the paper $P1$ is cited by twelve different papers. Here, we used google page rank algorithm to rank the papers firstly and then from the ranking value of papers we will calculate the rank value of the authors.

$$P(i) = (1-d) + d \sum_{j \epsilon S(i)} \frac{P(j)}{L(j)} \tag{4.1}$$

Here d is the damping factor for the calculation of the rank. The rank value of paper i is dependent on rank values of other papers j contained in a set $S(i)$, divided by $L(j)$ of out-links from paper j. Here S is define as a set of papers. The set $S(i)$ contains all papers linking to paper i. After getting the rank value from the page rank algorithm [1] we drew a scatter diagram 4.2.

4.2 Initial Evaluation

We used google's page rank algorithm [1] to rank all papers in the network presented in Figure 4.1. From the above diagram, we can see that the rank value 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 paper. But from the Figure 4.1 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 2.1 and scatter diagram of Figure 4.1 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 the similar value and ranked just after $P0$. All the other paper also have the similar value and ranked consecutively.

From Table 4.1 we can see the ranking value of the papers. But our objective is to rank authors. At the beginning of the thesis we used these ranking values for rank the authors. Here 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. But after doing that we did not achieve any better result. So, we decide not to do that in any furthermore.

Rank	Papers	PageRank Ranking Values
1	P1	0.2235
2	P0	0.2061
3	P12	0.0689
3	P13	0.0689
3	P14	0.0689
4	P2	0.0303
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

Table 4.1: Paper Ranking Value from PageRank Algorithm

$$Ar(i) = \sum_{j \epsilon P} \frac{Pr(j)}{n(a_j)} \tag{4.2}$$

Here, $Ar(i)$ is the ranking value of author i. P is a set of papers written by author i. Ranking value of paper j is defined as $Pr(j)$. $n(a_j)$ is the number of authors who have written paper j.

$$Pr(i) = \sum_{j \epsilon P_i} \frac{Pr(j)}{O(p_j)} \tag{4.3}$$

We are here introducing two new algorithm and the formula we used to calculate paper rank from a data set. In the first formula (4.3) we simply used the summation of ranking value of all paper $Pr(j)$ that are citing another paper P_i division by total number of paper P_j is citing. Here $Pr(i)$ 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.

$$Pr(i) = \frac{Pr(i)}{\sum\limits_{i \epsilon P} Pr(i)} N \tag{4.4}$$

We used the above formula (4.4) to normalize the rank value we derived in formula (4.3). Here *sum* means the summation value of $Pr(i)$ and N means the total number of paper.

$$Pr(i) = \frac{Pr(i)}{\sqrt{I(p_i)}} \tag{4.5}$$

$$N = \sum_{i \epsilon P} Pr(i) \tag{4.6}$$

For our second algorithm we firstly used the formula (4.3) to calculate a value of $Pr(i)$ then we used the formula (4.5) to get average and better value of $Pr(i)$. Here *count* means the number of paper that is cited paper P_i. Again here we used the formula (4.4) to normalize $Pr(i)$ such as $\sum Pr(i) = N$.

4.3 Comparison

Rank	Papers	Ranking Values
1	P0	16.997870891488006
2	P1	0.0013306553192619387
3	P14	0.0002661343972512374
3	P13	0.0002661343972512374
3	P12	0.0002661343972512374
4	P9	4.166748559616056e-09
4	P8	4.166748559616056e-09
4	P7	4.166748559616056e-09
4	P6	4.166748559616056e-09
4	P5	4.166748559616056e-09
4	P4	4.166748559616056e-09
4	P3	4.166748559616056e-09
4	P2	4.166748559616056e-09
4	P16	4.166748559616056e-09
4	P15	4.166748559616056e-09
4	P11	4.166748559616056e-09
4	P10	4.166748559616056e-09

Table 4.2: Our Algorithm Ranking Value with summation

If we see our three table which are 4.1 , 4.2and 4.3 we can see that in pagerank algorithm $P1$ is the first ranked with the value of 0.2235 but in our algorithm paper

Rank	Papers	Ranking Values
1	P0	4.824633043707412
2	P1	1.7487871301157036
3	P14	1.1772923730721754
3	P13	1.1772923730721754
3	P12	1.1772923730721754
4	P9	0.5745585589133633
4	P8	0.5745585589133633
4	P7	0.5745585589133633
4	P6	0.5745585589133633
4	P5	0.5745585589133633
4	P4	0.5745585589133633
4	P3	0.5745585589133633
4	P2	0.5745585589133633
4	P16	0.5745585589133633
4	P15	0.5745585589133633
4	P11	0.5745585589133633
4	P10	0.5745585589133633

Table 4.3: Our Algorithm Ranking Value with square root

$P0$ is the first with 4.824633043707412 and 16.997870891488006 respectively. So the reason behind $P0$ being the top in our ranking system is, In 4.1 we can see that $P1$ is cited by many papers but those papers don't 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 are cited carries more value. So in our algorithm we are getting the value which we think should be the actual rank of the papers.

