Journal of Theoretical and Applied Information Technology
31st January 2023. Vol.101. No 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

# PROPOSED APPROACH FOR ACADEMIC PAPER RANKING BASED ON BIG DATA AND GRAPH ANALYTICS

**NAGWA YASEEN HEGAZY [1], MOHAMED HELMLY KHAFAGY [2], AYMAN E.KHDER [3]**

[1] Ph.D. Candidate, Faculty of Computers & Information, Department of Information System, Fayoum University, Cairo, Egypt.

[1] Lecturer Assistant faculty of computer science, Canadian International College, Cairo, Egypt

[2] Professor Department of Computer Science, Faculty of Computers & Information, Fayoum University, Fayoum, Egypt.

[3] Professor of Information Systems and Data Science, Department of Information Systems, Faculty of Computers and Information Technology, Future University in Egypt (FUE).

E-mail: [1]ny1158@fayoum.edu.eg, [1]nagwa.yaseennm@gmail.com, [1]nagwa_y_mohamed@cic-cairo.com, [2]Mhk00@fayoum.edu.eg, [3]ayman.khedr@fue.edu.eg

**ABSTRACT**

The outburst growth of technology in the academic environment and the widespread use of digital libraries have generated big scholarly data. Ranking and measuring the impact of academic papers grants higher importance to the academic environment that is required for promotions, hiring, awards, grants, scholarships, and ranking university procedures. Google Scholar ranking depends mainly on the citation count of academic papers; therefore, some papers are ranked low even if they are qualified papers. Identifying the most important articles in the field is considered a critical issue for researchers, journals, and academic institutions. The goal of this study is to create a ranking system for big scholarly data (RBSD) that integrates network analysis based on graph analytics, citation analysis, and similarity between papers. The proposed model ranks papers based on the paper citation network to get the central papers. It also ranks authors to identify the top authors in the computer science citation network and analyzes the similarity between academic papers to get the relevancy between papers. A new methodology is proposed to rank papers based on a weighted score that considers paper information, author information, and publication venue information. The proposed model also considers the complex relationship between papers, overcoming the limitations of other ranking systems that rely only on the traditional PageRank algorithm. To produce a more accurate ranking system, our suggested model excludes authors' self-citation and collaboration citations, which are often used by authors to increase their citation count. To evaluate the RBSD model, four real-world datasets were used: ACM, MAG, DBLP, and Scopus Elsevier, for publication venue information. The proposed model was applied to 2,092,356 papers, with 8,024,869 citations. This was implemented using Apache Spark Graphx to accelerate the execution time for graph analysis and to explore the nature of scholarly data. The experimental results show that our proposed model outperformed the Google Scholar Ranking procedure based citation count and returns reasonable results.

**Keywords:** *Scholarly Data, Big Data, Graph Theory, Citation Analysis, Ranking Systems, Bibliographic Coupling, Co-Citations.*

## 1. INTRODUCTION

Academic paper ranking system is an essential and challenging task for academic universities and institutions. It has also begun attracting greater attention in the academic environment and recommendation systems. Academic members' promotions, awards, hiring, and scholarship procedures depend mainly on evaluating their scholars [1], [2]. Publication venue information that authors publishing their research on it has an important impact on the evaluation of their academic papers. Prestigious publication venues led to a higher weight for academic papers than the lower-ranked venue. Finding an accurate ranking system that

considers the most important criteria for academic paper ranking is required for researchers and academic institutions. Previous ranking systems have focused on ranking papers based on the traditional PageRank algorithm [3], [1], which has many limitations related to implementation complexities with a limited number of data. Also, it only relays on citation count when ranking papers. To address these issues, updated PageRank for graph distributed processing has been used. In addition, updated PageRank for parallel processing can help solve the implementation complexities. However, Focusing on the citation count when ranking papers led to unfair rank for recent and qualified papers.

This study proposes a different ranking system that considers the quality of authors and publication venue information in academic paper ranking. It also considers the similarity of papers in ranking to get an accurate academic ranking system for papers. The proposed approach aims to overcome the drawbacks of previous ranking systems that mainly depend on PageRank, which ranks papers based on the citation count for a paper without any concern for authors, journals, and paper similarity information. Citation count is considered an inaccurate base for ranking because it includes self-citations and collaborator citations. In our model, we consider calculating an accurate citation count for papers and authors after removing self-and collaborators citations, author ranking based on the modified citation count using a modified version of PageRank algorithm, calculating weight for each paper that consider authors, journals, similarity information, and rank paper according to paper weight.

