Deanship of Graduate Studies
Al-Quds University

Time Aware Ranking Algorithm for Scientific Publications

Moath Dawood Mahmoud Abudayeh

M.Sc. Thesis

Jerusalem- Palestine

1443/2021

Time Aware Ranking Algorithm for Scientific Publications

Prepared by:

Moath Dawood Mahmoud Abudayeh

B.Sc. Computer Science Al-Quds University Palestine

Supervisor: Dr. Badie Sartawi

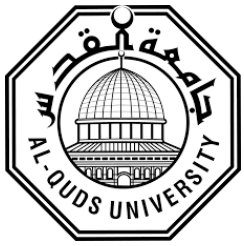
A thesis submitted in partial fulfillment of requirements for degree of Master of Computer Science at Al-Quds University

1443/2021

Al-Quds University

Deanship of Graduate Studies

Computer Science Department



Thesis Approval

Time Aware Ranking Algorithm for Scientific Publications

Prepared By: Moath Dawood Mahmoud Abudayeh

Registration No: 21710749

Supervisor: Dr. Badie Sartawi

Master thesis submitted and accepted, Date: 22/8/2021

The name and signatures of the examining committee members are as follows:

1. Head of Committee: Dr. Badie Sartawi Signature 
2. Internal Examiner: Dr. Saeed Salah Signature 
3. External Examiner: Dr. Dirar Eleyan Signature 

Jerusalem- Palestine
1443/2021

Dedication

I dedicate this thesis work to my parents, my wife and my daughter. Thank you for your continuous support and love.

Moath Dawood Mahmoud Abudayeh

Declaration

I certify that this thesis submitted for the degree of Master is the result of my own research, except where otherwise acknowledged. I also certify that this study or any part of it has not been submitted for a higher degree to any other university or institution.

Signature, 

Moath Dawood Mahmoud Abudayeh

Date: August 22, 2021

Acknowledgments

Many thanks for my advisor Dr. Badie Sartawi for all the guidance, patience and support during this work.

Great thanks for Al-Quds University, especially the Computer Science Department and its staff.

Abstract

After the enormous evolution of the Web, the emergence of digital information resources and the universities and libraries have allowed access to their contents via the Internet, a large amount of data became available to users to conduct searches and queries. However, this huge content made it difficult to quickly access the required data using traditional search methods. These methods depend on matching keywords or determining the extent of relevance. As a result, the need for ranking algorithms emerged in information retrieval systems.

The terms ranking and evaluation are related because the ranking process is based on certain evaluation criteria and indicators. One of the most widely used algorithms for ranking scientific publications is the PageRank algorithm. It evaluates publications using popularity metrics based on the linking analysis approach. However, this algorithm was designed mainly to rank Web pages rather than scientific publications. Therefore, due to the different nature of Web networks and citation networks, it resulted in unfair rankings and bias in favor of old publications. The reason for this bias is in its heavy reliance on the number of citations as an indicator of popularity.

This study focuses on solving the problem of bias in favor to old publications by introducing a new indicator called Citation Change Rate and integrating it with PageRank algorithm. Time information such as publication date and citation occurrence time are used along with citation data in the ranking process in order to produce time aware rankings.

The proposed ranking method was tested on a dataset of scientific papers in the field of medical physics. They were published in the Dimensions database from 2005 to 2017. The results showed that the proposed ranking method took into account the characteristics and dynamic nature of the publishing network. This resulted in fair rankings for publications of different ages, and less bias against recent publications. The results have shown that 13 papers published in the last four years based on the new ranking scores, are now among the top 100 ranked papers of this dataset. In addition, there were no radical changes or unreasonable jumps in the ranking process. Therefore, the correlation rate between the results of the proposed ranking method and the original PageRank algorithm was 90% based on the Spearman Correlation Coefficient. This is an indication of the quality and accuracy of the results.

خوارزمية تصنيف مدركة للوقت لتصنيف المنشورات العلمية

إعداد: معاذ داود محمود أبوديه

إشراف: د. بديع السرطاوي

ملخص:

بعد التطور الهائل لشبكة الويب وظهور العديد من مصادر المعلومات الرقمية وسماح الجامعات والمكتبات بالوصول إلى محتوياتها عبر الإنترنت، أصبحت كمية كبيرة من البيانات متاحة للباحثين والطلاب لإجراء عمليات البحث والاستعلام. ولكن في الوقت نفسه يصعب الوصول الى هذا الكم الهائل من البيانات بسرعة تلبي حاجة المستخدم وذلك لاستخدام طرق البحث التقليدية التي تعتمد على مطابقة الكلمات الرئيسية أو تحديد مدى الصلة، وهنا ظهرت الحاجة إلى خوارزميات الترتيب لدعم أنظمة استرجاع المعلومات.

التصنيف والتقييم مصطلحان مرتبطان ببعضهما لأن عملية التصنيف تعتمد بالأساس على معايير ومؤشرات تقييم معينة لإعطاء الدرجات إحدى الخوارزميات الأكثر استخدامًا لتصنيف المنشورات العلمية هي خوارزمية PageRank، تجري هذه الخوارزمية عملية التصنيف عن طريق تقييم المنشورات باستخدام مقاييس الشعبية و نهج تحليل الروابط، ولكن تم تصميم هذه الخوارزمية بشكل أساسي لتصنيف صفحات الويب وليس المنشورات العلمية، لذلك نظرًا للطبيعة المختلفة لشبكات الويب وشبكات الاقتباس، فإنها تؤدي إلى تصنيفات غير عادلة ومنحازة للمنشورات القديمة، ويأتي سبب هذا التحيز من اعتمادها الشديد على عدد الاستشهادات كمؤشر على جودة وشعبية المنشورات العلمية.

تركز هذه الدراسة على حل مشكلة التحيز للمنشورات القديمة من خلال تقديم مؤشر جديد يسمى معدل التغير السنوي في الاقتباس ودمجه مع خوارزمية PageRank، حيث يتم استخدام معلومات الوقت مثل تاريخ النشر ووقت حدوث الاقتباس جنبًا إلى جنب مع بيانات الاقتباس في عملية التصنيف من أجل إنتاج تصنيفات مدركة للوقت. أظهرت النتائج أن طريقة التصنيف المقترحة تأخذ في الاعتبار الخصائص والطبيعة الديناميكية لشبكة النشر، كما انها تنتج تصنيفا أكثر عدلاً للمنشورات من مختلف الأعمار، وتقلل التحيز ضد المنشورات الحديثة بدرجة كبيرة.

تم اختبار طريقة التصنيف المقترحة على مجموعة اوراق علمية في مجال الفيزياء الطبية منشورة في قاعدة بيانات Dimensions، من سنة ٢٠٠٥ وحتى سنة ٢٠١٧، واظهرت النتائج ان ١٣ ورقة منشورة خلال السنوات الأربعة الاخيرة تحسن تصنيفها لتحتل مركزا ضمن أفضل ١٠٠ ورقة من مجموعة الاوراق. ايضا لم يحدث تغييرات جذرية او قفزات غير منطقية في عملية التصنيف بحيث كانت نسبة الارتباط بين نتائج طريقة التصنيف المقترحة وخوارزمية PageRank الاصلية ٩٠% بناء على معامل ارتباط سبيرمان.

Table of Contents

List of Figures

List of Tables

List of Equations

List of Appendices

List of Abbreviations

IR: Information Retrieval.

API: Application Programming Interface.

DSL: Dimensions Search Language.

IF: Impact Factor.

PR: PageRank Algorithm

HITS: Hyperlink-Induced Topic Search

MPR: Modified PageRank

AM: Adjacency Matrix

RAM: Retained Adjacency Matrix

NR: NewRank Algorithm

WC: Weighted Citation

ACR: Annual Citation Change Rate

$\Delta \boldsymbol{C}$: The amount of change in citation over the years.

$\Delta \boldsymbol{T}$: change in time.

Spearman's ρ: Spearman's rank correlation coefficient

Chapter 1

Introduction

1.1 Introduction

Hundreds of scientific publications are published daily in all fields of research. Especially with the increase in publishing methods and the evolution of the Web. Despite the great benefits of the abundance of data, this makes it difficult for the new researchers and regular users to search and find the required information. Therefore, retrieving information by relying on the traditional search methods which depend on matching keywords or determining the extent of relevance, is not sufficient to meet the user's desires. This is mainly because the number of returned records will be huge, and it may take plenty of time to find what is required among these results. Hence, the emergence of the need to develop information retrieval systems that take into account the mechanism for arranging these results by using ranking algorithms.

Digital libraries are considered the most important information sources currently available. They include a huge number of academic papers and scientific publications. It is also called bibliometric databases and link-based databases. However, as mentioned earlier, in order to facilitate the search process and enable researchers and students to reach their needs, these publications must be ranked based on their impact and importance. The concept of importance is broad as different methods can be used to measure the importance of the paper. But in general, the paper popularity is a strong indicator of the extent of its impact and importance in the scientific community (Singh et al., 2011).

Citations are often used to measure paper popularity. It is the number of times a particular paper received citations from other publications. Citations are a rich source of data that can be analyzed in various ways to indicate the importance of a scientific publication or journal (H. Cavaillon and G. Gak 2009). They are considered one of the Bibliometrics indicators, which are a term given by the scientific community to a set of indicators and measures that are used to refer to the popularity and quality of a scientific publication. Bibliometrics are also used by institutions to evaluate researchers. In addition, they are used by funding agencies to distribute funds and so many other ranking purposes. Mathematical and statistical methods are used to calculate these

indicators. The most used approach in citation analysis is the link-based analysis. Both the PageRank (PR) and Hyperlink-Induced Topic Search (HITS) algorithms are the most popular algorithms that use this approach (Joshi, 2014).

1.2 Problem Statement

As mentioned previously, the PageRank algorithm is the most widely used algorithm for ranking scientific publications based on citation analysis. However, this algorithm was primarily designed to deal with Web pages rather than scientific publications. Publication networks differ in their characteristics from Webpage networks, as the PageRank algorithm treats them as a static networks and does not take into account their dynamic nature. Publications are represented by nodes in the network, while the edges represent citations. This network is also called a citation network and it changes with time as there are constantly new publications and new citations.

There are many citation-based metrics such as citation count, which is the number of citations that a particular paper has received. The PageRank algorithm is highly dependent on the citation count. It is good that PageRank algorithm implicitly takes into account the importance of the citing paper. As this means that it assigns weights to citations instead of treating them equally. However, this algorithm still depends on the number of citations. Therefore, the old papers that took enough time to collect a large number of citations, even if these citations are of little importance will get a high score. This causes the problem of bias in favor of the old papers.

Several solutions have been proposed to solve this problem, such as CiteRank (Walker et al., 2007) and FutureRank (Sayyadi and Getoor, 2009). Some of these algorithms rely on time information, such as publication date to reduce bias by raising new publications scores. However, recent publications do not always deserve high scores, so relying only on publication date to solve this problem may transfer the bias in favor of recent publications. In some solutions, recent publications are awarded higher rankings by using side information. Among the information included are author data, journal or conference data and paper metadata. This information is very valuable and helps to produce more accurate assessments. However, it does not solve PR algorithm drawback because it is still dependent on the citation count. Citation data and citation networks must be analyzed in other ways, not just counting their totals.

1.3 Objectives of the Study

1. The main objective of this thesis is to improve the PageRank algorithm to be more convenient to ranking scientific publications.
2. Helping researchers to know the extent of spread and reputation of their publications.
3. Helping funding agencies to determine the impact of any funded research, as well as identifying future funding trends.
4. Supporting research databases with an appropriate ranking mechanism that ensures queries results are ranked in a fair manner

1.4 Motivation

1. Using the citation count as the only indicator of the relevance and popularity of scholarly publications might lead to unfair rankings.
2. Citation data can be used to generate new indicators other than the citation count that might produce unbiased rankings.
3. Using time information such as publication date in order to influence the ranking process arbitrarily with the aim of reducing bias, which leads to creating new problems such as imparting bias in favor of recent publications. This is because they will get high scores even if they are not worthy, but only because the date of publication is recent.
4. Generalizing assumptions such as an author with a good reputation always produces high quality papers. Or that recent papers are more valuable because the researcher always begins his/her research by reading recent papers, in order to raise the scores of recent papers might not give correct results. This is due to publishing and citing behavior change over time.

1.5 Thesis Contributions

1. Proposing a new time-aware ranking method that takes into account the dynamic nature of the publishing network for ranking scientific publications by modifying the PageRank algorithm as follows:

1.1 Presenting a new indicator, called the annual citation change rate. Which uses the time information as well as the citation network in its calculation.

1.2 Producing fair rankings by reducing the bias to old publications without transferring the bias in favor of recent publications.

1.6 Research Question and Working Hypotheses

The main research question that this study tried to answer is:
How can we use citation data and time information to enhance PageRank algorithm to produce time aware ranking for scientific publications?

The working hypotheses in this study are:

1. Relying only on citation count to rank scientific papers will not give correct results.
2. Using the citation change rate will reduce bias to old papers and give fair results.

1.7 Research Methodology

After reviewing the most important quantitative ranking methods of scientific publications, and studying in depth the PageRank algorithm and time-aware ranking algorithms. The Empirical approach was used in order to test our hypotheses and answer the research question, by modify the PageRank algorithm and applying it to a real dataset using the quantitative research methods.

1. **State-of-the-art:** It summarizes, classifies, and compares different ranking mechanisms, and prepares a comprehensive review of the PageRank algorithm and its modified versions. It also identifies the challenges and gaps in these versions.
2. **Dataset Exploration and Processing:** Determining the necessary data using Dimensions web app and then collecting it from Dimensions database using Dimensions API, preprocessing data using Bibxcel tool by removing the unwanted and missing fields and keeping the necessary ones, collect the bibliography data for all citing papers by using Dimcli language in order to build the citation network.
3. **Building the model:** Building the citation network using Gephi tool which is an open-source network visualization and analysis software, using Python language to extract the values needed to calculate the annual citation change rate such as the paper age, the time each citation occurred, and number of citations in each year. Then calculating the new rank score based on the new equation.

4. **Applying the model:** Applying the PageRank algorithm on the citation network to get the PR score for each paper, and calculating the new rank score based on the annual citation change rate, the results may be affected positively if the change rate is high or negatively if it is low or is a negative value.
5. **Validating the results:** In order to evaluate the results and ensure the achievement of the study objectives, a set of measures was used. In addition, a comparative analysis was conducted between the results of the original algorithm and the modified one.

1.8 Thesis Structure

The remaining parts of this thesis are structured as follows:

Chapter 2 (Background and Literature Review): This chapter contains an overview of the scientific research impact, its types and its evaluation purposes. It also reviews ranking models used in information retrieval systems and bibliometric databases. In addition, it presents literature review about studies related to the scientific publications ranking and PageRank algorithm.

Chapter 3 (Theoretical Framework): This chapter introduces the main concepts, terms, theories and approaches related to ranking problems. In addition to a brief explanation of the PR algorithm and a review of the enhanced versions. Then a discussion of the proposed ranking method.

Chapter 4 (Experiments): This chapter explains the implementation steps of the proposed Ranking algorithm, starting with collecting and preparing data, building the citation network, calculating the original PR values and presenting the proposed ranking method.

Chapter 5 (Discussion of Results and Validation): This chapter contains a comparative analysis between the results of the original PR algorithm and the modified one. In addition to discussing and validating the results.

Chapter 6 (Conclusion and Future Work): This chapter presents a summary of the study results, and suggestion for future works.

Chapter 2

Background and Literature Review

2.1 Background Knowledge

This section provides an overview of the concepts and key terms related to this study. It starts with the research impact in general, its types, and ranking process of each type. It also reviews ranking models used in information retrieval systems and bibliometric databases. In addition to that it introduces the tools and theories used in this study such as Graph theory and Citation metrices.

- **Research Impact**

When conducting research, significant impacts are expected. The primary goal of research is to generate knowledge that will be beneficial for both the academic field and society as a whole. Scientific research impact is a broad term. It is difficult to find a comprehensive definition that includes all areas and aspects of research impact. In general, we can distinguish between two main categories of impact. The first one is the academic impact that is considered the contribution of research within academia. The other is the external impact on society in various fields such as economic, political, social and health fields.

- **Major Impact Categories**

1. Direct Impact (Academic Impact):

It is considered the first type of impact to be observed, because it does not require a long time to appear. It is easy to measure because it mainly depends on quantitative and Bibliometrics indicators such as the number of publications, the citation count and the number of views or downloads. The most important frameworks that are concerned with this category of impact are Research Excellence Framework (REF) (Hubble, 2015), Research Assessment Exercise (RAE) (Barker, 2007), and Research Utilization Ladder. Moreover, this type of impact is an important indicator for measuring the spread of research, producing new knowledge, indicating the reputation of the researcher, and it is considered as a reference for researchers.

2. External Impact (Impact Beyond Academia):

This type of impact is more difficult to track and takes a long time to appear. However, it is a more accurate indication of the research impact on society, economy, policy making, general culture and quality of life. Measurement of this type of impact usually requires qualitative indicators. Moreover, most methodological frameworks use the narrative approach such as case study to describe this type of effect. It depends on expert review instead of quantitative measures to analyze these aspects.

- **Citation Indexing and Bibliometric Databases**

Citations are one of the most widely used bibliometric indicators for evaluation, assessment and ranking purposes. They are used to measure productivity, prevalence, reputation and relevance for the research and researcher as well as the institution itself. Citation index is a type of bibliographic index. It is used in bibliometric databases such as Web of Science (Garfield, 2016) and Scopus (Elsevier B.V., 2020) so that the user is able to access related documents in an easy and smooth way (Bienert et al., 2015).