Now if we compare between our two algorithms and analysis the table and scatter diagram which are 4.2, 4.3, 4.4 and 4.3 we can see that in the algorithm with the summation, paper $P0$ is getting higher value but the other papers are getting very low value which seems almost identical and difference with other paper is very high. On the other hand, If we see the algorithm with square root here paper $P0$ is also high ranked and the differences between other papers are low which seems more acceptable to us. That is why we used the algorithm with square root for our main database [5] to calculate the rank.

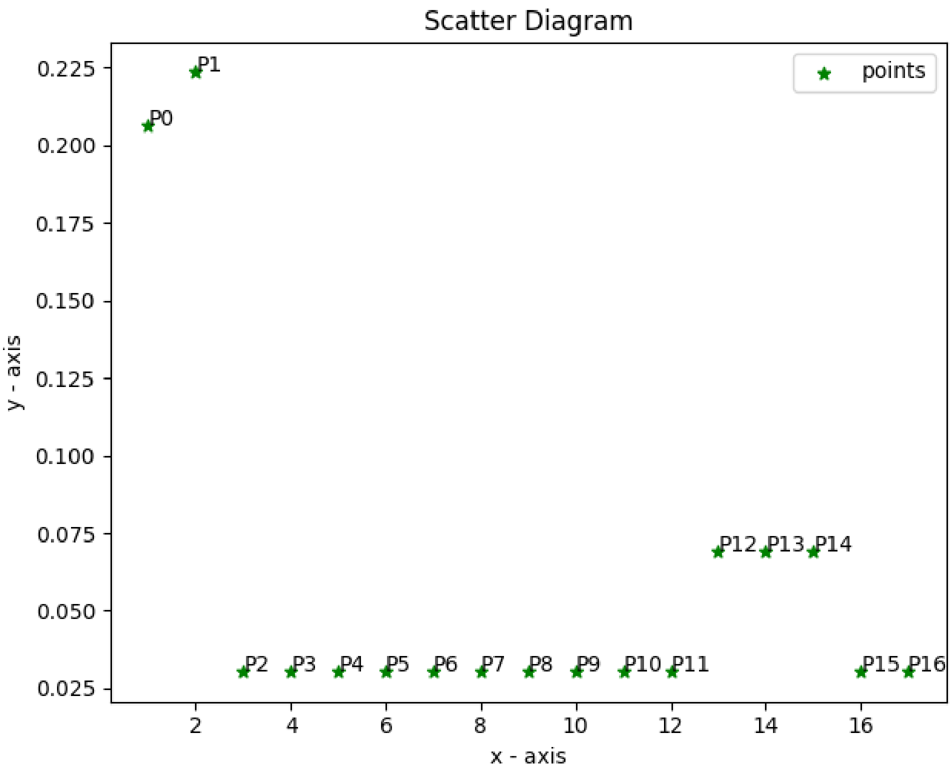


Figure 4.2: Paper Ranking(PageRank Algorithm)