## 2. BACKGROUND

The size of Academic information generated around the World Wide Web exceeds millions of papers documented, which represents a challenge for Ranking systems, information retrieval, and big data. Considering the relevancy of the paper in ranking Big Scholarly Data (BSD) based on retrieval criteria, providing qualified ranked documents in traditional methods is considered time-consuming due to the complexity of data and its relations. The volume of research papers produced by the academic environment, digital libraries, and academic social networks is known as Big Scholarly Data (BSD). Millions of articles, authors, co-authors, citations, journals, conferences, and the intricate links between them the citation Network can represent are all part of big scholarly data [3][4]. Big scholarly data has been overgrowing due to the digital transformation in technology that makes authors publish their papers and share them via digital libraries and academic networks [5], [4].

BSD is growing vastly due to academic social networks [6] and digital libraries [4] such as ResearchGate. M.khabsa et al.[7] Estimates that the number of scholarly documents published on the public web in 2014 is 114 million scholarly documents available on the public web. In 2022, an academic social network such as ResearchGate announced on its website that they have more than 135 million publications, more than 17 million authors, and 700,000 research projects on its network. Microsoft academic in 2020 have a 241,170,095 publication document, 244,552,188 authors available on their network [8]. The semantic scholar has 206,138,656 papers for all fields of science [9].Big data provide an ability to manage, process, and explore a large volume of data [10],[11],[12],[62],[63]. Due to the increasing size of big scholarly data, a suitable approach is needed to analyze this complex scientific data. Big data challenges represented in the five V's make BSD an essential and vital research topic [13]. 'Volume' refers to the continuously growing volume of scholarly data that involves millions of authors, citations, figures, and metadata. The term 'Velocity' refers to the fast rate at which scholarly data is generated [12] with hundreds and thousands per day and submitted to journals and digital libraries. According to Feng et al. [5], the growing rate average of scholarly data generated in 2016 is 6.3% per year. 'Variety' refers to the variety of scholarly data entities [14] and the complex relationship between them that make it challenging to analyze and explore. 'Veracity' refers to the accuracy and quality of data that can represent scholarly data in author name duplication and disambiguation, which has a higher impact on the analysis results. The 'value' indicates the ability to gain essential knowledge and insights from big scholarly data gathered and combined from different data sources [5], [13].

Researchers always search for papers related to their field that satisfies their information need. However, finding the most important and relevant paper in the field is considered difficult and time-consuming, especially with the available millions of articles on the internet and in digital libraries. Information Retrieval (IR) is a process of helping users to find a document that satisfies the user information need from a vast collection of data [15]. PageRank algorithm is an IR algorithm used to rank webpages according to their importance [16]. PageRank is a popular graph processing algorithm

that works well on graph analysis and is mostly used in ranking web pages[17],[8].

Big scholarly data is represented in the term of citation network; citation network is a graph representation of the relationship between literature citations for the connection between papers as a cite-in and cite-out. Cite-in for paper represents the incoming citations links from other papers called in graph theory in-degree. Cite-out; in citation network represents the outgoing citations for other papers that are the out-degree in graph theory. Citation network connections are used to calculate the relatedness among academic papers; two standard similarity measures are used to get the similarity between scholarly data based on the literature; bibliographic coupling and co-citation. Bibliographic coupling measures the similarity of two papers that cite the same paper, whereas co-citation measures the similarity of two papers that are cited in the same document[18],[19].

Due to the complexity of analyzing big scholarly data, the most suitable way is to use graph analysis based on graph theory. Graph theory; is a representation of data objects and their relationship between them; in the concepts of graph vertex and edge. Each data object; is represented as graph vertices or nodes, and graph edges are the relationship between vertices. Graph analysis is a suitable method to store, analyze and query big data; that has a complex interconnected relationship [20]. Graphical modeling for big data can help to understand and explore the complex relationships between data that cannot be analyzed using traditional methods due to its complex nature [21].