- **Ranking Process**

The process of ranking scientific papers varies according to its purpose. In general, comprehensive evaluations that aim to study the impact of scientific research on society must be based on qualitative measures and qualitative analysis of data. The data is usually collected by narrative methods and case studies. It is then analyzed by peer review. This is the most reliable way to assess the external impact of research (Penfield et al., 2014).

On the other hand, ranking process is one of the most important pillars of information retrieval systems. It ranks the query results and identifies the most relevant papers to meet the user's needs. In this case, statistical and mathematical methods are usually used, depending on some numerical and bibliometric measures such as citation analysis (Kelly and Sugimoto, 2013).

- **Ranking Approaches**

The main function of the ranking model is to assign scores to documents, Web pages, scientific papers, or whatever the results of the query. It represents the amount of correlation and similarity

between results and queries. Ranking algorithms can be divided into the following approaches.

1. Content Based Ranking:

In this approach, papers are ranked based on content and keywords, so that results are analyzed individually and compared to keywords. When entering a query by the user, the root words are specified. Then a dictionary is created from the words synonymous with each root. The keywords on the results page are compared to the dictionary and then the weight of each word is determined based on the match found. The final step is to summarize all weights for keywords to calculate the overall relevancy of a given link versus a user's query (Arora and Govilkar, 2016).

2. Usage Based Ranking:

The ranking algorithms that rely on usage data are usually recommendation algorithms. It mainly aims to anticipate and provide the next pages to the user based on the usage data represented by his/her current visits, and the movement patterns between pages by similar users. The suitability of the Web page or paper with the user's options is determined by the number of times it is viewed or selected. However, relying on this indicator independently to make recommendations is not accurate to indicate real importance. There are other indicators that can be used to make more accurate recommendations such as time spent reading, saving or printing, frequency of downloads or addition of pages to a bookmark. Combining these indicators is the best and gives suitable recommendations (Arora and Govilkar, 2016).

3. Linking Based Ranking:

Link structure algorithms present the documents in a structured manner. The goal is to give them scores according to their relevance and importance through correlation analysis. Using this approach, the quality of ranking can be significantly improved by exploiting links between pages in search engines or links between papers to improve the ranking quality of scientific publications. Any paper that has received so many citations must have something to express. These algorithms calculate scores offline, and do not wait for the query to be received from the user. This improves the search process and speeds up the retrieval of results. Link structure algorithms calculate the popularity and spread of a Web page or a scientific publication by creating graphs consisting of nodes. The nodes represent the entities to be ranked and the links represent relationships between these entities, Figure (2.1) shows the linking structure of web

pages. The most popular algorithm belonging to this category is the PR algorithm which is used to rank Web pages (Arora and Govilkar, 2016).

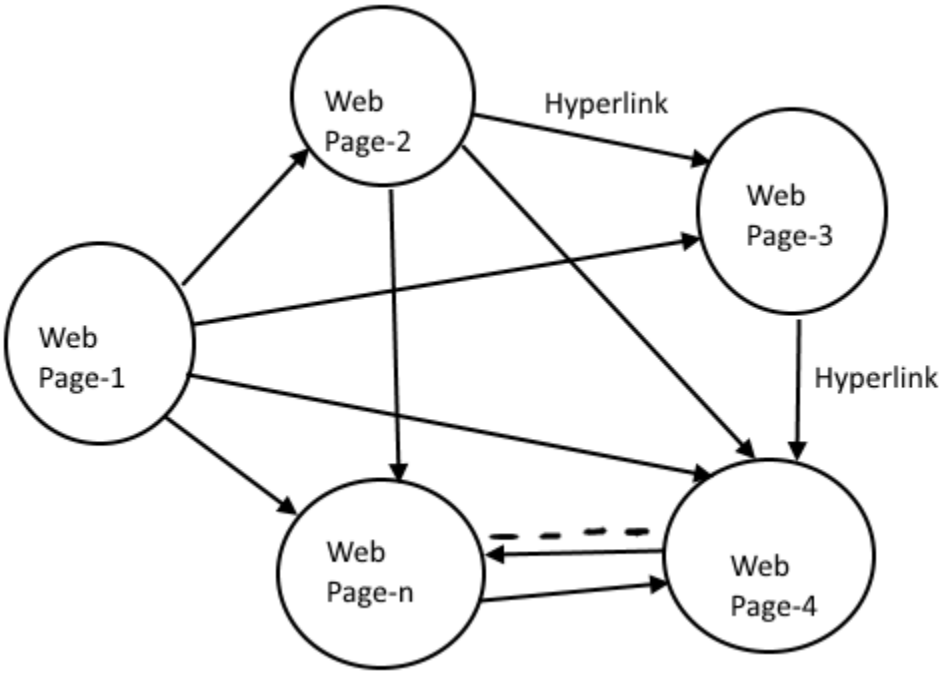


Figure 2.1: The linking structure of web pages.
Source (Alom, 2016).

- **Ranking Algorithms**

As mentioned earlier, the ranking mechanism depends on the purpose of ranking. In information retrieval systems, the main goal is to deal with the query rank problem. Most algorithms used in IR systems are divided into three types: Similarity based models, probabilistic models and link-based models (Liu, 2011). Figure (2.2) shows a classification of IR ranking models.

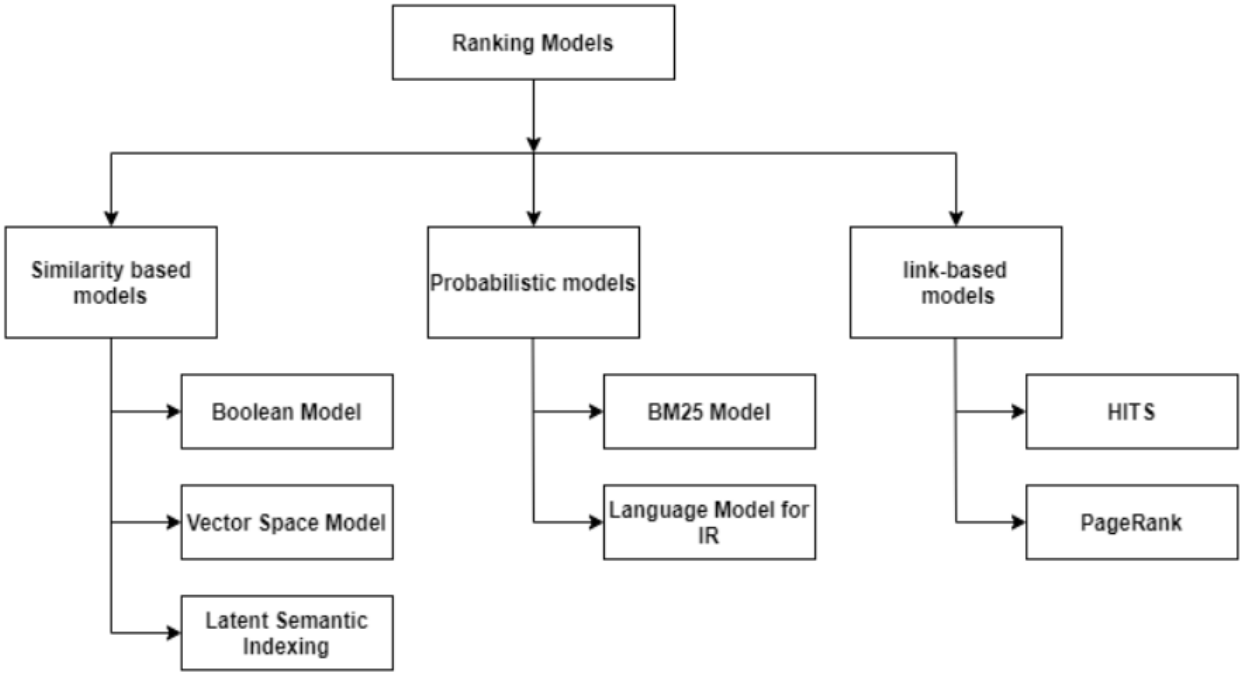


Figure 2.2: A classification of information retrieval ranking mechanisms.
Source: (B.Yates and R.Neto 2011)**.**

1. Similarity based Models

Generally, these algorithms rely on measuring the amount of similarity between a query and the documents. It does so by counting the number of similar words and the repetition of key terms along with their locations. Next the results of the query are given scores to be ranked in descending order, starting from the most similar document (Liu, 2011). Examples of these models include:

- **Boolean model:** is one of the most famous similarity-based models. It is a query model based on Boolean algebra that deals with the query results separately. This model forms an index of words or phrases for each document to compare it with the query (Hiemstra, 2001)
- **Vector space model**: is an algebraic model based on the concept of similarity. This model assumes that the similarity between the query and the document represents the amount of relevance. Initially the bag of words model is used to represent both the query and the document. For each set of documents, a set of terms is defined and weighted using Term Frequency- Inverse Document Frequency (TF-IDF) method. Then the documents and queries are represented as vectors. The similarities can be measured using the inner product of two vectors (Salton, Wong and Yang, 1975).

- **Latent Semantic Indexing:** It is a natural language processing method that analyzes the relationships between documents by analyzing the terms they contain. The basic principle here is that words and terms that appear in similar texts and contexts have similar meanings. Documents and words are represented in the form of a matrix in which the columns represent the documents while the rows represent the words contained in each document. Then a technique called singular value decomposition (SVD) is used to reduce the number of rows and discover patterns in relationships (Deerwester et al., 1988).

2. Probabilistic models

Probabilistic models estimate the likelihood that a given document is related to the requested query. The user requests information by entering a query, which is then translated into query representations. Moreover, documents are converted into document representations. This model assumes that the probability of relevancy depends on these two representations. In addition, it assumes that a partial set of documents are preferred by the user and considers them the most appropriate results for the entered query (Manning et al., 2009). Examples of these models include:

- **BM25 model:** It is a probabilistic ranking model also called Okapi. The ranking process for a collection of documents is done based on the terms of query that are present in the document without considering the interrelationship between the query terms within the document. Actually, it is not a single function, but a whole group of scoring functions, with different parameters and components. It is usually used by IR systems to classify documents based on their relevance for a particular query (Robertson et al., 1995).
- **The Language Model of IR (LMIR):** is an application of information retrieval based on the statistical language model. For each document, a linguistic model is created, and then the probability of generating the query is estimated according to each model, and based on these probabilities, the documents are ranked. As this model assumes that the user who generated the query has prior knowledge of the terms that may be present in the documents that will meet the needs, and therefore it is assumed that the query will distinguish the required document from others in the group (Ponte and Croft, 1998).

3. Link-based models

These algorithms are based on the link analysis approach. Documents are represented by a graph that contains a set of nodes and edges that represent relationships. These algorithms are independent of the query and operate in offline mode. Therefore, the document is scored based on its importance within the document set. The PR algorithm (Page et al., 1999) is the most popular algorithm that belonging to this category, where citations represent the relationships (links) between the nodes. Another example is the HITS algorithm (Kleinberg, 1999) which gives each paper two scores, the first one called the authority score which estimates the content value of a given paper, and the second one called the hub score which estimates the value of its links to other papers. The link analysis approach especially the PR algorithm will be discussed in more detail in Chapter 3.

- **Citations**

Citations are an important source for research, researchers, educational institutions and scholarly journals. Citations are used as an indicator of the research importance and the quality of its outputs (Moed, 2006). Through citations, the authors can trace the development stages of ideas in their research, to verify authenticity, originality, accuracy and then measure impact, relevance, diffusion and reputation. Citations also preserve the intellectual rights of the author who developed the idea and facilitate the process of loaning intellectual credits. (Shah and Mahmood, 2017).

- **Dimensions API**

Dimensions is an indexing database developed by Digital Science (Digital Science, no date). It is designed with the aim of providing a different view of research and its information. It is an interconnected research system with more than 100 million records and millions of links between them (Mouratidis, 2019). Dimensions provide data related to research in various forms and are stored in separate repositories such as papers, scientific articles, research metadata, books, grants, data sets and patents as shown in Figure (2.3).

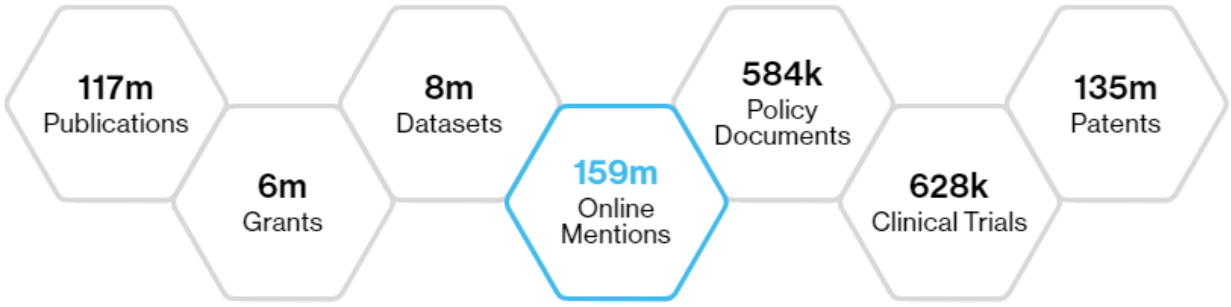


Figure 2.3: Dimensions Data Repositories.
Source (Dimensions, 2020).

One of Dimensions' priorities is to provide research data in the best and most useful way for the scientific community. Therefore, Dimensions provide a set of API services that meet the user's needs. Through it, data extraction can be carried out in various forms that enable the user to use it for analysis, visualizations and complex operations. Dimensions also provide the ability to analyze references and calculate some metrics that are at the article level. Using analytical API, we can access programmatically to these metrics. (Mori and Taylor, 2018). Table 2.1 contains some of these metrics.

Table 2.1: An article-level Metrics.

Metric	Description
Times_cited	This indicator shows the citation count for a particular publication that received citations from other publications indexed within Dimensions.
Recent_citations	While the times_cited refers to the citations that occurred in all the years, recent_citations refer to those that have occurred in the last two years.
Relative_citation ratio	It is an indicator to measure the relative performance of citation for a specific publication when compared to other publications in the same research field.
Field_citation ratio	It is an indicator to measure the relative performance of citation for a specific publication when compared to other publications with the same age and in the same research field.

- **Dimensions Search Language (DSL)**

Dimensions database has its own search language called DSL. It allows users to write programming expressions called queries to obtain data and return it from the Dimensions database. Dimcli (Dimcli documentation n.d.) is a command line tool that aims to facilitate learning of the DSL language. Dimcli also is a Python library (Python client). Through it, DSL queries can be created interactively (Hook, et al., 2018).

2.2 Literature Review

2.2.1. Introduction:

For many years there have been many attempts and studies related to the ranking of scientific publications. It has resulted in a wide range of approaches, methods, and algorithms where each method has its own ranking mechanism. Some methods designed for special evaluation purposes, such as accountability purposes, advocacy, higher education institutions overview. Meanwhile, other methods were designed for learning purposes. There are also evaluation methods concerned with studying the impact. First, there is the societal impact in various fields such as economic, political, and health. Second, the academic impact which is considered the contribution of research within academia in addition to the reputation, spread and outreach (Penfield et al., 2014).

our study is concerned with ranking based on academic impact, which is based on statistical, mathematical and bibliometric methods to measure it. This section also provides an overview of these ranking methods. In particular, the PR algorithm and its improved versions. In addition to an overview of its related problems, specifically the problem of bias towards old publications. Afterwards, we present the modified algorithms that provide solutions to reduce the bias by producing time-aware rankings.

The academic impact based ranking algorithms are depend on quantitative metrics as a measure of impact. For example, the number of publications, the number of views or downloads and the most used ones the citation counts. Garfield (Garfield, 1972) made the first effort when he proposed a metric, called an Impact Factor (IF) which is measured by counting the number of citations for a publication in a given scientific journal in the last two years. It is used to measure

the prestige of scientific journals and their ranking among the journals. The measure is done by counting the number of times the articles published in this journal are cited. After that, several studies were conducted to enhance the IF. Garfield used the same idea on the authors network to rank authors (Garfteld, 1984). Also, Narin and Pinski (Pinski and Narin, 1976) introduced a new modification by giving the citations different weights. They assumed that citations from a more important scientific journal have more value when calculating the impact of other journals. They then developed the cross-citation matrix based on this idea, which is a new method to calculate the impact of journals.

Later, the most popular algorithm was introduced, the PR algorithm (S.Brin and l.Page., 1998). It is mainly provided for the evaluation and ranking of Web pages. This algorithm simulates the user behavior in browsing and navigating between Web pages. The PageRank scores are calculated using a mathematical equation based on the graphic representation of Web pages. In simple terms, this model tracks the links between pages so that the rank value of a Web page depends on the number of pages linked to it. Its rank value also increases when the value of these pages is high. After that, the PR algorithm was applied to many other applications. Among of the applications was measuring the impact of authors by applying it on the authorship network (Liu et al, 2005). Moreover, it has also been applied to the citation network for the purposes of ranking scientific publications by (Chen et al., 2007) and (Ma et al., 2008).