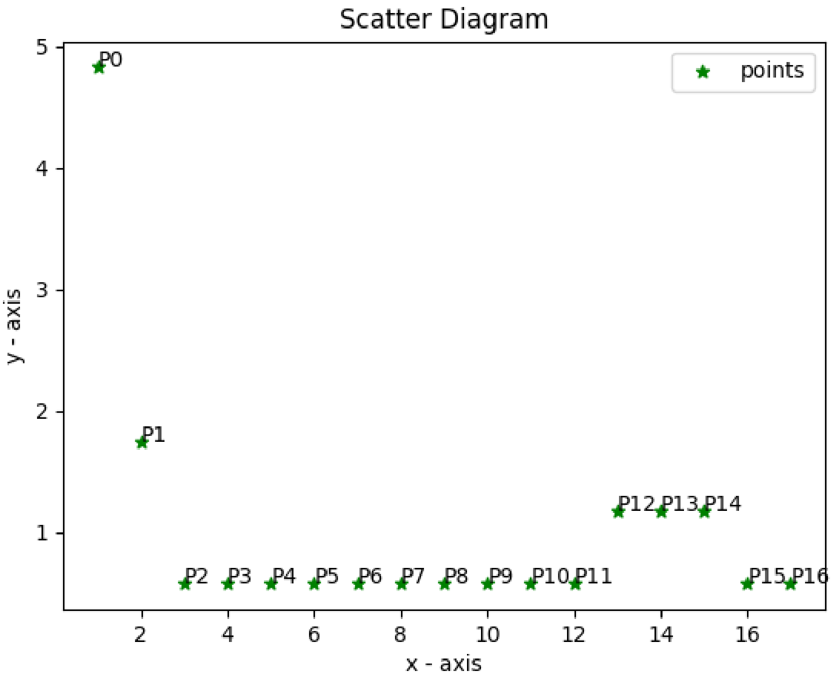


Figure 4.3: Our Algorithm Ranking Value with square root

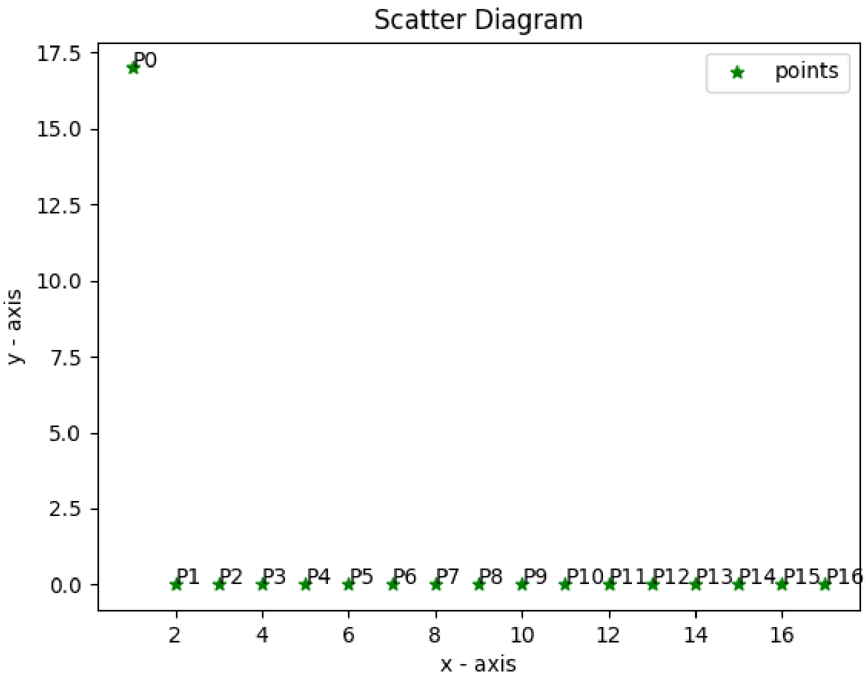


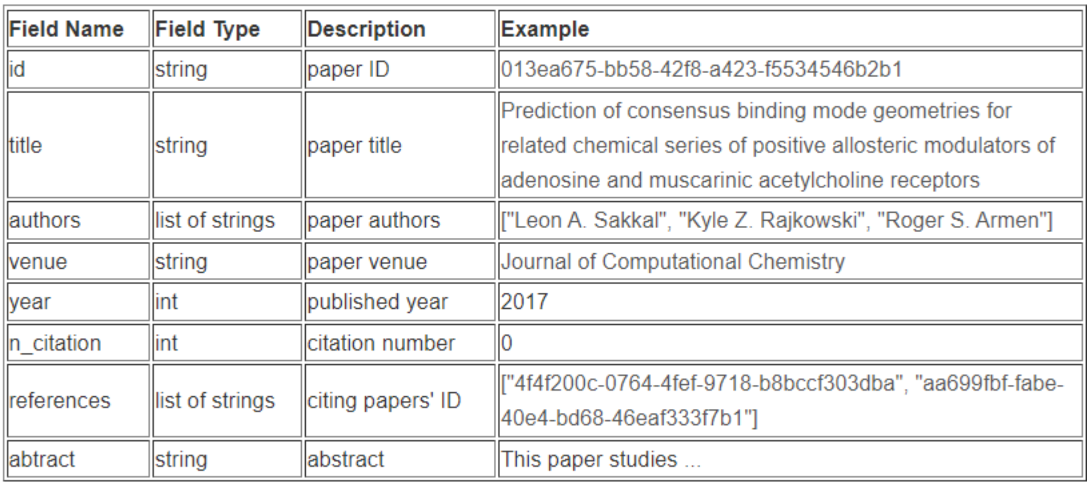
Figure 4.4: Our Algorithm Ranking Value with summation

Chapter 5

Implementation and Result Analysis

5.1 Dataset

The data set was extricated from ArnetMiner website [5]. We utilized the version 10 from 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. We extracted the dataset and used it in a way so that we can calculate the rank. In our associated information subset the total number connections of paper to author is 282525 and paper to paper connection is 634395 from (2017-10-27).