The advantage of using a graph is to explore and uncover the structural relationship in scholarly data and citation networks. Graph data preprocessing techniques include loading, transforming, and filtering data. It also includes graph creation, analysis, and post-processing. Several frameworks work with graph data preprocessing, among them GraphX, built on top of apache spark. Apache spark is a powerful tool for big data analytics that provides parallel processing for big scholarly data using a cluster computing system through two types of operations transformation and actions [22],[10],[23]. Apache spark combines two API components. The first low-level API includes RDD resilient distributed dataset, which is the main data structure for Apache spark. In addition, Apache spark provides fault tolerance and in-memory caching. The second component is structured APIs for a dataset, Dataframe, and SQL [24]. Spark GraphX is a spark component built on top of the spark that integrates ETL processes and data exploration through graph parallel computations in a single system [23], [25]

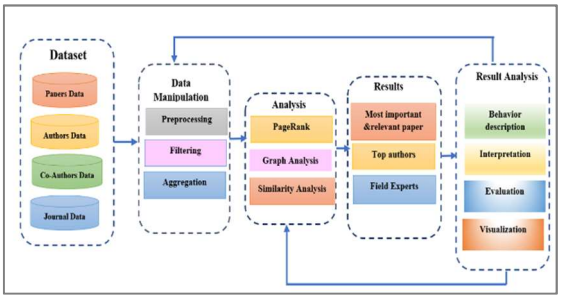


*Figure 1: Proposed Framework for ranking Big Scholarly Data Analysis*

based on graphframe. All graph algorithms are provided through Graphframe's Python, Scala, and Java APIs, which also assist users in building queries using Spark SQL APIs and dataframes. Graphframe also supports writing and reading graphs supported by dataframe[26],[27].

The proposed ranking model is constructed based on the citation network analysis to determine the most important papers and authors in the academic network, which helps to determine the top authors and top papers. The main advantage of any ranking system is to help academic members in the research field by helping them to reach the top qualified papers in their research field and help academic institutions in the promotions, awards, and scholarship procedures on fair biases. It also increases researchers' awareness to select the qualified references that let them output qualified work and give them insights regarding the important features which led to a higher rank of their papers. Figure 1 represents the proposed framework for a big scholarly data ranking system.

## 3. LITERATURE REVIEW

There are many studies conducted on ranking systems. Most of them are based on search engine ranking techniques that are built based on various ranking algorithms like PageRank, CiteRank, YetRank, and NewRank [28],[29],[1]. These studies rely on the approach of internet web pages, but it has complexities with scholarly ranking due to the interrelating relationship between various entities. Several other techniques focused on recommender systems to retrieve the most relevant papers for researchers such as [30],[31],[32]. Finally, other studies studies focused on self-citation analysis such as [33],[34] .

### 3.1 Studies Relaying On Ranking Systems For Scholarly Data

More studies construct ranking systems to rank academic papers; studies rely on PageRank depending on the citation count to rank papers. Recent papers consistently ranked low, even if it is qualified papers or eminent literature, regardless of the prestigious venue that was published on it. The implementation complexity for PageRank algorithm when working with a large number of data also a drawback for the algorithm. Applying PageRank algorithm through distributed computing and graph PageRank helps solve this problem with time and space complexity. CiteRank algorithm [58] has emerged based on the idea of PageRank algorithm to overcome the limitations of PageRank algorithm by considering the publication date of an academic paper [3], [16], [35]. PageRank algorithm is the foundation for Google's search engine Recommendations [16], [35]. Google's search engine recommendations are also built based on the PageRank algorithm. In addition, Google Scholar uses PageRank, which offers suggestions for all connected publications and gives each work a citation score[36],[37].

Dunaiski et al. [28] have suggested the CiteRank algorithm, which is built based on the PageRank algorithm's principle, to address the issues with PageRank. CiteRank algorithm considers the paper's publication date and the aging effect in the citation network, although this technique has issues because of time and space complexity. The YetRank algorithm was created by Hwang et al. [38] to address PageRank and CiteRank issues by taking the impact factor of the publication venue into account. The research published in a reputable journal receives a higher ranking than the paper published in a journal with a lesser reputation, but the algorithm yielded a complexity to calculate the impact factor for each publication venue per year.

T. Abdeltief. et al. [2] developed a model using the Fair Paper Ranking algorithm (FPRT), which attempts to address issues with earlier ranking techniques by considering seven critical factors into account. The seven factors that the model uses are; authors number in paper, publication year, H-index of authors, citation score, journal impact factor, and paper field and the maximum impact factor value of paper field. This study also developed a normalized impact factor to eliminate the impact factor gap between various scientific fields. However, it ignores the structural relationships in the citation network; the model also does not consider the journal information quartile and SJR, which is an essential criterion for paper ranking. The model takes the author's information into account based solely on the author's H-index, regardless of the author's rank. T. Abdeltief. Model designed to deal with a limited number of data.