2.2.2. Integrating the PageRank algorithm with other models:

Several models have combined PageRank (PR) and Hyperlink-Induced Topic Search (HITS) algorithms, which is another linking based ranking algorithm. The basic idea behind the HITS algorithm is to divide the network nodes into two types, hubs and authorities. The paper is a hub when it directly points to other related papers, while the paper is an authority when it points to a group of hubs (Kleinberg, 1999). One of these models is the PaperRank, which is an extension of PR and HITS algorithms. PaperRank depends on the indirect relationships between scientific papers, instead of the traditional relationships that are represented by citations (Du et al., 2009). Another example of using hubs and authorities alongside the PR algorithm is the framework suggested by (Wang et al., 2013), which uses in the ranking process other information besides citations such as journals, authors, and time information. The network is made up of three different types of nodes which are journals, authors, and citations. In addition to the time

information that takes into account the dynamic nature of the publishing network.

Shubhankar et al (Shubhankar et al., 2011) provided a new algorithm called TopicRanc to detect and rank topics in a wide range of research papers. This algorithm uses the closed frequent keyword- set model for topic detection purposes alongside the PR algorithm. It assumes that the title of any document well summarizes the content and gives a perfect description. They also used a modified time independent PR algorithm to rank the papers, and give each one an authoritative score as a first step in ranking the topics. After that, an authoritative score is assigned to each topic based on the related papers scores. This algorithm has the ability to rank a topic based on its importance in the scientific and research community rather than depending on the topic's popularity. Based on this approach, any paper can belong to one or more natural cluster, and each cluster contains a set of papers that share the same topic.

With the increase in the amount of documents and scientific research published on the Internet in various databases and information retrieval systems, the need for more sophisticated tools and methods has increased to discover and rank information that meets the user's need. Haddadene et al (Haddadene et al., 2012) developed a new approach to rank scientific publications by adapting the PR algorithm alongside with the similarity measures. A representation model of scientific production presented. The similarity between two papers was calculated using the Jaccard Index (Jaccard, 1902) also known as the Jaccard similarity coefficient. And the modified PR algorithm was used to rank documents based on their relative importance.

In some ranking systems, the PR model has been integrated with the N-linear model for the purpose of ranking multiple classes of objects. Through this integration, the ranking of each class is dependent on other classes by a linear constraint system. Le Anh et al (Le Anh et al., 2014) proposed a new ranking system based on this approach (PR alongside N-linear ranking model). In this ranking system, scientific publications are ranked according to the scores of four different classes, publication itself, scientific journal or conference, authors and citations. The system has two models that tested using datasets are built form Digital Bibliography and Library Project (DBLP). The first one is a simple DBLP 3-star ranking model (SD3R) in the case there is no citation information, and the second one is a simple citation 4-star ranking model (SC4R) to rank datasets with citation information. Another example of using this approach is N-star ranking

model proposed by Sohn and Jung (Sohn and Jung, 2015) called Universal-Publication rank (UP rank). It deals with three classes: publication, keyword and citations. The model takes into account the interrelationship between them.

2.2.3. Modified PR Versions:

In addition to using the PR algorithm in many applications and using it along with other models for ranking purposes, there are many modified versions that aimed to address its drawbacks. For example, its failure to take into account the dynamic nature of the citation network and aging characteristics, which causes the problem of bias in favor of older publications. Also, the PR algorithm treats all citations as equal and does not consider the value or quality of the citation. A quick review of the most important modified algorithms that dealt with these problems are mentioned below.

Sidiropoulos and Manolopoulos (Sidiropoulos and Manolopoulos, 2006) introduced SceasRank algorithm which is a modified version of the original PR. It contains two additional parameters. The first one is called the direct citation enforcement factor while the second parameter's primary task is to control the speed at which an indirect citation enforcement converges to zero. Taking the papers publication date into account, converges are usually faster than algorithms similar to PR algorithm.

Sun and Giles (Sun and Giles, 2007) introduced a new ranking algorithm based on the PR algorithm with better ranking performance for scientific publications. The main goal was to overcome the problems caused by the venue's IF. They proposed a new factor called the popularity factor that reflects the effect of the publishing place. For each paper, the popularity factor score is defined by the weight of citations coming from other papers in addition to the publication venue popularity factor.

When calculating the PageRank score for a given paper (P), the score of each citing paper (citing from paper P) is divided by the number of its references (the forward links). This is because the value taken from this citation decreases as the number of references increases. Then the PageRank score for paper P is calculated by summing all scores taken from the citing papers. This approach causes a problem called effect of forward links. So if paper P is cited by a group

of papers that are highly ranked, but has a large number of references, i.e., the forward links, it will result in reducing the rank of paper P. To solve this problem Krapivin and Marchese (Krapivin and Marchese, 2008) proposed a new algorithm called Focused PageRank algorithm for ranking scientific publications based on the Focused Surfer model. The possibility of moving to one of the references increases with the increase in its citation count. FPR model combines the traditional PageRank with citation count approach. Based on this model, papers with a high citation count will receive more citations in the future. Therefore, they will get a better ranking.

In order to deal with the problem of ignoring the value/quality of citation, scholars have continued attempt to weight them through a variety of factors. The most used approach to solve this problem was the content analysis-based approach. An example of this is the study carried out by (Bornmann and Daniel, 2007). This study looks into the extent of benefit that frequently and infrequently cited papers gives to the citing papers, where each reference in the reference list of the citing paper was classified based on two main categories, the section in which the citation took place, and is this mention was significant or cursory. The results show that a paper with high number of citations had greater relevance for the citing paper than a paper with low number of citations. Another content-based citation analysis study conducted by (Taşkın and Al, 2018) for Turkish citations to avoid treating them equally, the main goal of this study is to propose an evaluation model that can analyze citation structures (semantic and syntactic) to define taxonomic citation categories to be used instead of traditional citation based evaluation methods, The citations are divided into a set of main categories, under each category falls a set of sub-categories, the main categories are citation purpose, citation meaning, citation array, and citation shape. Machine learning were used to apply text classification methods for the automatic detection of these categories from the texts. Also, regarding the question are all citations equal? Giuffrida et al (Giuffrida, Abramo and D'Angelo, 2019) suggested a model for evaluating citations by the impact of the citing papers, taking into consideration two points. First the length of the citing paper reference list should not influence the measurement of impact (citation value is independent of the number of publications the citing article cites). Second, the value of a paper that have only one citation can be higher than a paper that have two or more citations.

Many studies have been conducted in order to reduce the problem of bias towards old publications, which is the main problem that our study is trying to deal with. These studies

resulted in new algorithms that take into consideration time information to produce rankings with less bias. One of these algorithms is the CiteRank, suggested by Walker et al (Walker *et al.*, 2007). The basic idea of this algorithm is to predict future citations, taking into account the publication date to give higher scores to recent publications which are expected to receive more citations in the future. Another algorithm called the FutureRank was proposed by Sayyadi and Getoor (Sayyadi and Getoor, 2009). It also estimates the future score for a scientific publication by using additional information besides the citation network. It uses time information including the time of publication in addition to the author's reputation. Another modified version was presented by Wang et al (Wang, Tong and Zeng, 2013), which focused on addressing the problem of ranking scientific publications in a heterogeneous network. In addition, it presents the problem of ignoring time information in the ranking process. The authors suggested a ranking method similar to that used in the FutureRank algorithm. It depends on using multiple networks that include citations, journals, authors, and time information.

Another modified version of PR algorithm conducted by Wei et al (Wei *et al.*, 2021) the main goal of this study is to bypass restrictions of the traditional PR algorithm in the context of standard citation networks, by integrating the text similarity approach (TSA). Where the original PR algorithm gives equal PR values to the downstream nodes, which was improved in this study by giving different importance weights for the downstream nodes using the cosine similarity algorithm which calculate the text similarity score between each pair of nodes (publications) with a citation relationship.

Table 2.2 summarizes all previous studies related to the research topic. Some of these studies suggest integrating the PR algorithm with other models, and other studies address the limitations of PR algorithm by suggesting modified versions to improve performance. Also, some studies dealt with these limitations with the aim of suggesting a new Ranking model as an alternative to the traditional methods of calculating citations, especially studies that try to solve the problem of ignoring the value/quality of citation. It is worth noting that some studies did not refer to the proposed algorithm or model with a specific name, so it was indicated in the table with the authors' names

Table 2.2-A: Summary of ranking algorithms inspired by PR algorithm.

The study	algorithm	Technique	Features
Generalized comparison of graph-based ranking algorithms for publications and authors (2006)	SceasRank	PageRank	• it gives higher scores to papers cited by other important papers and recent papers
Ranking scientific publications using a model of network traffic (2007)	CiteRank	PageRank with landing probabilities approach	• Time-aware ranking. • Ranks authors and conferences based on papers score
Popularity weighted ranking for academic digital libraries (2007)	Popularity weighted ranking algorithm	PageRank	• overcome the problems caused by the venue's impact factor (IF)
Functional use of frequently and infrequently cited articles in citing publications. A content analysis of citations to articles with low and high citation counts (2007)	(Bornmann and Daniel, 2007)	content-based citation analysis	• consider the value of citation based on 1. the section in which the citation took place 2. significant or cursory mention
Focused page rank in scientific papers ranking (2008)	Focused PageRank (FPR)	PageRank with Focused Surfer model	• Combining Pagerank and traditional citation count approach. • Reduce the impact of forward links.
Paperrank: A ranking model for scientific publications (2009)	PaperRank	PageRank with HITS	• depends on the indirect relationships between scientific papers, instead of citations
Futurerank: Ranking scientific articles by predicting their future pagerank (2009)	FutureRank	PageRank with multiple networks	• Time-aware ranking. • generate future citations for the papers.
Weighted citation: An indicator of an article's prestige (2010)	Weighted Citation (WC)	Citation count with weighted citation matrix	• It uses a time quantity called citation gab. Which is the elapsed time from the publication date of the cited paper until the citation occurs
An efficient algorithm for topic ranking and modeling topic evolution (2011)	TopicRank	PageRank with Closed frequent keyword-set	• rank the topic based on its importance in the scientific and research community rather than depending on the topic's popularity.

Table 2.2-B: Summary of ranking algorithms inspired by PR algorithm.

The study	algorithm	Technique	Features
Time-aware ranking in dynamic citation networks (2011)	Retained Adjacency Matrix (RAM)	citation count variable	• Time-aware ranking. • citations are not treated equally
On the PageRank algorithm for the articles ranking (2012)	(Haddadene H, et al. 2012)	PageRank with jaccard Index.	• Calculates the similarity between papers. • PageRank algorithm was used to rank documents based on their relative importance.
Comparing paper ranking algorithms (2012)	NewRank	PageRank with CiteRank	• Time-aware ranking. • direct the random researcher to recent papers in order to cite it more than the old papers
Ranking scientific articles by exploiting citations, authors, journals, and time information (2013)	(Wang et al., 2013)	PageRank with HITS	• Time-aware ranking. • Ranks papers in heterogeneous network
Evaluating scientific publications by N-Linear ranking model (2014)	SD3R, SC4R	PageRank with N-star Ranking Model	• Evaluates everything much more detail based on the context of their relationships.
A novel ranking model for a large-scale scientific publication (2015)	Up Rank	PageRank with N-star Ranking Model	• Considers the query/topic, and the content.
A content-based citation analysis study based on text categorization (2018)	(Taşkın and Al, 2018)	content-based citation analysis	• avoid treating citations equally • define taxonomic citation categories to be used instead of traditional citation based evaluation methods
Are all citations worth the same? Valuing citations by the value of the citing items (2019)	(Giuffrida, Abramo and D'Angelo, 2019)	Citation count and proposed a new indicator (to account for the different contribution of citing publications).	• avoid treating citations equally • evaluating citations by the impact of the citing papers
An Improved PageRank Algorithm Based on Text Similarity Approach for Critical Standards Identification in Complex Standard Citation Networks (2021)	(Wei et al., 2021)	PageRank with text similarity approach (TSA)	• Overcome the limitations of traditional PageRank in the context of standard citation networks • giving different importance weights for the downstream nodes

In summary, great efforts have been made during the past years to improve the process of ranking scientific publications. This is due to their benefits to the scientific community as a whole, including researchers, authors, financiers and institutions. Also, due to the huge volume of data and scientific publications on the Internet, ranking methods have become the most important part of information retrieval systems to meet the users' needs.

The PR algorithm is the most popular model being widely used for ranking scientific publications. But it has some limitations including its preference for old publications over recent ones, treating the publishing network as a static network and not considering its dynamic properties. To bypass these restrictions, the PR concepts was used in conjunction with other models such as N-linear ranking model, HITS model, Jaccard index and Focused Surfer model. Moreover, several modified versions of PR algorithm have been introduced in order to address its drawbacks, especially the problem of bias which we are trying to address in our study, such as CiteRank, FutureRank and NewRank algorithms. However, most of the solutions that were introduced to bypass the problem of bias created new problems, such as transferring bias to recent papers. As was presented in the literature review, some works use time information such as publication date to influence the ranking. In this case, old publications that are still valuable and frequently cited will not be ranked fairly. Also, some works have relied on certain assumptions in order to anticipate future citations. Among of these assumptions is that recently published papers are more useful, and assuming that an author with a good reputation will always have valuable publications. Citation behavior is not fixed, and may be affected by many factors that may not be taken into account at the time of creating these expectations.

Chapter 3

Theoretical Framework

3.1 Introduction

Ranking scientific papers is an important process. Firstly, for the research itself in order to improve the quality of research. Secondly, for the institution as it gives an overview of the output's quality. Therefore, it is considered a strong indicator for the evaluation of educational institutions. The authors also need to prove the impact of their research for several reasons such as satisfying or persuading the funding agencies. In addition, the researcher's reputation is determined by the quality of his/her research. Also, a new researcher or a regular reader who wants to learn often turns to research databases or search engines, which uses their own ranking models to rank query results. This problem is called query rank and there are many algorithms to deal with it. The most used approach is the Link-Based Ranking algorithms, specifically the PR algorithm. However, most of them do not take into account the dynamic nature of the network and treat it as a static network that does not change with time. This leads to the problem of bias. Therefore, this study focused on solving the problem of bias against recent publications by utilizing time information to produce time aware ranking.

This chapter contains the necessary and relevant theoretical information needed to understand the idea of this thesis. It includes the concepts and models related to the research problem. To achieve the objectives of the thesis and answer the research questions, this study relied upon literature reviews. It reviews relevant models and algorithms that others have developed and worked on to convincingly explain and generalize the main thesis's findings.

This chapter contains three sections:

1. It discusses the main concepts, terminology, theories and approaches related to ranking problems and scientific papers ranking such as information retrieval, bibliometrics analysis, citation impact, network theory, ranking approaches, link based ranking and PR algorithm.
2. It provides a brief explanation of the PR algorithm and a review of the enhanced algorithms.
3. It thoroughly Discussion of the proposed ranking method.

3.2 Main Concepts and Terms

3.2.1. Information Retrieval:

It is the process of searching to obtain the required information from its sources, which might be information systems, databases, text files and multimedia, or any other source on the World Wide Web (WWW). The search process can be based on full texts when the need is to search for the documents themselves, or based on other content-based indexing when the required data are metadata, images, sounds, etc. (Ceri et al, 2013). Figure (3.1) shows the basic process in IR system.

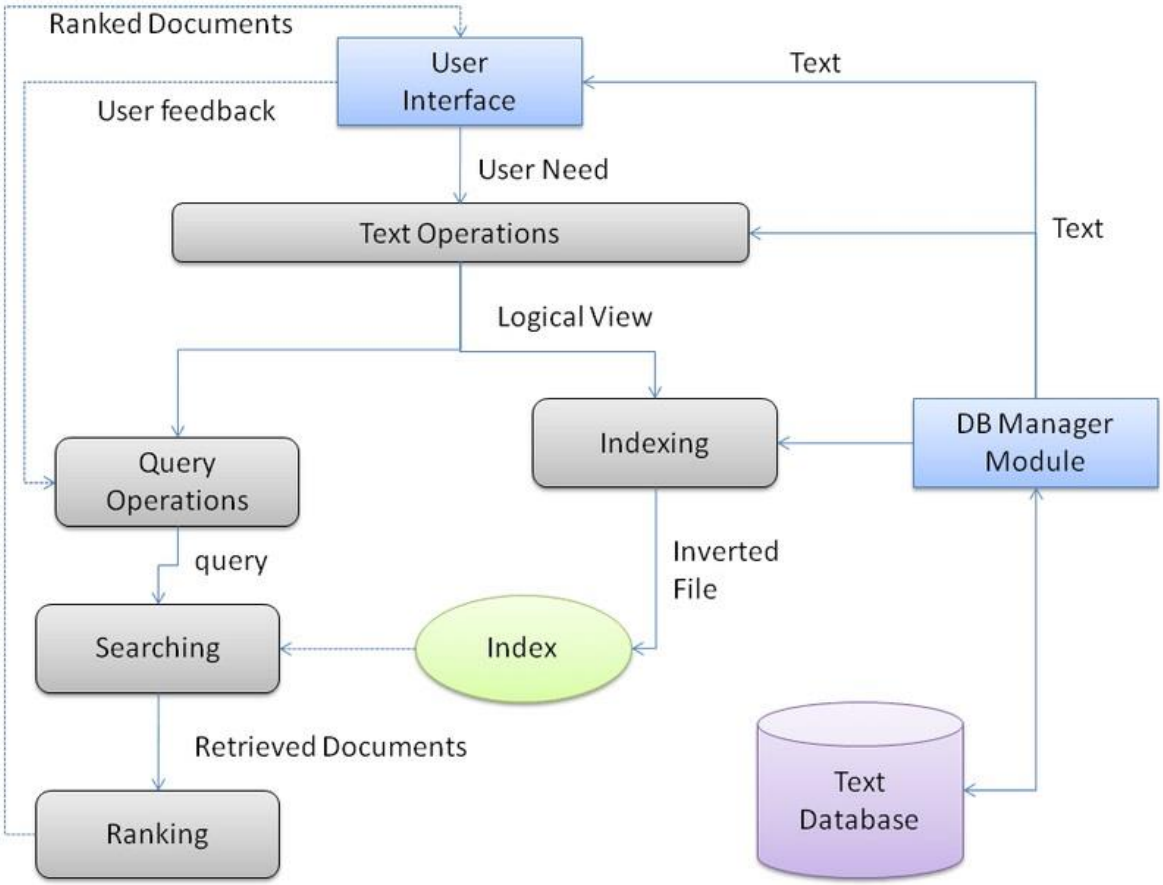


Figure 3.1: The basic process in IR system.
(Buscaldi, 2011).