Field Name	Field Type	Description	Example
id	string	paper ID	013ea675-bb58-42f8-a423-f5534546b2b1
title	string	paper title	Prediction of consensus binding mode geometries for related chemical series of positive allosteric modulators of adenosine and muscarinic acetylcholine receptors
authors	list of strings	paper authors	["Leon A. Sakkal", "Kyle Z. Rajkowski", "Roger S. Armen"]
venue	string	paper venue	Journal of Computational Chemistry
year	int	published year	2017
n_citation	int	citation number	0
references	list of strings	citing papers' ID	["4f4f200c-0764-4fef-9718-b8bccf303dba", "aa699fbf-fabe-40e4-bd68-46eaf333f7b1"]
abtract	string	abstract	This paper studies ...

Figure 5.1: JSON Dataset Schema (Source: Tang et al. [5])

5.2 Data Training and Preprocessing

The data set was in a JSON file format. First of all, we converted the JSON file into a CSV file by the help of python's pandas library. On that csv file we had 8 columns containing header abstract (Paper's abstract), title (Paper's title), authors (Paper's authors), venue (Paper venue), year (Paper published year), n-citation (citation number of paper), id (every paper's unique id) and references (ids of references papers). From this data set of csv file, we made two different cvs files which each contained two columns. One named as `paperToAuthor.csv` which had one column with id (every paper's unique id) and authors (Paper's authors). Second one also had two columns one with id (every paper's unique id) and references (ids of references papers), file named as `paperToPaper.csv`. Then, we trained file (`paperToPaper`) that as each row has one paper and it's one reference paper. Similarly for the other file (`paperToAuthor`) we trained as each row has one paper and one author of the paper. After training the data we had 282525 row in the `paperToAuthor` file which indicates 282525 connections between the papers to author and in the other `paperToPaper` file we had 634395 rows which indicates 634395 connections between the papers to papers. Therefore, We used these two files as inputs in our algorithms.

5.3 Result Analysis

The top 10 papers ranked by page ranks from the data base we trained are listed below:

1. Kocher, Paul, Joshua Jaffe, and Benjamin Jun. "Differential power analysis." *Annual international cryptology conference.* Springer, Berlin, Heidelberg, 1999.
2. Friedman, Nir, et al. "Using Bayesian networks to analyze expression data." *Journal of computational biology* 7.3-4 (2000): 601-620.
3. Cook, Matthew. "Universality in elementary cellular automata." *Complex systems* 15.1 (2004): 1-40.
4. Kim, SungHwan, et al. "Meta-analytic support vector machine for integrating multiple omics data." *BioData mining* 10.1 (2017): 1-14.
5. Voyiatzis, Ioannis, and Costas Efstathiou. "Low Cost Boolean Function generation." *Proceedings of the 20th Pan-Hellenic Conference on Informatics.* 2016.
6. Ésik, Zoltán, and Werner Kuich. "Continuous semiring-semimodule pairs and mixed algebraic systems." *Acta Cybernetica* 23.1 (2017): 061-079.
7. Geraci, Giovanni, et al. "Operating massive MIMO in unlicensed bands for enhanced coexistence and spatial reuse." *IEEE Journal on Selected Areas in Communications* 35.6 (2017): 1282-1293.
8. Kogler, Dieter Franz, Gaston Heimeriks, and Loet Leydesdorff. "Patent portfolio analysis of cities: Statistics and maps of technological inventiveness." *European Planning Studies* 26.11 (2018): 2256-2278.
9. Meng, Wenjia, et al. "Two-bit networks for deep learning on resource-constrained embedded devices." *arXiv preprint arXiv:1701.00485* (2017).
10. Rossi, Giovanni. "Near-Boolean Optimization: A Continuous Approach to Set Packing and Partitioning." *International Conference on Pattern Recognition Applications and Methods.* Springer, Cham, 2016.

The top 10 papers ranked by Our Algorithm with Square-Root and summation from the data base we trained are listed below:

1. Saxena, Vidit, et al. “On the achievable coverage and uplink capacity of machine-type communications (MTC) in LTE release 13.” *2016 IEEE 84th Vehicular Technology Conference (VTC-Fall).* IEEE, 2016

2. Saxena, Vidit, et al. “Reducing the modem complexity and achieving deep coverage in lte for machine-type communications.” *2016 IEEE Global Communications Conference (GLOBECOM).* IEEE, 2016.