An author and journal ranking model was created by M. Rathor et al. [40]. The purpose of Rathors' model is to suggest an appropriate reviewer for the submitted work based on that reviewer's field expertise. A modified version of the page rank algorithm was used to implement the journal and author ranking, and it can take self-citations from authors and journals into account. The model skipped author quality indicators like the author's h-index and the quality of his publications because author ranking depends on node weight and the number of publications the author has.

An alternative citation context article influence ranking methodology was developed by Chen et al. [41] in 2019 to overcome information redundancy in the semantic vector space and improve article retrieval. Chen converts the context of an article citation to a word vector representation using the word2vector model and natural language processing techniques. The model's objective is to enable users the capability to understand, compare article ranking outcomes, and enable them to investigate their desired paper influence. Unfortunately, other major elements, including the authors, journal, and publication venue's impact factor, have not been taken into account by the Chen model, which also neglected the influential article pattern that aids in analyzing the influence of work. The ranking systems, algorithms, benefits, and drawbacks for each technique are compared in Table 1.

### 3.2 Studies Relaying On Recommendation Systems For Scholarly Data

All academic researchers today have to deal with the growing number of research publications, conferences, journals, proceedings, white papers, and others. These complex relations make it difficult for researchers to find the research paper they are looking for quickly due to the information explosion, which leads to a waste of time. Big scholarly data recommendation systems can assist scholars in information filtration and finding the ideal publication for their academic research.

Da. Zhang et al. [30] proposed an approach that employs a distributed infrastructure for hardware and software to analyze large amounts of scholarly data. The system's objective was to identify relationships between papers and authors to recommend citations, identify prospective collaborators, and recommend an article to a suitable publication venue and an expert Reviewer. Da.

Zhang et al. model used a mixed and weighted metapath (MWMP) to investigate the interaction between entities in order to achieve the system goals. The shortcomings of the Da. Zhang model does not take into account the structural relationship between entities, and they also do not consider neighbor information for each object.

Network (MSCN) that aims to build a recommendation system. The approach used by Jieun incorporates citation analysis, content analysis, content filtering, and collaborative filtering techniques. The concept behind citation analysis is to examine the links that are directly cited in or mentioned by other papers for the multilevel

*Table 1: The Advantages and Limitations of Scholarly Data Ranking Systems*

| Author | Used Technique | Advantage | Disadvantage |
|---|---|---|---|
| Liua,Hasani, fiala [21],[9],[22] | PageRank algorithm | -Uses importance and relevance to rank web pages.<br>- Used to determine the central webpage in the network.<br>- Performs effectively with online web pages.<br>-Utilized by Google Scholar to produce a list of all relevant articles. | -Performs effectively with online web pages but has many complexities with ranking papers.<br>-It has a lot of constraints when working on citation network with its different entities.<br>- Rank papers based on the number of citations.<br>-Recent papers have consistently received low Rank. |
| Dunaiski, Bonchi [20] ,[30] | CiteRank Algorithm | - Based on the PageRank algorithm's principle but taking into account the ageing of the citation network.<br>- The publication date of paper has considered. | - Time and space complexity.<br>- More expensive. |
| Hwang, [31] | YetRank | - Take into account the publication venue's impact factor.<br>- Gives published papers in prestigious venues a higher ranking than those in lesser-known venues. | - Calculating the impact factors for each journal and year involves a lot of time and space complexities.<br>- Does not consider authors information and citation Relationship. |
| Dunaiski, [20] | NewRank Algorithm | - It significantly enhances CiteRank issues<br>- Improve the paper's initial value based on its reference list. | - Does not take into account the age of referencing articles, the impact factor of the publication venue, or the author's h-index.<br>- Citations from widely read articles should be given greater weight than citations from less significant articles. |
| T.abdelatief [32] | FPRT Algorithm | - Take into account the following seven variables: the total number of authors, publication year, author h-index, citation score, journal impact factor, paper field, and the highest possible impact factor value in the paper field.<br>- It is based on three factors, the normalized impact factor, the average h-index, and the citations factor. | - The algorithm performs well when dealing with large amounts of data, however it is not applied to enough huge amounts of academic data.<br>- Does not take the citation network's structural relationship into account. |
| M.Rathor [33] | Modified version of page rank algorithm & proposed new impact factor | - Use author ranking to solve issues with the traditional PageRank algorithm.<br>- Assist in selecting subject-matter specialists.<br>- A new impact factor that exclude self-citations from authors and publications is being proposed.<br>- Allow for the avoidance of conflicts of interest. | - The quality of authors does not considered authors indicators such as author H-index.<br>- The ranking of authors depends on the biases of the expert's field after taking the node weight and the number of publications into consideration. |
| Chen et al. [34] | Visual analysis VAIR for citation analysis & SPEAR model for ranking. | - The model analyze article context.<br>- Improve article retrieval.<br>- Reduce information redundancy. | - Impact of paper has been neglected.<br>- Ignore the publication venue information and authors information<br>- The model not suitable to deal with the massive size of BSD. |

Jieun Son et al. [32] proposed a novel technique for a Multilevel Simultaneous Citation network. The implementation of content analysis makes use of a keyword-matching procedure.