The process of retrieving data from the search engine through the IR system includes several

steps as follows:

1. The main step is to provide a text database (a set of documents). Then these documents are analyzed and transformed by text operations such as the stemming process that converts the word into its root or basic form. For example, "playing", "plays", and "played" would all be transformed into the root "play". There are other operations on the text such as getting rid of stop words in order to filter the text and remove useless words such as pronouns.
2. The results of the text operations are considered as the logical view of the text database, which is used by the indexing process to create the index. This allows quick searching to be performed on large amounts of data.
3. After that, the data retrieval process can be started, where the user determines his/her needs, then the text operations that used in the indexing process are also applied on it in order to produce a query that represents the user's needs. After that the query is processed to determine the required documents.
4. The retrieved documents should be ranked according to probability or relevance, in order to calculate the relevance, IR systems assign weights to the terms in each document, this helps to determine the importance of the document for a specific query. To do this process, there are many proposed models such as Vector space model and Boolean model.
5. Finally, the ranked documents are returned to the user, who in turn provides a feedback, whether in the case of being satisfied or not in order to improve the results.

3.2.2. Query Rank Problem:

Query rank problem is one of the most important problems that the retrieval system deals with. Ranking algorithms play an important role in choosing the most relevant results to the user's needs, which is based on specific criteria that differ from one algorithm to another. Query ranking is an important process in computer science and is used in various fields such as search engines, relational databases, recommendation systems, document classification systems and scientific papers ranking (B.Yates and R.Neto 2011).

3.2.3. Bibliometrics Analysis:

Bibliometrics analysis was introduced by Garfield in the 20th century with the aim of analyzing, organizing and understanding the major components of each research area. This methodology

includes several disciplines such as social sciences, biology, medicine, in addition to natural, environmental and management sciences. It then expanded to include most disciplines such as computer science and engineering. Bibliometrics methods rely on statistics and numbers in analyzing research trends, and determining the importance of scientific publications through mathematical tools. The Bibliometrics indicators are divided into three main types: quantitative, importance and structural indicators. Quantitative indicators are related to productivity and the volume of publications, whether for the institution or the author. Importance indicators measure the academic impact of research in its field. While structural indicators investigate the existence of collaborative research networks within and outside the research institution (Furner, 2014).

Bibliometric measurements concerned with analyzing scientific publications are called scientometrics. Citation analysis is the most used bibliometric method in the ranking of scientific publications. Citation analysis depends on the creation of citation networks, which are graphs that represent the relationship between documents. Through bibliometric measurements, it is possible to explore the impact of a specific research field, the impact of a specific paper and measure the influence of a researcher or group of researchers. Funding agencies can also verify the results of their funding (Diem and Wolter, 2013).

3.2.4. Network Theory:

Network theory is an analytical study of graphs that represent relationships between a group of discrete entities. These entities are linked by either symmetric or asymmetric relations. It is considered as a part of the graph theory in computer science. A network is a graph that contains a set of nodes that represent entities and edges for the purpose of representing relations. Applications of network theory include many fields such as computer science, electrical and electronic engineering, particle physics, statistical physics, biology, social and cognitive sciences (Estrada, 2012). Figure (3.2) represents an example of a social network analysis.

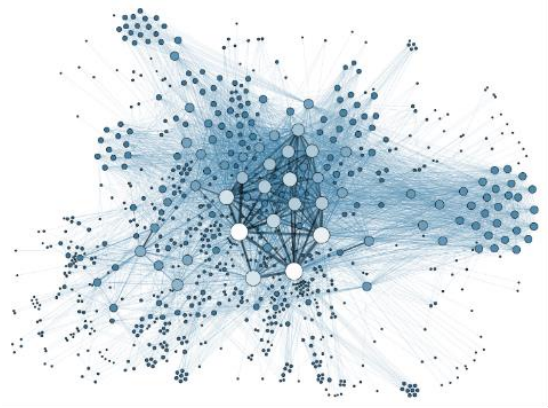


Figure 3.*2*: Visualization of social network analysis.
Source (Grandjean, 2014)**.**

Linking analysis is a subset of network analysis. It is used by many ranking algorithms, including the ranking algorithms used by Google to rank web pages such as PR algorithm (Tsonis et al., 2006).

3.2.5. Citation Network:

Citation network is one of the network theory applications which depend on link analysis. The network consists of nodes that represent papers and links that represent citations. The citation graph is directed so that each edge is oriented from one paper towards another that it cites. A citation network is represented by the adjacency matrix. If we assume a citation network contains N nodes, the presence or absence of an edge between two nodes in the network is represented in the adjacency matrix by entering 1 or 0 (Kanellos et al, 2019).

3.2.6. Citation Impact:

It is an indicator to measure the number of times a scientific publication or an author was mentioned in other scientific publications, whether it's a book, paper, article or even another author. The citation impact is used to measure the impact of academic work, analyze patterns and study the characteristics of scientific publications. Many document ranking systems also rely on citation impact as an indicator to measure importance (Tang et al., 2017).

3.2.7. PageRank Score:

PageRank was developed by Sergey Brin and Lawrence Page (S.Brin and l.Page., 1998) for the purpose of ranking Web pages and query results on search engines such as Google. Then it was

used extensively for evaluation purposes in various applications. For example, it was applied on authors' network to investigate the influence of authors by Liu et al (Liu et al., 2005). Moreover, it was applied by Bollen et al (Bollen et al., 2006) and Chen et al (Chen et al., 2007) on citation networks to evaluate scientific papers. PR score refers to the possibility of choosing a scientific publication by a random user to read it by simulating the random search process. This is to enable the researcher to begin reading a random paper and then moving to another from the references.

3.2.8. Time Aware Ranking:

Ranking algorithms can be divided into two types according to time awareness. The first type does not take into account the properties of the citation network, which change over time. An example of this type is the PageRank algorithm. The second type uses time information to produce more accurate rankings. Therefore, it takes into consideration the citation network changes over time. Some algorithms assign weights to the edges. These weights are quantities of time such as paper age or citation age. Another method is to set unequal landing probabilities for papers and decrease continuously with the age of the paper (Ghosh et al, 2011).

3.3 PageRank Algorithm and Time Aware Ranking Algorithms

3.3.1. Basic PageRank Working Mechanism:

As previously described, PageRank relies on links between pages that refer to the citation. The links are divided into two types, backlinks and forward links as shown in Figure (3.3). A paper is highly rated if it has a large number of backlinks, and it also increases whenever these links come from papers with a high rating (Page et al., 1999). PageRank is an iterative algorithm and its values are calculated using Equation (3.1).

$$R(u) = c \sum_{v \in B_u} \frac{R(v)}{N_v} \qquad (3.1)$$

Where:

- $\boldsymbol{B_u}$: is the set of nodes (papers) that connect with paper u by a backlink. In other words, they are the papers that cite paper u.
- $\boldsymbol{R(v)}$: It is the citing paper value (its importance). Where equal initial values are assigned to

all papers at the beginning. Then these values change with iterations, so that each paper in the graph gets its final value when the difference between the previous iteration and the current iteration becomes too small. This difference is called Epsilon and is preset to stop a repeat when it is reached.

- N_v : It is the number of forward links of the paper v, i.e., the number of paper's references. The goal is to distribute this page's vote evenly across its entire forward links. Assuming that the value provided by the cited paper decreases as the number of references increases, and thus the value of citation decreases.
- **C**: The normalization factor to make ||R||L1 = 1 (||R||L1= |R1 + … + Rn|).

Although the PR algorithm implicitly takes into account the value of citation through the value of the citing paper, this algorithm is still biased to old publications. This is because recent publications do not have a large number of citations. Figure (3.3) shows an example of backlinks and forward links.

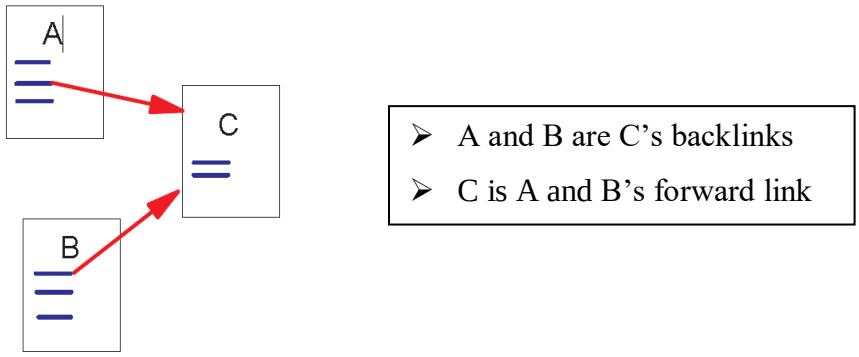


Figure 3.3: Backlinks and Forward links.

3.3.2. Time Aware Versions of the PR Algorithm:

This section reviews the proposed algorithms to address drawbacks in the PR algorithm, especially those related to the ones not taking into account the dynamic properties of the citation network and its changes over time. Most of these algorithms rely on two approaches. The first is to create time-aware ranking methods that consider time information in the evaluation process. The second method does so by exploiting side information such as the papers metadata to create other types of networks, and conduct analysis across multiple networks.

- **CiteRank Algorithm**

In order to take into consideration the ageing characteristics of publication network, Walker et al., (2007) introduced the CiteRank algorithm. It uses the time aware landing probabilities approach. A random walk model was developed to predict future citations by relying on time information, and also assuming that the researcher always starts his/her research from a recent publication and then moves to an older publication and so on until he/she is satisfied. Therefore, higher ratings will be given to recent publications to reduce bias. The CiteRank scores are calculated using Equation (3.2) (Walker et al., 2007).

$$S = 1 \,.\, \bar{p} + (1 - a)W \,.\, \bar{p} + (1 - a)^2 W^2 \,.\, \bar{p} + \cdots \qquad (3.2)$$

$$p_i = e^{-(t_c - t_i)/\tau}$$

Where:

- $\boldsymbol{p_i}$: The probability of choosing paper i.
- $\boldsymbol{t_c - t_i}$: The age of paper i.
- $\boldsymbol{W}$**:** The adjacency matrix that represents the citation network.
- $\boldsymbol{a}$ and $\boldsymbol{\tau}$: Constant values.

- **FutureRank Algorithm**

Another algorithm designed to capture the dynamic nature of publication networks is called FutureRank (Sayyadi and Getoor, 2009). In addition to the citation network, the author's reputation and time information are used in order to generate future citations for recent papers based on several assumptions. They include that good research papers are written by highly reputable researchers and newly published papers are more useful; hence getting more citations in the future. The approach used is time-aware and multiple networks (author-paper network and paper-paper network) as shown in Figure (3.4). The value of a particular author is distributed on the papers that he/she authored, and the value of the papers is distributed to their authors.

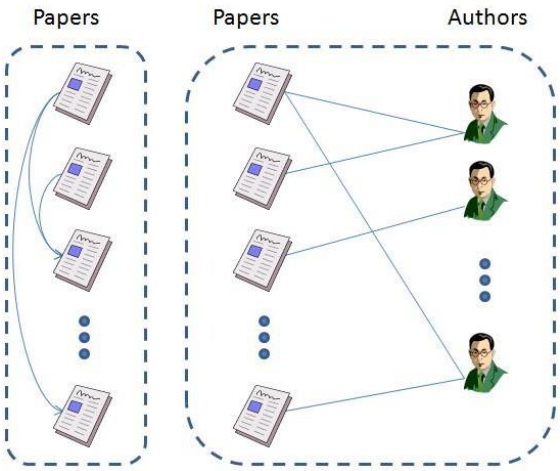


Figure 3.4: Paper-Paper network and Paper-Author network.
Source (Sayyadi and Getoor, 2009)**.**

- **Retained Adjacency Matrix (RAM)**

This method uses a citation count variable. But citations are not treated equally, as the cited paper age affects the citation value. It gives a higher value to the link coming from a recent paper and the paper's associated value decreases with age. This algorithm is based on the assumption that more recent information is often preferred by people. Where the parameter ($\gamma < 1$) is used to give a higher weight to a recent paper, and this weight decreases with the age of the paper. If v is the correlated value with the citation link for a paper published in year t_n, a scaled down value $\gamma^{n_i} v$ is the correlated value with a citation link paper published in year t_{n-n_i}. Therefore, a paper published in year t_n will be given a higher weight than a paper published earlier. The Retained adjacency matrix is constructed using Equation (3.3) (Ghosh et al., 2011).

$$R_{n,\gamma}(i,j) = \begin{cases} \gamma^{N-n_i}, & if\ p_i\ cites\ p_j\ and\ t(p_i) = t_{n_i} \leq\ t_n \\ 0, & otherwise \end{cases} \qquad (3.3)$$

Where:

- $\boldsymbol{\gamma}$: The retention probability
- **N**: The current date
- $\boldsymbol{n_i}$: Publication date of paper i

- **NewRank (NR)**

This method assigns weights to citations depending on the cited paper age and also uses landing probabilities. This algorithm follows the same approach as the CiteRank algorithm by measuring the probability of choosing a particular paper (Kanellos *et al.*, 2019). Suppose that p represents the vector that includes the probabilities of choosing paper i where $p_i = e^{-t_i/\tau}$. t_i represent the paper age and τ is the characteristic decay time.

Let $D(p_j)$ the probability of reaching a reference from paper j, we can calculate it using Equation (3.4) (Dunaiski and Visser, 2012).

$$D(p_j) = \frac{p_j}{\sum_{pk \in N+(p_i)} p_{pk}} \qquad (3.4)$$

The previous equation basically normalizes the paper's initial value through the initial values of the papers in their references list. The goal is to direct the random researcher to recent research in order to cite it more than the old research. This algorithm adopts the iterative approach used in the PageRank algorithm. As a result, recent research has the potential to obtain more citations than old research (Dunaiski and Visser, 2012).

- **Weighted Citation (WC)**

This algorithm depends on the number of citations. It uses a weighted citation matrix by the time quantity called citation gab. Citation gab is the elapsed time from the publication date of the cited paper until the citation occurs (Kanellos et al., 2019).

3.4 The Proposed Ranking Method

The new method uses the linking Analysis based approach with Time Aware Ranking to produce rankings that take into account the citation network dynamic nature and its change over time. The new algorithm avoids generalizing assumptions to solve the problem of bias to old publications, such as using author's reputation or the paper's metadata. These data are very valuable for generating more comprehensive assessments of the scientific research impact. However, they cannot be relied on with the aim of reducing bias by producing the expected future citations of recent papers without considering the fact that citation behavior changes over time.

Moreover, time information such as publication date cannot be used to influence ranking results arbitrarily by increasing or decreasing the score of a particular paper based on its recency. Instead, a new indicator should be adopted, which is the citation change rate over time. This indicator measures the change in reliance on a particular paper, whether it is recent or old. The new indicator ensures fairness and minimizes bias in favor of old publications. In other words, if the paper was published in the past and obtained a large number of citations during its life, but few of these citations occurred recently and the number of its citations are constantly decreasing, this means the paper is no longer important thus reliance on it decreases. Therefore, its value must be reduced. While papers that still receive continuous citations, will receive good values and their value may not be underestimated only because the date of their publication is old.

3.4.1. The Citation Change Rate:

This indicator gives a clear perception of the ability of an old scientific paper to keep giving and being important in its field by consistently appearing in the reference list of recent papers. It is not sufficient for the paper to receive a large number of citations to obtain a good ranking, because the citation date also matters. On the other hand, for recent papers that are still in the growth process, we can identify the nature of this growth. If the citation rate is high and is increasing year after year, it indicates that the paper is valuable and will receive many citations in the future. Therefore, it must be given good rankings, instead of solely relying on the citation count. As these papers are still new and did not take enough time to collect many citations. In order to calculate the citation change rate, the paper age should be at least two years old. The accuracy of the results increases with the increase of paper age, because the citation behavior of this paper is stabilized. Equation (3.5) (Adams and Essex, 1999) calculates the average rate of change.

$$\frac{f(x_2)-f(x_1)}{x_2-x_1} \qquad (3.5)$$

where f is a function depends on x, in our case x is time (t) and f is the number of citations (C) in t years. so, the modified formula becomes as shown in Equation (3.6).

$$ACR = \frac{C(t_2)-C(t_1)}{t_2-t_1} = \frac{\Delta c}{\Delta t} \qquad (3.6)$$

Where:

- ACR: The annual citation change rate
- $C(t_2)$: Number of citations in the current year.
- $C(t_1)$: Number of citations in the Publishing year.
- t_2: The current year.
- t_1: The publishing year.

3.4.2. The additional Information in the citation network:

The citation network should contain the publication date for each paper. This is in order to calculate the cited paper age. Also, to identify the time when each citation occurred using the publication date of the citing paper.