3. Haddad, Ziad S., et al. “Derived observations from frequently sampled microwave measurements of precipitation—Part I: Relations to atmospheric thermodynamics.” *IEEE Transactions on Geoscience and Remote Sensing* 55.6 (2017): 3441-3453.

4. Sy, Ousmane O., et al. “Derived observations from frequently sampled microwave measurements of precipitation. Part II: Sensitivity to atmospheric variables and instrument parameters.” *IEEE Transactions on Geoscience and Remote Sensing* 55.5 (2017): 2898-2912.

5. Bhardwaj, Kartikeya, and Radu Marculescu. “K-hop learning: a network-based feature extraction for improved river flow prediction.” *Proceedings of the 3rd International Workshop on Cyber-Physical Systems for Smart Water Networks.* 2017.

6. Ding, Ruizhou, et al. “Enhancing precipitation models by capturing multivariate and multiscale climate dynamics.” *Proceedings of the 3rd International Workshop on Cyber-Physical Systems for Smart Water Networks.* 2017.

7. Moreno-Schneider, Julian, Peter Bourgonje, and Georg Rehm. “Towards user interfaces for semantic storytelling.” *International Conference on Human Interface and the Management of Information.* Springer, Cham, 2017.

8. Rehm, Georg, et al. “Designing user interfaces for curation technologies.” *International Conference on Human Interface and the Management of Information.* Springer, Cham, 2017.

9. Seneff, Stephanie, et al. “Development and preliminary evaluation of the MIT ATIS system.” *Speech and Natural Language: Proceedings of a Workshop Held at Pacific Grove, California, February 19-22, 1991.* 1991.

10. Theakston, Anna, and Elena Lieven. “Multiunit sequences in first language acquisition.” *Topics in Cognitive Science* 9.3 (2017): 588-603.

If we see here, in page rank the most cited papers are on top. On the other hand, in our algorithms are giving the top rank comparatively to the reputed author’s papers which does not have huge citation as these papers are not easy to cite.

Chapter 6

Conclusion and Future Work

6.1 Fututre Work

In this research paper, we focused on coming up with a unique method paper ranking can be done by its quality, not by the quantity of its citation. But still this is also true that no algorithm or method is perfect for ranking scholarly articles. With the help of linear equations, we approached a system where we can score the author's contribution in scientific research more effectively. However, we faced some downfall while studying this topic. We started with the DBLP dataset for authors' metrics. However, due to the database being too large for the computing resources available to us, we decided to use a smaller dataset by ArnetMiner. So for further research we might need a Computer with better configuration, or use a more resource efficient approach for handling big dataset. Next if we want to run our system accurately and efficiently we might need to develop an easy yet reliable algorithm for the system. So that our work can be easier to demonstrate and retrieve better results from the algorithm. Our results look promising for eccentric authors with lowly cited but important papers. However, there is more room for improvements as a proper study needed to check the quality of ranking computed by our algorithm. In this thesis we have introduced a new approach for computing ranking scores of scientific papers. For future work, when determining the score of a paper, instead of only taking a look at the score of the papers that cited it, we should consider the score of the authors of the said papers. This may allow us to more accurately distinguish lowly cited papers that have been predominantly cited by the experts of the specific field.

6.2 Conclusion

Our goal is to build a better and more efficient ranking system for authors which can give the actual value and credit to that author or researcher for his or her work. Currently we are working with the flaws of current author ranking algorithms and trying to resolve the flaws. If we become successful to build a new one resolving the current drawbacks then there will be a dramatic change on the current ranking system. So many underrated researchers will be on the top and maybe sharing a mutual ranking with many well known researchers of current time. It will also improve the categorization for websites for each search engine.

Bibliography

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

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

[3] 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.

[4] 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.

[5] 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.

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

[7] 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.

[8] T. Amjad, A. Daud, and A. Akram, "Mutual influence based ranking of authors," *Mehran University Research Journal of Engineering & Technology*, vol. 34, no. S1, pp. 103–112, 2015.

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

[10] 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.

[11] 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.

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

[13] 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.

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

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

[16] 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.

[17] 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.

[18] M. Kosmulski, “Nobel laureates are not hot,” *Scientometrics*, vol. 123, no. 1, pp. 487–495, 2020.

[19] A. Daud, S. Arabia, T. Amjad, H. Dawood, and S. H. Chauhdary, “Topic sensitive ranking of authors,”