In Jieun's approach, candidate papers are chosen after computing the candidate score for each candidate paper and creating a multilevel citation network.

Jieun's model needs to be improved to handle big scholarly data because it was developed using a small amount of data. The limitations of this model include the candidate's scores ignoring information for authors and journals.

### 3.3 Studies Relaying On Self-Citation Analysis For Scholarly Data

Faiz et al.[33] proposed a system for self-citation analysis for google scholar winner author's data to evaluate authors fairly. Faiz system considers removing self-journal citations from google scholar data. Faiz's model found that the rate of author self-citation is 2.86%, co-author citations are 3.33%, and the rate of journal self-citation is 3.95%.

According to Yurko et el. [34], a study for self-citation analysis found that self-citation is considered good practice if it is at an acceptable rate, but if it is at an excessive rate, it will be a bad practice and lead to unfair evaluation. Therefore, according to the literature [34], [33], self-citation and collaborators citations are essential aspects to consider in ranking and evaluating authors.

Finally, it has been noticed that there are many limitations and challenges for previous literature, which enables us to understand that all prior ranking systems did not consider the quality of authors, the article's influential features, and the significance of academic papers. Limitations with traditional PageRank algorithm that can be solved using distributed computing and graph PageRank; help in solving this problem with time and space complexity. PageRank measures the centrality of paper compared to other papers in the network, and centrality is a measurement for potential impact ranking, not a measurement for actual impact. Due to the importance of citation count in paper and author ranking, we cannot ignore it, so we will use it as an initial step to determine the centrality rank based on citation count. To get the qualified ranking of a paper, we consider authors' information and journal information to get a weight for each paper and rank it according to the proposed weight score.

On the other hand, every ranking system must take into account the information of the journal that publishes the academic article because it has an impact; on the paper's quality and ranking. In addition, the complexity of massive scholarly data, which represents time and space complexity for most ranking algorithms, is another major problem. Finally, the ranking system should consider self-author citations and collaborator citations that authors use to inflate their citation count.

Our work aims to develop a new approach that can handle enormous amounts of scholarly data through distributed computing and parallel processing. Ranking academic papers based on paper information, author's quality, publication venue information, and paper relevancy.

## 4. PROPOSED MODEL

The proposed approach helps academic members and institutions to uncover the most