Figure (3.5) shows a simple citation network containing 15 papers published over 5 years. By making a simple comparison between node 1 and node 2, which are the oldest in the network, they both have 3 citations. But the annual citation rate for Node 1 is higher because it gets citations continuously from recent papers. As for Node 2, the citation on it stopped three years ago, which means that the reliance on it is declining; hence paper 2 should receive a lower ranking.

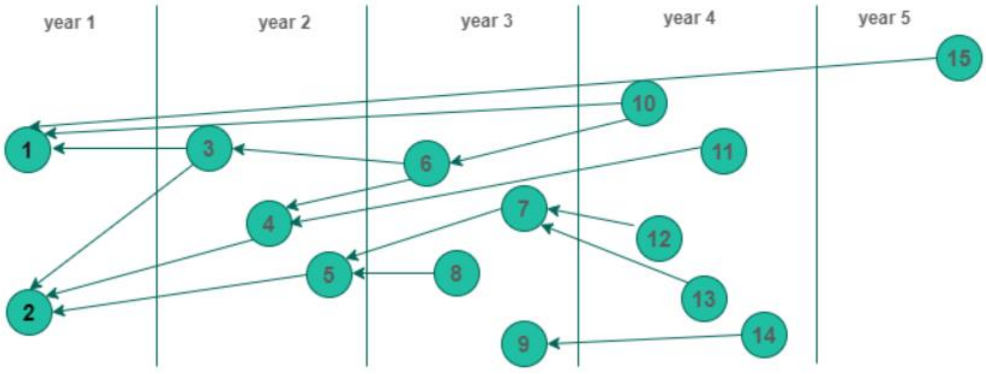


Figure 3.5: Time aware citation network.

3.4.3. The Modified PageRank Scores:

The PR algorithm does not depend on the citation count directly, but rather takes into account the citing papers values. The value is mainly calculated by depending on the number of citations. This leads to bias. But by adding citation change rate to the equation, the bias towards old

publications is reduced. The modified PageRank score is calculated by Equation (3.7).

$$MPR = c \sum_{v \in B_u} \frac{R(v)}{N_v} + (\frac{\Delta C}{\Delta T})\, s$$

$$\text{In which} \quad MPR = PR + ACR * S \qquad (3.7)$$

Where:

- **MPR**: Modified PageRank (final rank).
- **PR**: PageRank value.
- **ACR**: Annual Citation change rate.
- **S**: Scale (constant value).

Figure (3.6) shows a process flow diagram to implement the proposed ranking method and validates the results. The goal is to obtain a bias-free ranking of a specific search query results within an information retrieval system.

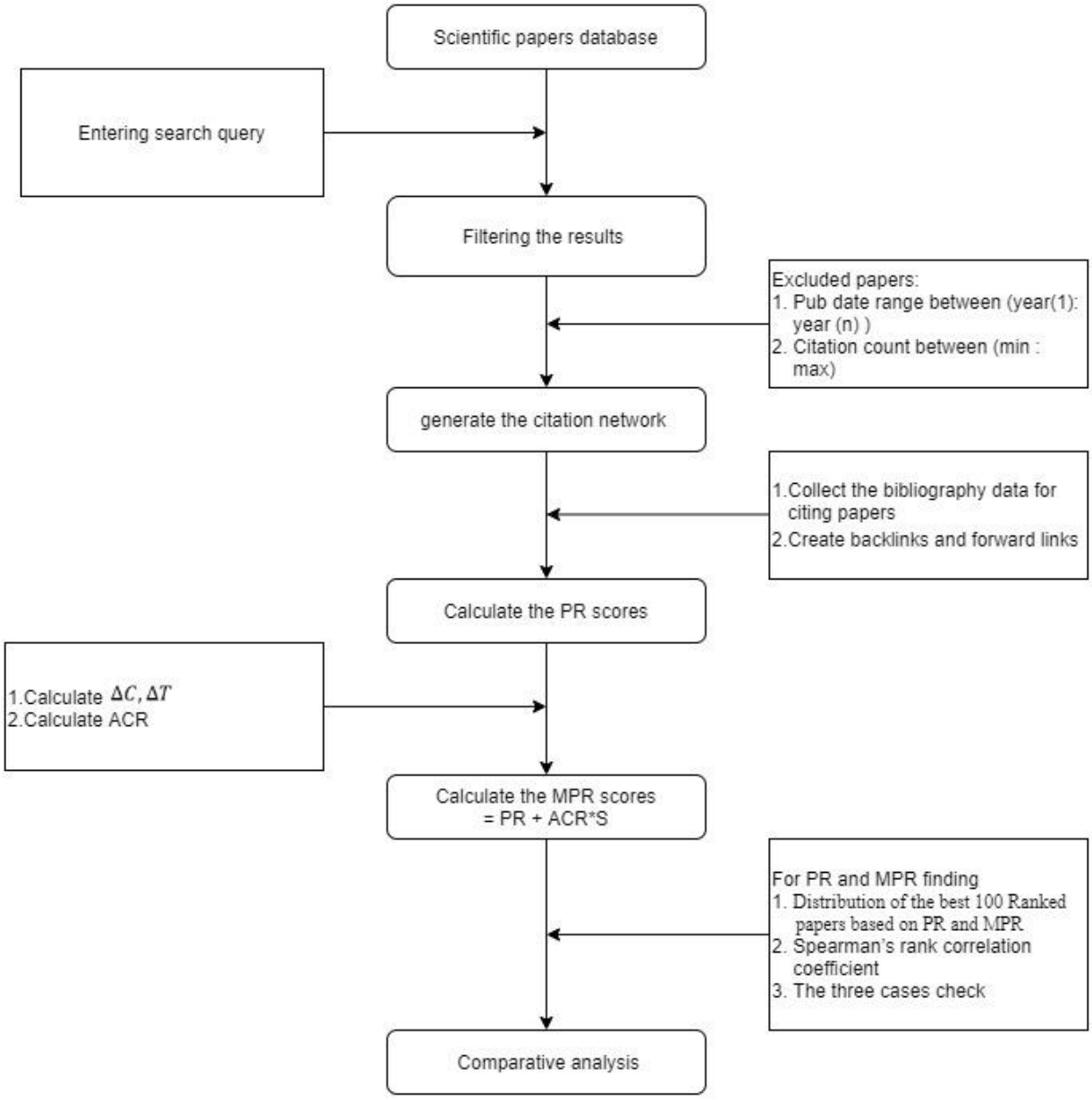


Figure 3.6: The proposed ranking method flow diagram.

Chapter 4

Experiments

4.1 Introduction

This chapter aims to explain the implementation of the modified version of the PR algorithm (MPR). Starting with collecting and preparing data to extract the required information. Then building the citation network to obtain the linking structure, calculating the original PageRank values and presenting the proposed ranking method. It also contains a comparative analysis between the results of the original PR algorithm and the MPR algorithm. In addition, it lists the evaluation metrics that are used to evaluate the proposed method.

4.2 Data Collection

In order to conduct experiments and validate the proposed method, Dimensions database was used to obtain the data. This database provides a wide range of research information. It contains an open and comprehensive data infrastructure that empowered users to explore connections between a wide range of research data.

4.2.1. Determine the Necessary Data Using Dimensions Web App:

The simplest way to get data from Dimensions database is via Dimensions Web app. It allows making limited queries, such as searching for scientific papers related to a specific field. Then retrieving the results, where the papers or the bibliometric data is provided. This is what we need. A sample is shown in Table (4.1).

Table 4.1: A Sample of bibliometric data collected by Dimensions Web app.

Publication id	DOI	Title	Pub year	Authors	Time cited
pub.1059030406	10.1088/0031-9155/59/16/4739	Measurement of the dielectric properties of the epidermis and dermis at frequencies from 0.5 GHz to 110 GHz.	2014	K Sasaki, K Wake, S Watanabe	41
pub.1059029692	10.1088/0031-9155/57/9/2555	Toward automatic detection of vessel stenoses in cerebral 3D DSA volumes.	2012	F Mualla, D Hahn, J Hornegger	1
pub.1002258494	10.1016/j.kjms.2011.08.006	Physics teaching in the medical schools of Taiwan	2012	Jiann-wien, Roy Hsu	3
pub.1079029042	10.1684/abc.2014.0985	Mass spectrometry: from physics fundamentals to laboratory medicine	2015	Roselyne Garnotel, Edgard Delvin	0
pub.1059026457	10.1088/0031-9155/51/6/013	MANTIS: combined x-ray, electron and optical Monte Carlo simulations of indirect radiation imaging systems.	2006	Aldo Badano, Josep Sempau	63
pub.1026711521	10.1118/1.2786860	The American Board of Radiology perspective on maintenance of certification: Part IV: Practice quality improvement in radiologic physics	2007	G. Donald Frey, Geoffrey S. Ibbott, Richard L. Morin -	5

By using Dimensions Web App, the required data cannot be obtained due to many limitations. This includes the inability to create accurate queries, as well as restrictions on the amount of data that can be obtained. But what is needed can be determined.

4.2.2. Return the Data Using Dimensions API:

Dimensions database provides an analytics API. The analytics API supports the extraction of Dimensions data for use in complex analyses and visualizations. The API uses a query language

called Dimensions search language (DSL) specifically developed for Dimensions data. So, we can retrieve, aggregate, and sort data from highly specific requests in a single API call.

Using Dimensions API, we got the required data based on the conditions that must be met to conduct the experiments. They are:

- The scientific papers must be related to one field.
- The papers must also be published in a number of years (a long time period).
- The papers must have citations.

Thus, the required query that meets the conditions will be as follows:

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

This query returns all scientific papers related to the topic of Medical physics published between 2005 and 2017, and the number of citations per paper is between 20 and 200. All papers are in the same field, because the characteristics of citation differ from one field to another. The number of researchers and the number of research varies between fields. Therefore, comparing scientific papers from different fields may be injustice, and this is what most ranking systems try to avoid. Also, these papers must be published over a wide period of time so that we can test whether the new method reduces bias or not. As for citations, it is necessary that the data set does not contain papers without citations or very few citations, because in this case they will take the same rank whether the original or modified algorithm is used. Figure (4.1) shows the process of capturing and filtering data based on the required conditions and Table (4.2) shows a sample of this data.

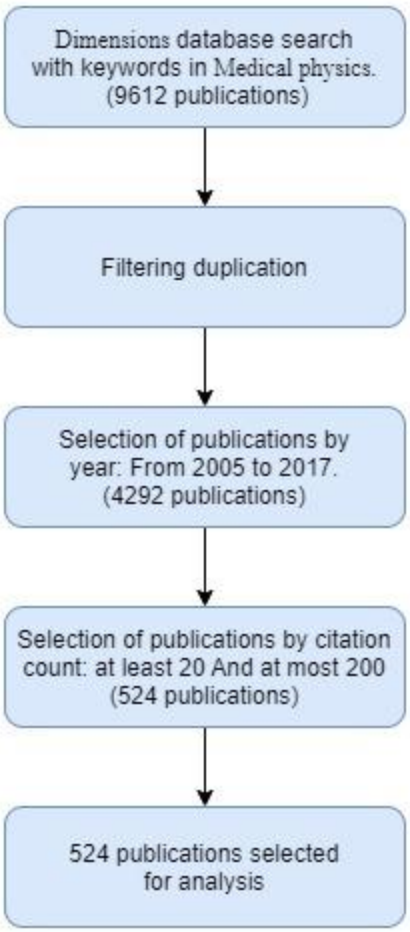


Figure 4.1: The process of capturing and filtering data.

Table 4.2: Sample of the collected data using Dimensions API.

id	year	doi	times_cited	title
pub.1099918061	2017	10.1109/tkde.2017.2785824	23	MCS-GPM: Multi-Constrained Simulation Based Graph Pattern Matching in
pub.1093028380	2017	10.1002/mp.12702	63	RECORDS: improved Reporting of montE CarlO RaDiation transport Studies
pub.1091615833	2017	10.1088/1361-6633/aa8b1d	55	Review of medical radiography and tomography with proton beams
pub.1092367839	2017	10.3762/bjoc.13.219	28	Phosphonic acid: preparation and applications
pub.1092226210	2017	10.1016/j.freeradbiomed.2017.10.003	82	Limitations of oxygen delivery to cells in culture: An underappreciated problem
pub.1091850388	2017	10.1088/1361-6595/aa8d4c	27	Foundations of low-temperature plasma physics—an introduction
pub.1092148337	2017	10.3390/polym9100494	101	Block Copolymers: Synthesis, Self-Assembly, and Applications
pub.1090670633	2017	10.1016/j.nima.2017.06.017	22	Proton beam characterization in the experimental room of the Trento Proton The
pub.1091274054	2017	10.1098/rsfs.2016.0159	28	Evolution viewed from physics, physiology and medicine
pub.1091085605	2017	10.1002/acm2.12146	30	AAPM-RSS Medical Physics Practice Guideline 9.a. for SRS-SBRT
pub.1090306251	2017	10.1097/hp.0000000000000674	31	Appropriate Use of Effective Dose in Radiation Protection and Risk Assessment
pub.1090837242	2017	10.1088/1742-6596/874/1/012029	57	Horizon 2020 EuPRAXIA design study
pub.1085591560	2017	10.1002/mp.12371	65	Future of medical physics: Real-time MRI-guided proton therapy
pub.1085475539	2017	10.1186/s41747-017-0006-5	25	Trends in radiology and experimental research
pub.1086050674	2017	10.1364/boe.8.003248	72	Twenty-five years of optical coherence tomography: the paradigm shift in
pub.1091274530	2017	10.1115/1.4037671	31	Applicability Analysis of Validation Evidence for Biomedical Computational Mode
pub.1085591656	2017	10.1002/acm2.12080	29	AAPM Medical Physics Practice Guideline 8.a.: Linear accelerator performance
pub.1090323410	2017	10.4324/9781315268897	42	A History of Technoscience
pub.1085292639	2017	10.1142/s0217732317400090	26	Overview of the future upgrade of the INFN-LNS superconducting cyclotron
pub.1084251244	2017	10.1186/s12938-017-0326-y	23	Computational medical imaging and hemodynamics framework for functional

4.3 Building Citation Network

4.3.1. Collect the Bibliography Data of Citing Papers:

To build citation network we need to collect the bibliography data for all citing papers using Dimcli language. This is because the citing papers must be a part of the network (Part of the network's nodes) so that we can make links between it and the cited papers (the papers returned from the previous query).

Using Dimcli, this query was created to obtain citing papers data:

```
%dsl search publications
where reference_ids in ["pub.1084251244","pub.1084131849","pub.1083961865",
pub.1021478049","pub.1062159556","pub.1059648986","pub.1092367839"]
return publications[id+doi+title+year+times_cited+reference_ids]
```

This query takes the paper ID, and searches for it in the references of all papers published in the Dimensions database. If it finds the ID in the reference list of one of the papers, the paper is returned because it cited the paper with this ID. This query will be applied on all papers that resulted from the first query until we get all of the citing papers.

Now we have all the network nodes. They include the cited papers that resulted from the first query (524 paper), and the citing papers that resulted from the second query (25250 paper).

4.3.2. Create Backlinks and Forward Links:

The links between them can be identified through the reference list of each paper. By using Excel power query editor, the reference list was divided into separate fields, each containing only one reference and corresponding to the citing paper. Then two separate files are created, one containing the nodes and the other containing the links (edges). Figure (4.2) shows these two files.

Nodes

	A
1	**id**
2	pub.1019393936
3	pub.1128227127
4	pub.1128217986
5	pub.1128222499
6	pub.1057799069
7	pub.1027376959
8	pub.1128220408
9	pub.1128216814
10	pub.1101373041

Edges

	A	B
1	**source**	**target**
2	pub.1134317108	pub.1019393936
3	pub.1134317108	pub.1128227127
4	pub.1134317108	pub.1128217986
5	pub.1134317108	pub.1128222499
6	pub.1134317108	pub.1057799069
7	pub.1134317108	pub.1027376959
8	pub.1134317108	pub.1128220408
9	pub.1134317108	pub.1128216814
10	pub.1134317108	pub.1101373041

Figure 4.2: Sample of nodes and edges.

4.3.3. Generate the Network:

After that, we enter the previous files into Gephi tool to create a citation network. Figure (4.3) shows a part of the resulting network.

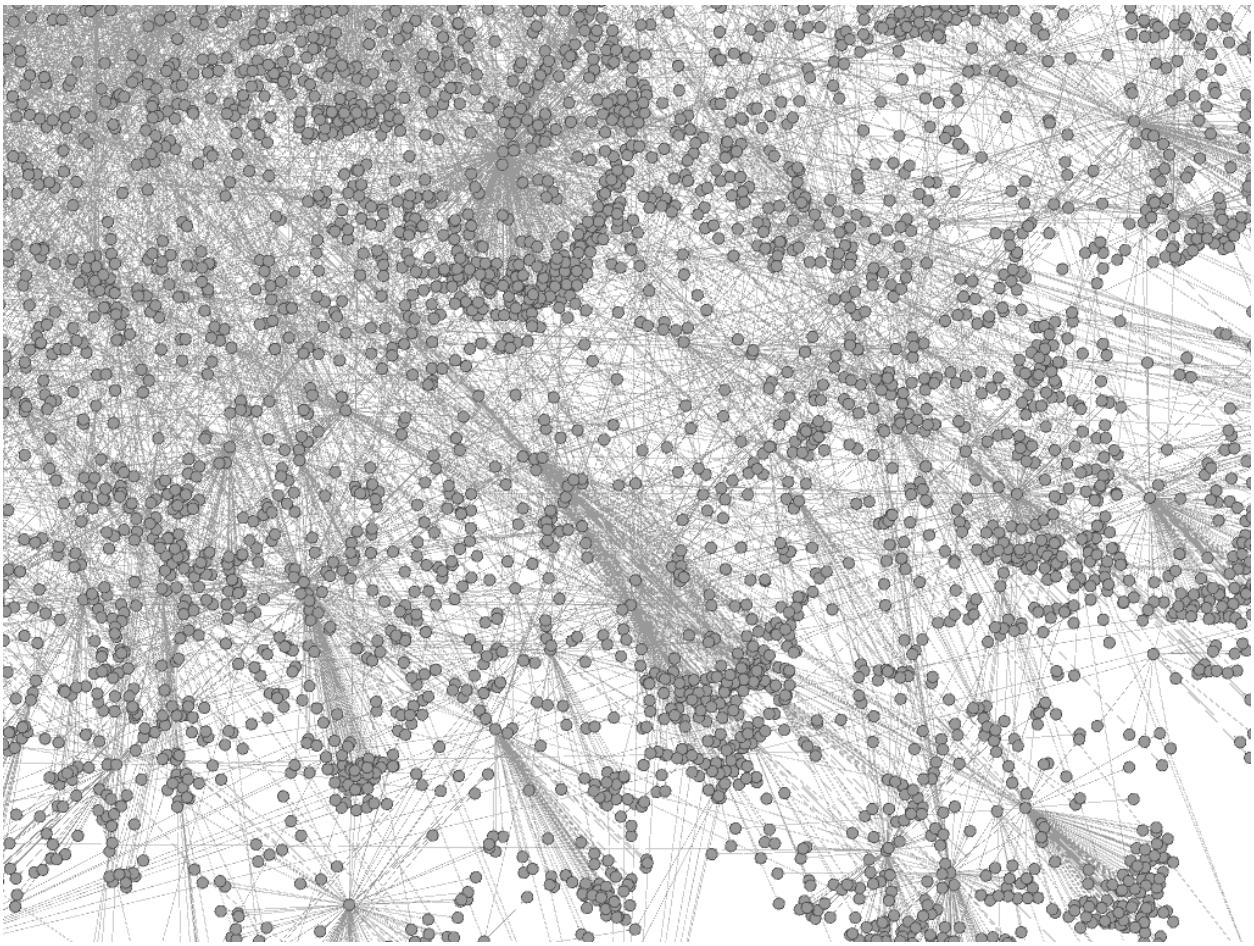


Figure 4.3: Citation network visualization for our dataset.

4.4 Calculate the PageRank Values

The PageRank algorithm is based on the linking structure of the papers. So, any paper containing many citations must have good ranking. The values are calculated using equation (3.1). Since we have the linking structure (citation network), the PR algorithm can be applied using Gephi to calculate the scores. Table (4.3) shows the resulting PR scores.

Table 4.3: A sample of the PR scores.

Id	Label	PageRank
pub.1021217544	pub.1133376887	0.000428
pub.1059026106	pub.1036176796	0.000193
pub.1059026111	pub.1091290698	0.000168
pub.1061694895	pub.1093379245	0.000168
pub.1046741936	pub.1085430147	0.000146
pub.1026762707	pub.1093381035	0.000146
pub.1047930863	pub.1093382213	0.000136
pub.1051578899	pub.1093382205	0.000124
pub.1009613690	pub.1113527918	0.000123
pub.1025773096	pub.1093380523	0.000121
pub.1091181335	pub.1012631019	0.000113
pub.1061694517	pub.1093379875	0.000106
pub.1061296582	pub.1093388395	0.00009

4.5 Calculate the Citation Change Rate

To calculate the annual citation change rate, which gives an indication of the amount of change in reliance on a scientific paper as a reference, we use equation (3.6). For each paper, the following steps were followed using Python. To calculate the annual citation change rate.

1. Fetch publication date of all cited papers, and put them in a python dictionary to be used later in calculating the papers ages.

2. Fetch publication date for all citing papers to each paper in our dataset. This is to determine the time of each citation.

3. After determining the occurrence date of each citation (by publication date for citing papers), we need to counting the number of citations that occurred each year.

4. Now we have the number of citations that occurred each year. Therefore, we can calculate ΔC values by subtracting the total citations in the last year from the total citations in the first year.

$$\Delta C = Ct_2 - Ct_1$$

5. Then calculate the paper age, by subtracting the current year from the paper publication year.

$$\Delta T = t_2 - t_1$$

6. Finally, calculate the annual change rate.

The paper ranking will be affected by this value, which can be positive or negative. There is a direct proportional relation between the citation change rate and the paper ranking. Figure (4.4) illustrates the previous steps that were needed to obtain the ACR.

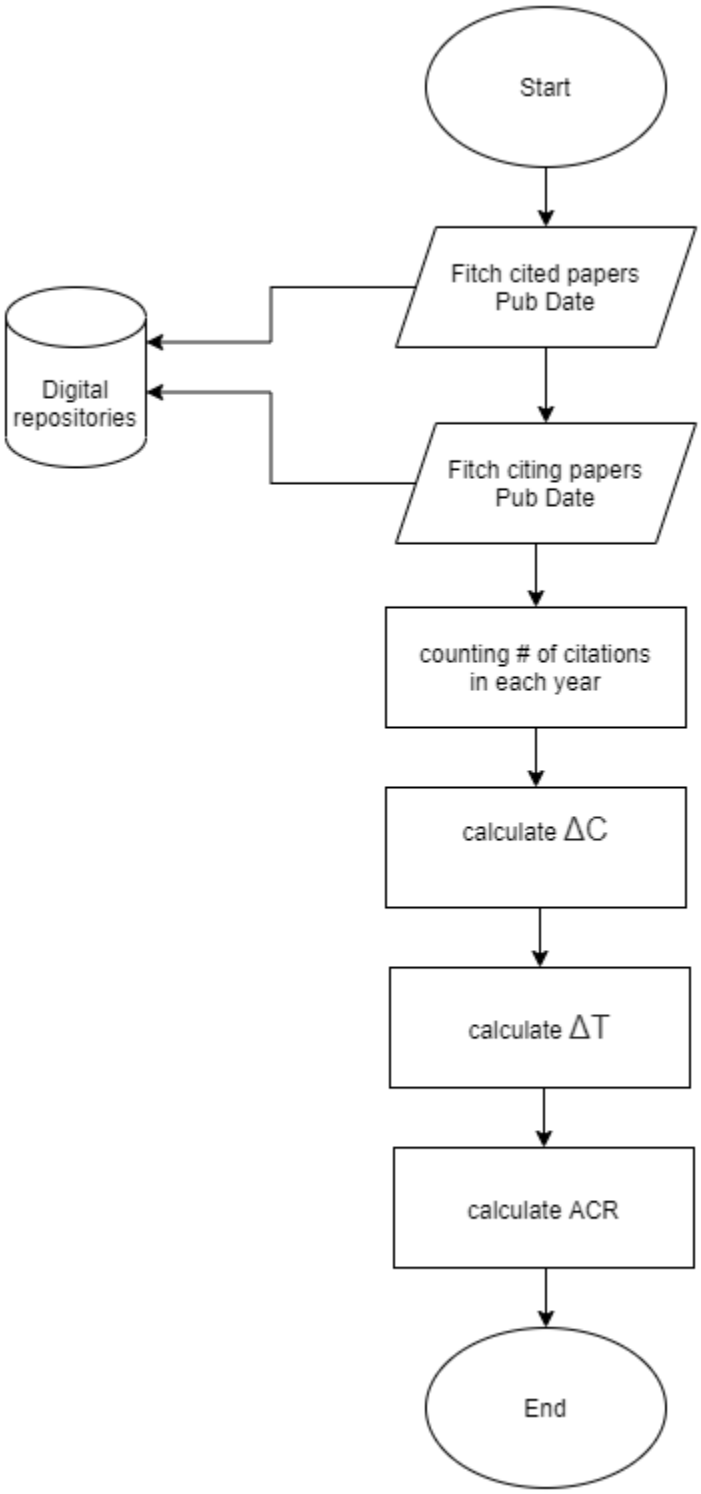


Figure 4.4: Annual citation change rate calculation steps.

The resulting values need to be scaled. This makes the equation sides more balanced where the other side of the equation contains the original PR value. The scale S was set to .01 after testing other values, and it gave the most balanced result compared to other tested values. Table (4.4) shows a sample of annual citation change rate (ACR) values before scaling.

Table 4.4: A sample of ACR values.

id	ACR
pub.1099918061	2.333333
pub.1093028380	11
pub.1091615833	7
pub.1092226210	14.66667
pub.1091850388	2
pub.1092148337	14
pub.1090670633	3.333333
pub.1091274054	1.666667
pub.1091085605	5
pub.1090306251	3
pub.1090837242	4.333333
pub.1086050674	7.13
pub.1091274530	3.3
pub.1085591656	2.2
pub.1090323410	0.666667
pub.1085292639	2
pub.1084251244	1.4

4.6 Calculate the Modified PageRank scores

To calculate the modified values after adding the new indicator, equation (3.7) was used. The pseudocode of the MPR is as the following:

Procedure: Calculate the Modified PageRank scores.

Required:

ID: Paper id.

PR: PageRank score.

t_1: Paper's publication date.

t_2: Current date.

$C(t_1)$: Number of Citations in the first year.

$C(t_2)$: Number of Citations in the last year.

S: Scale (Constant value).

For each paper in dataset

Get:

$PR, t_1, t_2, C(t_1), C(t_2)$

Compute ΔC = $C(t_2) - C(t_1)$

Compute $\Delta t = t_2 - t_1$

Compute ACR = $\frac{\Delta C}{\Delta T}$

Compute MPR= PR + (ACR * S)

Print MPR score.

End.

Chapter 5

Discussion of Results and Validation

5.1 Introduction

In order to evaluate the results and ensure the achievement of the study objectives, a set of measures was used to evaluate the performance of the modified algorithm. In addition, a comparative analysis was conducted between the results of the original algorithm and the modified algorithm. To analyze the results and make comparisons, a list of top 100 papers classified according to each algorithm will be used.

The evaluation process for ranking algorithms faces many challenges that make it difficult and non-standardized. Such as the absence of a ground truth of the actual Ranking (Wang, Tong and Zeng, 2013), and the lack of recognition by the research community of comprehensive evaluation standards (Dunaiski and Visser, 2012). Moreover, each ranking algorithm is designed to achieve specific goals and satisfy the desires and requirements of specific users. Some algorithms aim to assess impact, others aim to assess spread and reputation, while others aim to rank papers based on scientific or economic returns (Sidiropoulos and Manolopoulos, 2006).

5.2 Distribution of Ranking Results Among Papers Publication Date

To ensure that there is an improvement in the results in favor of recent publications that deserve a better ranking and reduce the bias to old publications, the publication years of the top 100 papers classified by each algorithm will be compared.

Figure (5.1) shows the number of papers published each year that ranked among the top 100 papers using the PR algorithm. All of these papers were published between 2005 and 2017. The results show that among all papers published during the last three years, only 14 ranked among the top 100 papers, and only 3 of them were published during the last two years (2016 and 2017). It is evident that original the PR is biased against recent papers and gives higher scores to old papers.

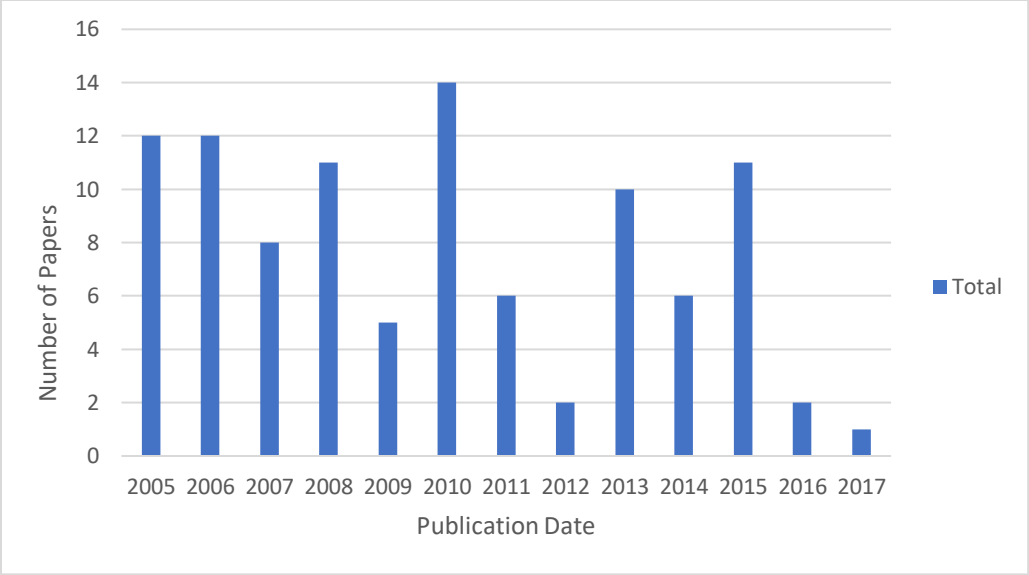


Figure 5.1: Distribution of the best 100 ranked papers based on the PR algorithm.

Figure (5.2) shows the number of papers published each year that ranked among the top 100 papers using the MPR algorithm, the results show an improvement in the scores of recent published papers, 26 papers instead of only 14 were ranked among the top 100 papers, and 12 papers of them were published during the last two years (2016 and 2017). So, the bias against recent publications has diminished, and the rapidly growing papers are taking better scores. On the other hand, the old publications that have become less reliable, even though they have a large number of citations obtained in the past, will taking less scores.

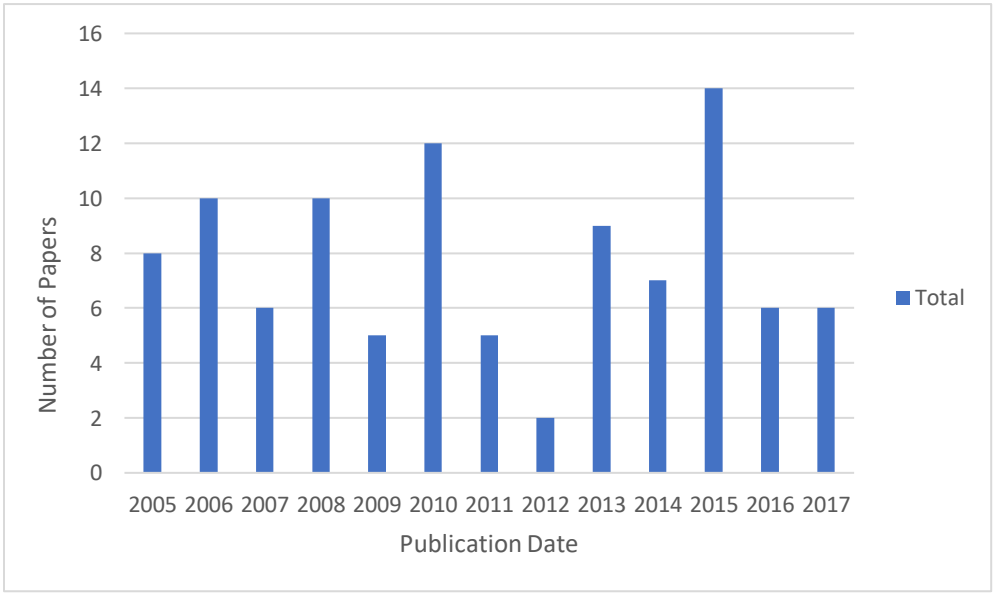


Figure 5.2: Distribution of the best 100 ranked papers based on the MPR algorithm.

To clarify the above in numbers, Table (5.1) shows the distribution of the top ranked papers according to the publication date.

Table 5.1: Distribution of the top ranked papers according to the publication date.

publication date	PR Papers	MPR Papers
2005	12	8
2006	12	10
2007	8	6
2008	11	10
2009	5	5
2010	14	12
2011	6	5
2012	2	2
2013	10	9
2014	6	7
2015	11	14
2016	2	6
2017	1	6

5.3 Assess the Similarity by the Spearman's Rank Correlation Coefficient

The change in results should be logical in which they don't differ radically, and don't cause large and illogical jumps in the papers' ranking. To achieve this, the similarity between the original PR algorithm and the MPR is assessed by the Spearman's correlation coefficient (Spearman's ρ). It is calculated by Equation (5.1) (Myers, Well and Lorch Jr, 2013).

$$R = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (y_i - \bar{y})^2}} \qquad (5.1)$$

Where:

- (x) and (y) are the ranks.
- (x) bar and (y) bar are the mean ranks.

The value of R is 0.92. This is a strong positive correlation. It indicates that the changes are logical and the results are reliable. The PR algorithm gives good results. Therefore, what is required is improvement on a certain part, without radical changes in the results. This is what happened here. Figure (5.3) shows the positive correlation between the PR algorithm and the MPR algorithm.

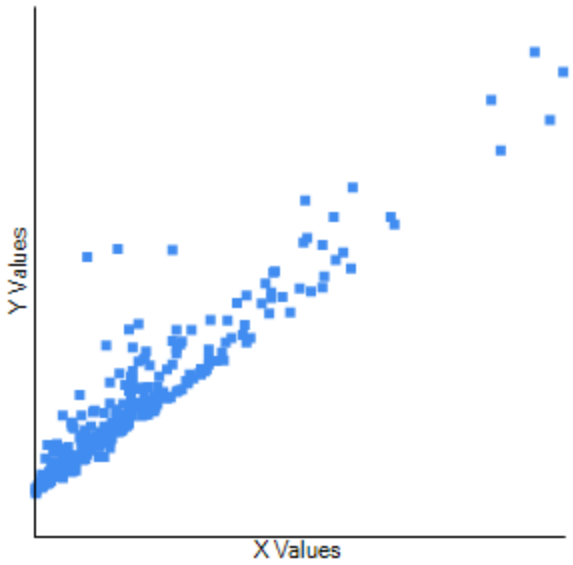


Figure 5.3: Positive correlation between X and Y.