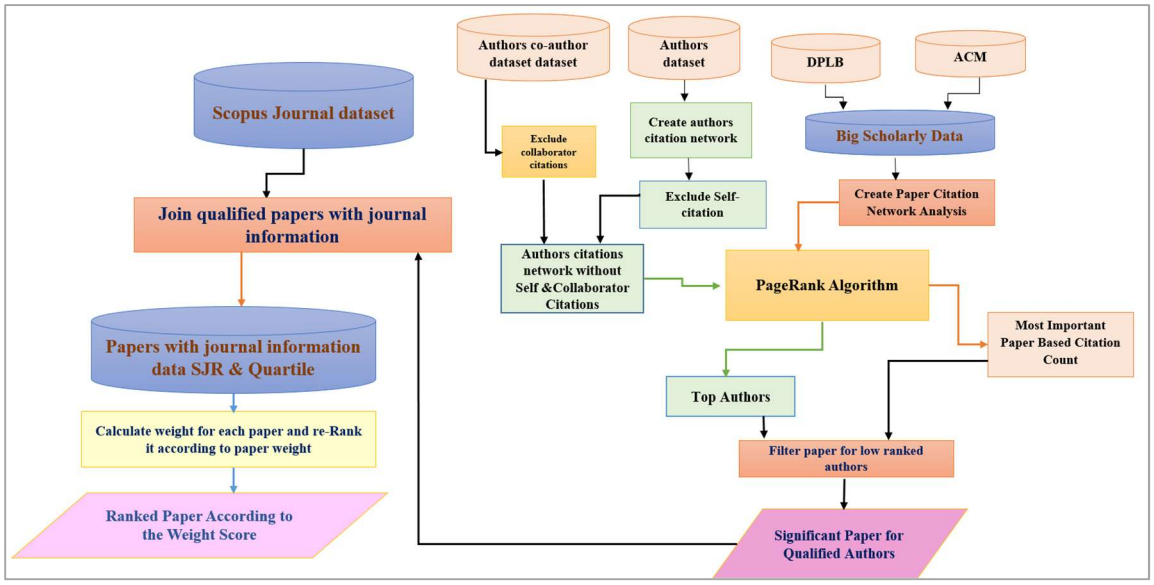


*Figure 2: Ranking System for Big Scholarly Data (RSBSD)*

Journal of Theoretical and Applied Information Technology
31st January 2023. Vol.101. No 2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

**Procedure-1: Calculate significant papers for qualified authors**

**Input**: source of papers& authors: DBLP, ACM.
**Output**: Ranked list of significant Papers P.Q

```
From sources:
Construct Graph.Paper = (Vp, Ep)
    Δ Paper Citation Network. Vp: set of Papers. Ep: Paper Citation
Construct Graph.Authors = (VA, EA)
    Δ Author Citation Network.
    VA: set of authors. EA: author citation
For each  e ∈ EA :
  If Srce = dste or  srce& dste are collaborator:
  Remove e from    EA      Δ Remove self-citation & collaborator citations
PR-A = [ ]              Δ Initialize list of PageRank values for authors
P-Q = [ ]               Δ Initialize list of significant papers
For each P ∈  Vp
      Calculate PR(P)              Δ Calculate PageRank
      Append P & PR(P) to P-Q
For each  a ∈  VA
            Calculate PR(a)
            Append  a & PR(a) to PR-A
From P-Q: filter papers with low Ranked PR(a) in PR-A
Return P-Q
```

**Procedure-2: Rank Papers according to weight score**

**Input**: Ranked list of significant papers P-Q.
Scopus Journal Dataset

**Output**: Ranked Papers according to weight score.

```
JP ◄— Join significant papers according to P-Q with journal information
    Δ Get Journal information SJR & Quartile
For Each P ∈  JR:
    Calculate Bibliographic coupling (B.B) for paper P
    Calculate Co-citation(Co.c) for paper P
    Calculate the Distance (D) for paper P
    W(P)=(B.B +Co.C +J.SJR +J.Quartile+author.H_index)/D
Rank papers according to W(p).
Return Ranked Papers according to weight score.
```

influential criteria for a qualified, accurate ranking system. It also allows researchers to get the most important paper related to their field and determine the experts. Furthermore, the proposed system helps the academic environment to evaluate the academic paper that is considered the core criteria of the award, promotions, and scholarship procedures in

academic institutions. The main goal of the proposed approach is to get a qualified ranking system that considers citation network analysis and explores the complex and structural relationship between papers, authors, coauthors, and journals. The proposed approach combines the citation analysis for big scholarly data, network analysis, and similarity analysis between papers to get a qualified ranking system for academic papers based on analysis of paper information, author information, relevancy of paper, and publication venue information.

The proposed model uses big data analysis tools and techniques, which include graph analysis that helps to explore and uncover the structural and complex relationship between papers using Apache Spark Graphx due to its ability to provide distributed processing through a dataflow framework that accelerates the execution time for graph algorithms on the paper references network, authors network after that, we use the similarity analysis measures to determine the most relevant papers.

There are many steps in our proposed model, which are illustrated in figure 2. These steps will be categorized and discussed in three phases;

The first phase focuses on paper, author, and co-author data analysis. There are two steps in this phase: The first step focuses on developing and examining paper citation network analysis. The second step; analyzing the author's citation network to explore the structural relationship in both networks.

The second phase combines four steps; first, PageRank for paper ranking based citation count to identify the most important papers based citation count. Second, rank authors used PageRank algorithm after removing self-citation and collaborators citations to obtain a fair rank for authors and identify the top authors in the author citation network.