5.4 Testing the Algorithm Accuracy in the Three Cases

In order to ensure that the modified version avoids the problems found in the previous solutions, including imparting bias to recent publications, a group of papers belonging to each of the three cases were selected. The three cases are recent publications, old publications that are still valuable and old publications that are no longer valuable. Then the citation behavior of these papers was analyzed to compare it with the changes in ranking, as shown in Tables 5.2, 5.3, and 5.4.

5.4.1. Recent Publications That Received a Better Ranking:

Table 5.2-A: Citation's behavior and ranking's results for a sample of recent publications that received a better ranking.

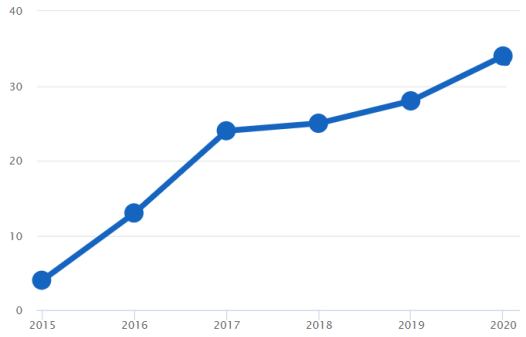
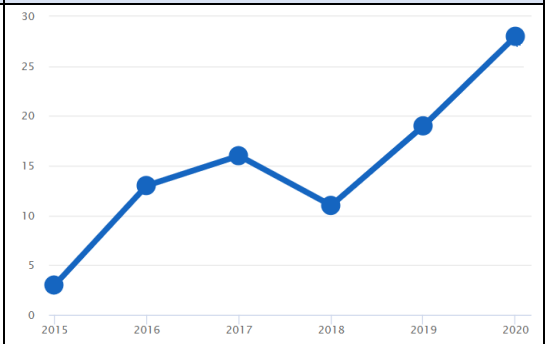
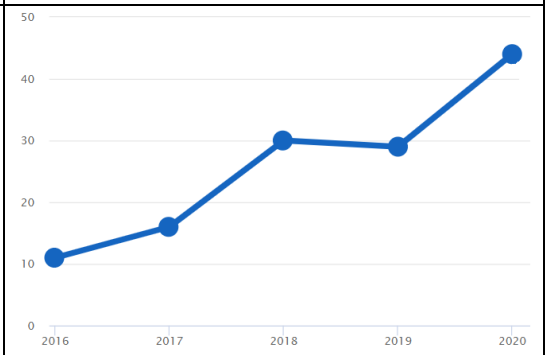
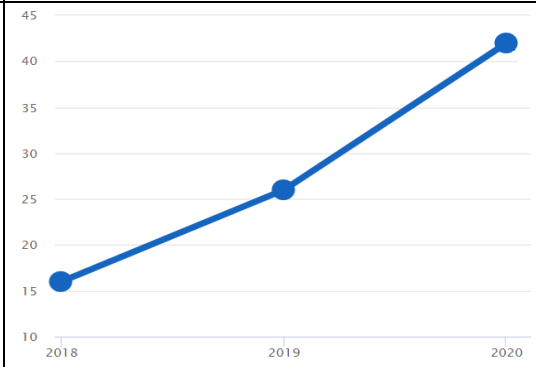
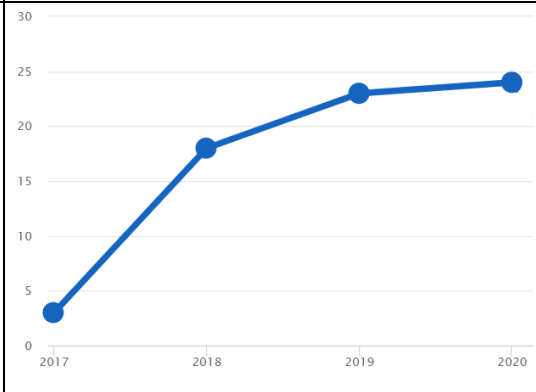
Paper ID	Publication date	PR ranking	MPR ranking	Citations over years
pub.1049556687	2015	31	18	

Table 5.2-B: Citation's behavior and ranking's results for a sample of recent publications that received a better ranking.

Paper ID	Publication date	PR ranking	MPR ranking	Citations over years
pub.1051273479	2015	39	26	
pub.1016190282	2016	48	22	
pub.1092148337	2017	280	34	
pub.1086050674	2017	122	79	

The previous publications are all recent, less than five years old. As shown, the ranking of these papers has improved, and now occupies better positions on the list. Looking at the citation behavior of these papers, we note that all of them share an ascending pattern of citation over the paper age. This indicates that they are in continuous growth, and dependence on them is also increasing. This explains the positive change in the ranking of these papers. Therefore, the modified algorithm is considered successful in ranking this group of papers.

5.4.2. Old Publications That Are Still Valuable:

Table 5.3-A: Citation's behavior and ranking's results for a sample of old publications that are still valuable.

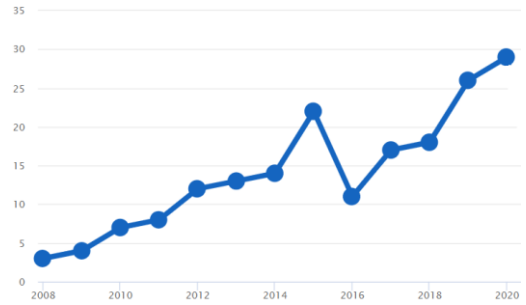
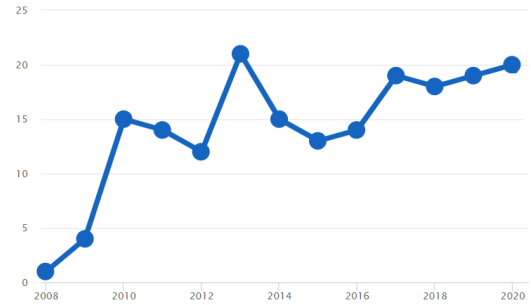
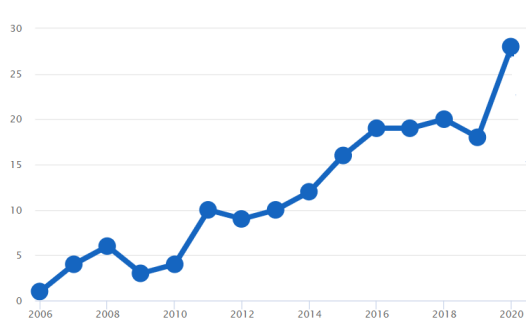
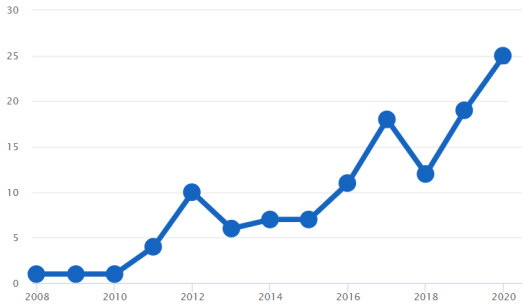
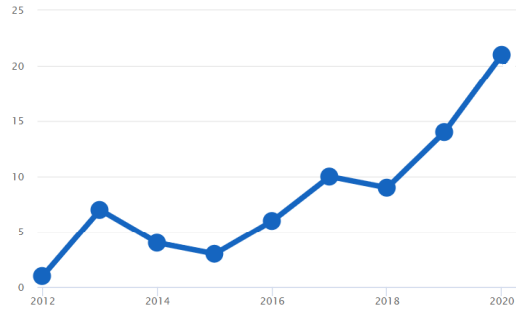
Paper ID	Publication date	PR ranking	MPR ranking	Citations over years
pub.1021827363	2007	20	15	
pub.1017361131	2008	22	24	
pub.1022680265	2005	19	16	

Table 5.3-B: Citation's behavior and ranking's results for a sample of old publications that are still valuable.

Paper ID	Publication date	PR ranking	MPR ranking	Citations over years
pub.1053343297	2008	27	27	
pub.1016603040	2010	56	57	

This group of papers has two things in common. They are old papers and all of them are still valuable. Their annual citation rate has not decreased. As it appears in the charts, they are still maintaining the same growth rate; therefore, it is not fair to underestimate their value only because their publication date is old. So, the results of the new algorithm are very close to the results of the original algorithm regarding this case of papers. It had obtained good scores using the original algorithm, and the goal here is not to use time information in a way that underestimates their value unlike the other solutions proposed.

5.4.3. Old Publications That Are Not Valuable:

Table 5.4-A: Citation's behavior and ranking's results for a sample of old publications that are not valuable.

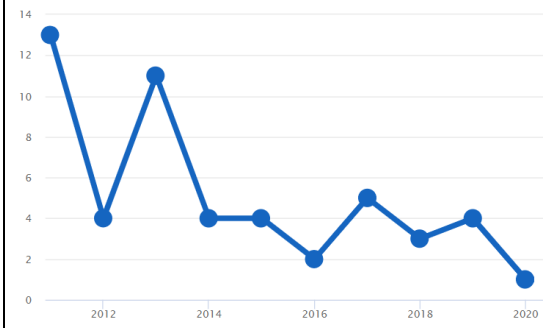
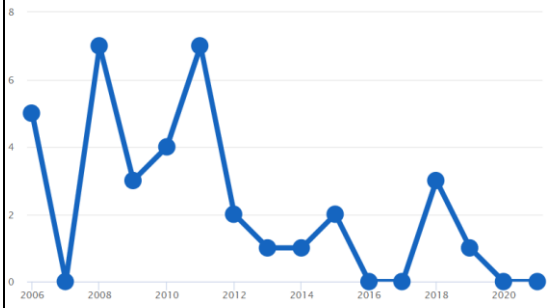
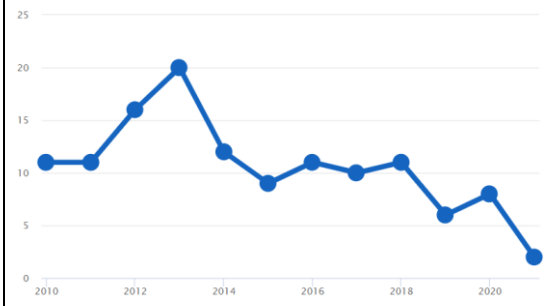
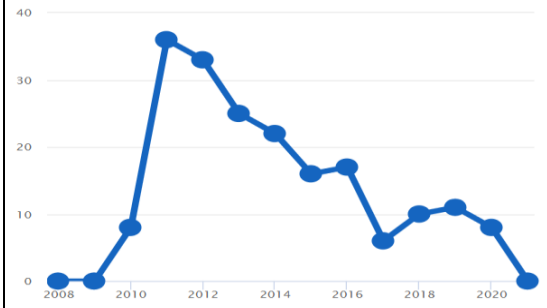
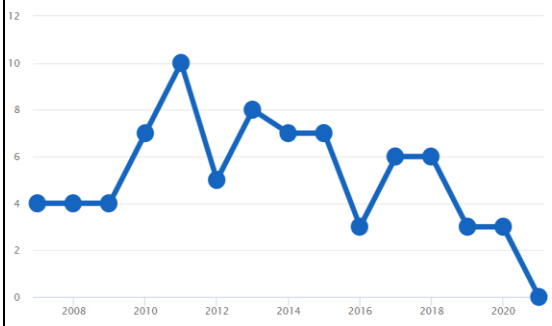
Paper ID	Publication date	PR ranking	MPR ranking	Citations over years
pub.1019598992	2010	28	37	
pub.1026762707	2006	76	101	
pub.1005766972	2009	54	72	
pub.1053137680	2008	74	95	

Table 5.4-B: Citation's behavior and ranking's results for a sample of old publications that are not valuable.

Paper ID	Publication date	PR ranking	MPR ranking	Citations over years
pub.1033194032	2005	29	41	

Also, this group of papers has two things in common, they are old papers and all of them have decreased in both their value and annual citation rate. As shown in the charts, the reliance on them is constantly decreasing, and has disappeared in some cases. Therefore, the new algorithm gives lower scores for these papers compared to the original algorithm scores. It had obtained good rankings using the original PR, because it relied on citation count in the ranking process.

Chapter 6

Conclusion and Future Works

This chapter summarizes the main conclusions and highlights for some of the future works for further improvement.

6.1 Conclusion

This thesis proposed a new modified version of the PR algorithm for scientific publications. It did so by adding a new indicator called the Citation Change Rate to the PageRank algorithm, where time information and citation data were used to calculate it. The aim was to reduce the bias in favor of old publications, which resulted from relying heavily on the citation count in the PageRank algorithm.

The results showed that the proposed ranking method was time aware. It took into account the citation occurrence time. As a result, recent publications that were still in the growth process but were continuously getting citations received better scores, even if they do not get enough time to collect large number of citations. This was because all of those citations were recent, and this was an indication that at the present time they are considered valuable and reliable papers. On the other hand, the results also showed that the old publications got fair rankings. Old publications were divided into two parts. The first part was the publications that got high scores, which based on their citation behavior, were still getting new citations until this time, and maintaining their value in the scientific community. This is the main reason why this part of old publications still had high scores. The second part was the old publications that got lower scores even though they had a large number of citations. The reason is that they were no longer getting citations as before, or not getting citations at all. Therefore, it was fair that these publications received lower scores as their values have decreased in the scientific community.

It is also worth noting that the new indicator did not cause anomalous behavior in the ranking process. There were no radical changes or unreasonable jumps in the rankings; this is an indication of the accuracy of the proposed ranking method.

6.2 Future Work

1. Testing the proposed algorithm on datasets of other journals, such as Scopus and Web of science.

2. Extending the new method to be able to rank set of publications from different fields, as the new method currently can only be applied on publications of the same field only.

3. Using other types of indicators, other than citation-based indicators, to produce more comprehensive evaluations of scientific publications for other evaluation purposes.

4. Using the proposed ranking method by one of the information retrieval systems and evaluating the ranking results of users' queries.

References

Adams, R. A. and Essex, C. (1999) *Calculus: a complete course*. 4th edn. Addison-Wesley Boston.

Alom, B. M. M. (2016) 'Web Data Mining: Views of Criminal Activities', *European Journal of Computer Science and Information Technology*, 4(4), pp. 28–40.

Le Anh, V. *et al.* (2014) 'A General Model for Mutual Ranking Systems', in Nguyen, N. T. et al. (eds) *Intelligent Information and Database Systems*. Cham: Springer International Publishing, pp. 211–220.

Arora, N. and Govilkar, S. (2016) 'Survey on Different Ranking Algorithms Along With Their Approaches', *International Journal of Computer Applications*, 135(10), pp. 1–5. doi: 10.5120/ijca2016908514.

Baeza-Yates, R. and Ribeiro-Neto, B. (2011) *Modern Information Retrieval: The Concepts and Technology behind Search*. 2nd Ed. Boston: Addison-Wesley Professional.

Barker, K. (2007) 'The UK Research Assessment Exercise: the evolution of a national research evaluation system', *Research evaluation*, 16(1), pp. 3–12.

Bienert, I. R. C. *et al.* (2015) 'Bibliometric indexes, databases and impact factors in cardiology', *Brazilian Journal of Cardiovascular Surgery*, 30(2), pp. 254–259.

Bollen, J., Rodriquez, M. A. and Van de Sompel, H. (2006) 'Journal status', *Scientometrics*, 69(3), pp. 669–687.

Bornmann, L. and Daniel, H.-D. (2007) 'Functional use of frequently and infrequently cited articles in citing publications. A content analysis of citations to articles with low and high citation counts', in *Proc 11th Int Conf Int Soc Scientometrics Informetrics. Spanish Research Council (CSIC)*. Madrid: Citeseer, pp. 149–153.

Brin, S. and Page, L. (1998) 'The anatomy of a large-scale hypertextual web search engine', *Computer networks and ISDN systems*, 30(1–7), pp. 107–117.

Buscaldi, D. (2011) *Toponym disambiguation in information retrieval*. polytechnic university of valencia.

Ceri, S. *et al.* (2013) 'An Introduction to Information Retrieval', in *Web Information Retrieval*. Berlin, Heidelberg: Springer Berlin Heidelberg, pp. 3–11. doi: 10.1007/978-3-642-39314-3_1.

Chen, P. *et al.* (2007) 'Finding scientific gems with Google's PageRank algorithm', *Journal of Informetrics*, 1(1), pp. 8–15.

Deerwester, S. *et al.* (1988) 'Improving information-retrieval with latent semantic indexing', in *Proceedings of the ASIS annual meeting*. Information today INC 143 old marlton pike, Modeford, NJ 08055-8750, pp. 36–40.

Diem, A. and Wolter, S. C. (2013) 'The use of bibliometrics to measure research performance in education sciences', *Research in higher education*, 54(1), pp. 86–114.

Digital Science (no date) *Dimensions*. Available at: https://app.dimensions.ai/discover/publication (Accessed: 25 February 2021).

Dimensions (2020) *The Next Evolution in Linked Scholarly Information*. Available at: https://www.dimensions.ai/ (Accessed: 25 March 2021).

Du, M., Bai, F. and Liu, Y. (2009) 'Paperrank: A ranking model for scientific publications', in *2009 WRI World Congress on Computer Science and Information Engineering*. IEEE, pp. 277–281.

Dunaiski, M. and Visser, W. (2012) 'Comparing paper ranking algorithms', in *Proceedings of the South African Institute for Computer Scientists and Information Technologists Conference*. Pretoria, South Africa, pp. 21–30.

Elsevier B.V. (2020) *Scopus preview - Scopus - Welcome to Scopus, Scopus Preview*. Available at: https://www.scopus.com/home.uri (Accessed: 20 February 2021).

Estrada, E. (2012) *The structure of complex networks: theory and applications*. Oxford University Press.

Furner, J. (2014) ‘The ethics of evaluative bibliometrics’, *Beyond bibliometrics: Harnessing multidimensional indicators of scholarly impact*, pp. 85–107.

Garfield, E. (1972) ‘Citation analysis as a tool in journal evaluation’, *Science*, 178(4060), pp. 471–479.

Garfield, E. (2016) *Trusted publisher-independent citation database - Web of Science Group*. Available at: https://clarivate.com/webofsciencegroup/solutions/web-of-science/ (Accessed: 20 February 2021).

Garfteld, E. (1984) ‘How to use citation analysis for faculty evaluations, and when is it relevant? Parts 1& 2’, *Es $ ay $ of an information scientisr*, pp. 354–372.

Getting Started with Dimcli — DimCli documentation (no date). Available at: https://digital-science.github.io/dimcli/getting-started.html (Accessed: 27 January 2021).

Ghosh, R. *et al.* (2011) ‘Time-aware ranking in dynamic citation networks’, in *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, pp. 373–380.

Giuffrida, C., Abramo, G. and D’Angelo, C. A. (2019) ‘Are all citations worth the same? Valuing citations by the value of the citing items’, *Journal of Informetrics*, 13(2), pp. 500–514.

Grandjean, M. (2014) ‘The knowledge is a network’, *Les cahiers du numérique*, 10(3), pp. 37–54.

Haddadene, H. A., Harik, H. and Salhi, S. (2012) ‘On the PageRank algorithm for the articles ranking’, in *Proceedings of the World Congress on Engineering*. Citeseer, pp. 4–6.

Haeffner-Cavaillon, N. and Graillot-Gak, C. (2009) ‘The use of bibliometric indicators to help peer-review assessment’, *Archivum immunologiae et therapiae experimentalis*, 57(1), p. 33.

Hiemstra, D. (2001) *Using language models for information retrieval*. Citeseer.

Hook, D. W., Porter, S. J. and Herzog, C. (2018) ‘Dimensions: Building context for search and evaluation’, *Frontiers in Research Metrics and Analytics*, 3, p. 23.

Hubble, S. (2015) *2014 Research Excellence Framework*. UK Parliament.

Jaccard, P. (1902) 'Comparative distribution of alpine flora in some regions of the western and eastern Alps', *Bulletin de la Murithienne*, (31), pp. 81–92.

Joshi, M. A. (2014) 'Bibliometric indicators for evaluating the quality of scientific publications', *The journal of contemporary dental practice*, 15(2), p. 258.

Kanellos, I. *et al.* (2019) 'Impact-based ranking of scientific publications: a survey and experimental evaluation', *IEEE Transactions on Knowledge and Data Engineering*.

Kelly, D. and Sugimoto, C. R. (2013) 'A systematic review of interactive information retrieval evaluation studies, 1967–2006', *Journal of the American Society for Information Science and Technology*, 64(4), pp. 745–770.

Kleinberg, J. M. (1999) 'Hubs, authorities, and communities', *ACM computing surveys (CSUR)*, 31(4es), pp. 5-es.

Krapivin, M. and Marchese, M. (2008) 'Focused page rank in scientific papers ranking', in *International Conference on Asian Digital Libraries*. Bali, Indonesia: Springer, pp. 144–153.

Liu, T.-Y. (2011) *Learning to rank for information retrieval*. 1st edn. Heidelberg: Springer Science.

Liu, X. *et al.* (2005) 'Co-authorship networks in the digital library research community', *Information processing & management*, 41(6), pp. 1462–1480.

Ma, N., Guan, J. and Zhao, Y. (2008) 'Bringing PageRank to the citation analysis', *Information Processing & Management*, 44(2), pp. 800–810.

Manning, C. D., Raghavan, P. and Schütze, H. (2009) 'Probabilistic information retrieval', *Introduction to Information Retrieval*, pp. 220–235.

Moed, H. F. (2006) *Citation analysis in research evaluation*. 1st edn. Netherlands: Springer Science & Business Media.

Mori, A. and Taylor, M. (2018) 'Dimensions metrics api reference & getting started', *Digital Science & Research solutions*.

Mouratidis, R. W. (2019) 'Dimensions', *Journal of the Medical Library Association: JMLA*, 107(3), p. 459.

Myers, J. L., Well, A. D. and Lorch Jr, R. F. (2013) *Research design and statistical analysis*. Routledge.

Page, L. *et al.* (1999) *The PageRank citation ranking: Bringing order to the web.* Technical Report. California: Stanford InfoLab. Available at: http://ilpubs.stanford.edu:8090/422/.

Penfield, T. *et al.* (2014) 'Assessment, evaluations, and definitions of research impact: A review', *Research evaluation*, 23(1), pp. 21–32.

Pinski, G. and Narin, F. (1976) 'Citation influence for journal aggregates of scientific publications: Theory, with application to the literature of physics', *Information processing & management*, 12(5), pp. 297–312.

Ponte, J. M. and Croft, W. B. (1998) 'A language modeling approach to information retrieval', in *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 275–281.

Robertson, S. E. *et al.* (1995) 'Okapi at TREC-3', *Nist Special Publication Sp*, 109, p. 109.

Salton, G., Wong, A. and Yang, C.-S. (1975) 'A vector space model for automatic indexing', *Communications of the ACM*, 18(11), pp. 613–620.

Sayyadi, H. and Getoor, L. (2009) 'Futurerank: Ranking scientific articles by predicting their future pagerank', in *Proceedings of the 2009 siam international conference on data mining*. Nevada: SIAM, pp. 533–544.

Shah, S. R. and Mahmood, K. (2017) 'Validation of journal impact metrics of Web of Science and Scopus', *Pakistan Journal of Information Management and Libraries*, 18(2), pp. 58–74.

Shubhankar, K., Singh, A. P. and Pudi, V. (2011) 'An efficient algorithm for topic ranking and

modeling topic evolution', in *International Conference on Database and Expert Systems Applications*. Toulouse, France: Springer Berlin Heidelberg, pp. 320–330.

Sidiropoulos, A. and Manolopoulos, Y. (2006) 'Generalized comparison of graph-based ranking algorithms for publications and authors', *Journal of Systems and Software*, 79(12), pp. 1679–1700.

Singh, A. P., Shubhankar, K. and Pudi, V. (2011) 'An efficient algorithm for ranking research papers based on citation network', in *2011 3rd Conference on data mining and optimization (DMO)*. IEEE, pp. 88–95.

Sohn, B.-S. and Jung, J. E. (2015) 'A novel ranking model for a large-scale scientific publication', *Mobile Networks and Applications*, 20(4), pp. 508–520.

Sun, Y. and Giles, C. L. (2007) 'Popularity weighted ranking for academic digital libraries', in *European Conference on Information Retrieval*. Springer, pp. 605–612.

Tang, M., Bever, J. D. and Yu, F. (2017) 'Open access increases citations of papers in ecology', *Ecosphere*, 8(7), p. e01887.

Taşkın, Z. and Al, U. (2018) 'A content-based citation analysis study based on text categorization', *Scientometrics*, 114(1), pp. 335–357.

Tsonis, A. A., Swanson, K. L. and Roebber, P. J. (2006) 'What do networks have to do with climate?', *Bulletin of the American Meteorological Society*, 87(5), pp. 585–596.

Walker, D. *et al.* (2007) 'Ranking scientific publications using a model of network traffic', *Journal of Statistical Mechanics: Theory and Experiment*, 2007(06), p. P06010.

Wang, Y., Tong, Y. and Zeng, M. (2013) 'Ranking scientific articles by exploiting citations, authors, journals, and time information', in *Proceedings of the AAAI Conference on Artificial Intelligence*. Washington.

Wei, Y. *et al.* (2021) 'An Improved PageRank Algorithm Based on Text Similarity Approach for Critical Standards Identification in Complex Standard Citation Networks', *Complexity*, 2021.

List of Appendices

1. Sample of data in Json format

"publications": [

{

"doi": "10.1002/cpe.888",

"id": "pub.1048380549",

"times_cited": 24,

"title": "Neuroscience instrumentation and distributed analysis of brain activity data: a case for eScience on global Grids",

"year": 2005

},

{

"doi": "10.1088/0031-9155/50/24/011",

"id": "pub.1059025860",

"times_cited": 67,

"title": "Monte Carlo study of Siemens PRIMUS photoneutron production.",

"year": 2005

},

{

"doi": "10.1109/tns.2005.862923",

"id": "pub.1061733339",

"times_cited": 52,

"title": "Large Size Lyso Crystals for Future High Energy Physics Experiments",

"year": 2005

},

{

"doi": "10.1016/j.icrp.2008.08.001",

"id": "pub.1040152440",

"times_cited": 87,

"title": "Editorial",

"year": 2007

},

{

"doi": "10.1088/0031-9155/52/23/017",

"id": "pub.1059026886",

"times_cited": 147,

"title": "A boundary-representation method for designing whole-body radiation dosimetry models: pregnant females at the ends of three gestational periods--RPI-P3, -P6 and -P9.",

"year": 2007

},

{

"doi": "10.1007/s00348-008-0603-4",

"id": "pub.1024315784",

"times_cited": 86,

"title": "Interactions of multiple spark-generated bubbles with phase differences",

"year": 2008

},

{

"doi": "10.1016/j.bone.2008.11.008",

"id": "pub.1047023495",

"times_cited": 38,

"title": "Reanalysis precision of 3D quantitative computed tomography (QCT) of the spine",

"year": 2008

},

```
    {
        "doi": "10.1118/1.3013555",
        "id": "pub.1017361131",
        "times_cited": 195,
        "title": "Anniversary Paper: History and status of CAD and quantitative image analysis: The role of Medical Physics and AAPM",
        "year": 2008
    },
    {
        "doi": "10.1016/j.medengphy.2010.11.013",
        "id": "pub.1015683209",
        "times_cited": 20,
        "title": "Assessment of function-graded materials as fracture fixation bone-plates under combined loading conditions using finite element modelling",
        "year": 2010
    },
    {
        "doi": "10.1088/0031-9155/56/1/016",
        "id": "pub.1059028685",
        "times_cited": 21,
        "title": "Validation of a small-animal PET simulation using GAMOS: a GEANT4-based framework.",
        "year": 2010
    },
    {
        "doi": "10.1118/1.4938097",
        "id": "pub.1016458998",
        "times_cited": 41,
        "title": "Low drive field amplitude for improved image resolution in magnetic particle imaging",
        "year": 2015
```

```
    },
    {
        "doi": "10.1016/j.nimb.2015.07.077",
        "id": "pub.1007991738",
        "times_cited": 32,
        "title": "APPA at FAIR: From fundamental to applied research",
        "year": 2015
    },
    {
        "doi": "10.17323/1998-0663.2017.4.17.28",
        "id": "pub.1103707103",
        "times_cited": 20,
        "title": "Digital economy: Conceptual architecture of a digital economic sector ecosystem",
        "year": 2017
    },
    {
        "doi": "10.1109/tkde.2017.2785824",
        "id": "pub.1099918061",
        "times_cited": 23,
        "title": "MCS-GPM: Multi-Constrained Simulation Based Graph Pattern Matching in Contextual Social Graphs",
        "year": 2017
    },
    {
        "doi": "10.1016/j.ejmp.2017.12.010",
        "id": "pub.1099726968",
        "times_cited": 20,
        "title": "A novel high-resolution 2D silicon array detector for small field dosimetry with FFF photon beams",
```

2. Python codes that used in the experiments

- Fetching the publication date of all papers and putting them in a python dictionary

```
seed = []
with open('seed.csv') as csv_file:
    csv_reader = csv.reader(csv_file, delimiter=',')
    line_count = 0
    for row in csv_reader:
        if line_count == 0:
            line_count += 1

        else:
            seed.append(row[1])
            line_count += 1


# fetch publications date
year_data = dsl.query(
    f"""search publications where id in {json.dumps(seed)} return publications[id
+year] limit 1000 """)

# get ids and associated years and put them in a python dictionary
year_of_publication = {}
for pub in year_data.publications:
    year_of_publication[pub.get('id')] = pub.get('year')
```

- Get publication date for all citing papers to determine citation occurrence time

```
data = dsl.query_iterative(
    f"""search publications where reference_ids in {json.dumps(seed)} return publ
ications[id+doi+title+year+reference_ids] """)


def build_network_dict(seed, pubs_list):
    network = {x: [] for x in seed}  # seed a dictionary
    for pub in pubs_list:
        for key in network:
            if pub.get('reference_ids') and key in pub['reference_ids']:
                network[key].append(pub['year'])
    return network



network1 = build_network_dict(seed, data.publications)
```

- Counting the number of citations that occurred each year

```
pubsRatesAverages = {}

for pub in network1.keys():

    current = year_of_publication[pub]
    years = []
    for y in range(current, 2021):
        years.append(y)

    years_counts = {}
    for yearOfCitation in years:
        years_counts[yearOfCitation] = 0

    for year in network1[pub]:
        for check_year in years:
            if year == check_year:
                years_counts[check_year] = years_counts[check_year] + 1
```

- Calculate the annual change rate

```
rates = []
    while current < 2020:
        thisYear = current
        nextYear = current + 1
        rateChangeInYear = years_counts[nextYear] - years_counts[thisYear]
        rates.append(rateChangeInYear)
        current = current + 1
    rate_average = sum(rates)/len(rates)
    pubsRatesAverages[pub] = rate_average
```

- Calculate the Modified PageRank values (PageRank with growth rate)

```
import csv

scale = 0.00001

# convert pagerank.csv to a python dictionary
pageRank = {}
with open('pagerank.csv') as csv_file:
    csv_reader = csv.reader(csv_file, delimiter=',')
    for row in csv_reader:
        pageRank[row[0]] = float((row[1]))

# convert rates.csv to a python dictionary
rates = {}
with open('rates.csv') as csv_file:
    csv_reader = csv.reader(csv_file, delimiter=',')
    for row in csv_reader:
        rates[row[0]] = round(float((row[1])), 2)

# find the new ranking which is modified by rates through adding rate muliplied w
ith a scale
m = {}
for i in pageRank.keys():
    for ii in rates.keys():
        if i == ii:
            mpr = pageRank[i] + scale * rates[i]
            m[i] = [pageRank[i], mpr]


# replace rates with the new ranking dic
with open('pageWithMod_WithoutMod.csv', 'w') as f:
    for i in m.keys():
        f.write("%s,%s,%s\n" % (i, m[i][0], m[i][1]))
```

3. Sample of the PR results (Top 30 papers)

Paper id	PR Score	Publication year
pub.1021217544	0.000428	2006
pub.1043135063	0.000398	2005
pub.1007569283	0.000377	2006
pub.1039521816	0.000354	2007
pub.1034014743	0.000314	2017
pub.1000200902	0.000313	2015
pub.1026580371	0.000306	2011
pub.1041650890	0.000298	2010
pub.1059030439	0.000298	2014
pub.1005766416	0.000288	2008
pub.1002643727	0.00028	2011
pub.1060839701	0.000278	2010
pub.1040551777	0.000275	2005
pub.1027347577	0.000275	2014
pub.1009553961	0.000261	2007
pub.1050018526	0.000257	2007
pub.1050860314	0.000256	2006
pub.1012076104	0.000252	2006
pub.1022680265	0.000249	2005
pub.1021827363	0.000249	2007
pub.1046563478	0.000241	2005
pub.1017361131	0.000229	2008
pub.1031759624	0.000229	2008
pub.1059301220	0.000224	2011
pub.1000621610	0.000222	2013
pub.1061584103	0.000219	2005
pub.1053343297	0.000218	2008
pub.1019598992	0.000212	2010
pub.1033194032	0.000207	2005
pub.1093176203	0.000206	2008

4. Sample of the MPR results (Top 30 papers)

Paper id	MPR Score	Publication year
pub.1021217544	0.0004416	2006
pub.1034014743	0.000404	2017
pub.1043135063	0.0004013	2005
pub.1007569283	0.0003877	2006
pub.1039521816	0.0003671	2007
pub.1059030439	0.0003663	2014
pub.1000200902	0.000351	2015
pub.1027347577	0.00033	2014
pub.1026580371	0.0003149	2011
pub.1041650890	0.000307	2010
pub.1002643727	0.0002922	2011
pub.1060839701	0.00029	2010
pub.1005766416	0.0002863	2008
pub.1040551777	0.0002777	2005
pub.1021827363	0.0002728	2007
pub.1022680265	0.000267	2005
pub.1009553961	0.0002664	2007
pub.1049556687	0.000264	2015
pub.1050018526	0.0002616	2007
pub.1050860314	0.0002589	2006
pub.1012076104	0.0002563	2006
pub.1016190282	0.0002545	2016
pub.1046563478	0.000247	2005
pub.1017361131	0.0002457	2008
pub.1000621610	0.000242	2013
pub.1051273479	0.000242	2015
pub.1053343297	0.000238	2008
pub.1059301220	0.0002362	2011
pub.1031759624	0.0002315	2008
pub.1051545459	0.000226	2015