Third, merging two datasets for the most significant papers with the top authors includes joining two datasets for the most significant papers features with the feature set of top authors. Fourth, filtering papers for low-ranked authors to get the most significant papers for qualified authors.

| paperKey | title | publication | refrenceNo |
|---|---|---|---|
| 10 | * The Three-Machi... | c Journal of the ... | [289259] |
| 12 | * Space-Time Trad... | c Journal of the ... | [289824, 408638, ... |
| 13 | * The VLSI Comple... | c Journal of the ... | [769, 289452, 319... |
| 47 | * Matings in matr... | c Communications ... | [25470, 289292, 2... |
| 53 | * A class of gene... | c ACM Transaction... | [317871, 598103, ... |
| 54 | * Static grouping... | c ACM Transaction... | [103, 288713, 317... |
| 86 | * Best approximat... | c ACM Transaction... | [317819, 320076] |
| 88 | * Efficient polyg... | c ACM Transaction... | [3104, 4093, 6028... |
| 116 | * Centralized ver... | c ACM Computing S... | [317884, 317895, ... |
| 117 | * Principles of t... | c ACM Computing S... | [2008, 287377, 28... |

*Figure 3: Graph Vertex Data from Apache Spark*

The third phase includes a joined feature set for the essential papers based on citation count for qualified authors with its publication venue

information to obtain a feature set for papers with journal information. This phase aims to examine paper similarity using similarity metrics for citation networks, then calculate the weight of each article to rank the paper according to its weight score based on the relevancy of the paper as in our proposed model procedure-1 and 2.

```
edges.show(10)
+---+------+------------+
|src|   dst|relationship|
+---+------+------------+
| 10|289259|       cited|
| 12|289024|       cited|
| 12|408638|       cited|
| 12|600828|       cited|
| 12|688[97|       cited|
| 13|   769|       cited|
| 13|289452|       cited|
| 13|319821|       cited|
| 13|408343|       cited|
| 13|598672|       cited|
+---+------+------------+
```

*Figure 4: Graph Edges Data from Apache Spark Graphx Using GraphFrame*

### 4.1 Proposed Model First Phase: Identify Central Papers And Important Authors In The Network

Our suggested approach's initial phase consists of various steps. The first step of our proposed RBSD model is to build citation networks for scholarly data; the second step is to construct an author citation network.

#### 4.1.1 Creating Paper Citation Network

The first step of our proposed model is to create a Paper citation network using Apache Spark GraphX based on the principles of graph theory, where each node or Vertex represents a paper, and graph edges represent the relationship between papers. Graph edges are the links connecting papers that are referred to as "cites." Each graph vertex contains paper information from our first data file. The first data file includes the Paper index, authors, affiliations, year, publication venue, reference id, and abstract. Graph property consists of two data files, Vertex and edges data as in figure 3 for vertices data and figure 4 for edges data.

As shown in Figure 3, the reference number column is represented in a list, so it needs some processing to get the list items of data to convert it into a record for each item to be suitable input for graph property data. Figure 4 represents a screenshot for edges date from Apache Spark GraphX, which consists of the source, destination, and Relationship. The source of edges data is the paper key, and the destination is the reference number. The property of graph operators includes vertex Indegree, outdegree, and vertex degree have been used to calculate the citation count for each paper where Indegree represents the cite-In and outdegree is the cite-out for each paper, as illustrated in Figure 5.

#### 4.1.2 Creating Author Citation Network

In this step, we have created an author's citation network utilizing the ideas of graph theory. In the author's citations network, each node represents an author, edges represent the citation relationship between authors as author A cites author B, and the edges score is the number of citations between them. Analyzing the citation network for authors is used to explore the relationship between authors, which helps to identify field experts. In the authors citation network, we are dealing with a directed graph where if author A cites author B three times, it does not mean author B has the same cite out score toward author A. If the score of citations between two authors is higher, this demonstrates that

```
+------+--------+ +------+---------+
|    id|inDegree| |    id|outDegree|
+------+--------+ +------+---------+
|686946|       4| | 17714|        1|
|289164|       1| | 20158|        3|
|408279|       3| | 20219|        1|
|320444|      12| | 20569|        1|
|289069|       7| | 20868|        1|
|  2069|       5| | 23318|       28|
|597991|      11| | 23918|       10|
|  2294|       7| | 25032|        4|
|374794|      12| | 27264|        4|
|323093|       6| | 29573|        2|
|318026|      20| |101205|        3|
|692974|       1| |102745|       11|
|599550|      16| |102944|       10|
```

*Figure 5: Graph Indegree and Outdegree that represents Cite-In and Cite-Out for each Paper*

they are works in the same scientific research field. In our proposed RBSD model, we have used in-degree and out-degree to determine each author's income and outgoing citations. In the author's graph, the vertex, which has a loop relationship, represents self-referencing, and the edge score is the number of times the author cites his works.

### 4.2 Proposed Model Second Phase: Determine The Most Important Papers For Top Authors.

The goal of this phase; is to identify the most significant and central papers based on citation count with its qualified top authors. To identify the top authors, self and collaborator's citations have been excluded from the author's dataset, which authors utilized to inflate their citation count.

The first step of this phase; is ranking papers using PageRank algorithm that ranks papers based on citation count. Ranking authors using PageRank algorithm after removing self-citation and collaborator's citations to obtain a fair rank for authors is the second step. The third step includes

merging two datasets for the significant papers with the qualified top authors. Finally, in the Fourth step, we filter papers for low-ranked authors to get the most significant papers for qualified authors.

#### 4.2.1 PageRank: Paper Initial Rank Based on Citation Count (Identify Central Papers in the Network)

In this step, we are applying PageRank algorithm for paper data to rank papers based on their citation count according to their importance in the citation network. The rank of papers is considered the initial rank to determine paper centrality. PageRank is a graph-based ranking system that evaluates both a vertex's inbound and outgoing links to assess the vertex's relevance within the graph[16],[5]. The algorithm estimates the vertex importance in a graph by supposing that an edge from paper A to paper B signifies B's approval of vertex A's significance. PageRank algorithm is used to find the influencers in any network, so it can be used in paper citation networks, academic networks, and social media analysis. In Big Scholarly Data, PageRank is used to assess an article's authority before ranking it according to the number of citations it has received, which is regarded as an indication of a high-quality paper. Higher PageRank values indicate that a paper is significant and has received more citations.

In This step, graph analysis theory has applied to paper datasets to process with PageRank as a graph centrality and importance analysis algorithm. The source for PageRank algorithm is a paper key, a destination is a reference number, and the relationship between source and destination is "cited" as represented in figure 4. The output of this step is ranked papers with PageRank values to determine the most significant papers.

#### 4.2.2 Author Ranking

In authors ranking, we have made two pre-steps for PageRank algorithm; these are excluding self-author citations and removing collaborator citations as represented in figure 6.

**A. Exclude Self Citation**

Self-citation is a method used by authors to cite their papers. It is made by authors in a way to inflate their citation count. Self-citations are typically identified as references or citations that are made to another publication produced by the same authors or author groups. The underlying reasons or motivations for self-citations have been the subject of investigations. For instance, authors may self-cite their work as a way to increase their visibility in their field of science, or they may self-cite regularly out of egotism. Since citation counts are frequently used to gauge the quality of scholarly outputs, self-citation is occasionally viewed with suspicion, if not outright manipulation [42].

In our proposed method, we have removed self-author citations to get a fair rank for authors. We have used the relationship between the paper data file, author's dataset, and paper2author dataset,

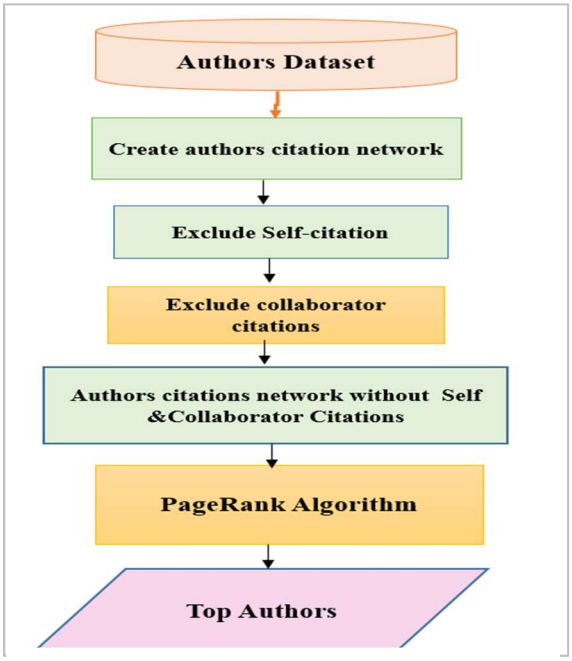


*Figure 6: Sub view of the proposed model for Authors Ranking Procedure*

which includes the paper key and author-id, to eliminate the author's self-citations from our data that are above the normal range.

On our dataset, the procedure is as follows: first, the paper id is obtained; next, the author id is discovered from the author2paper dataset; and last, the reference number and its authors are obtained from the paper and paper2authors datasets. Next, we determine if the author-ids are equal. The total number of self-citations that are excluded from our dataset is 590221. These connections between authors have been removed from the citation relationship that represents the self-citation considered a fraudulent citation and has an excessive rate for cation count.

**B. Remove Collaborator Citation**

Authors typically utilize the collaborator citation to cite their collaborators' articles in other publications. Therefore, it is used for other papers published by collaborator authors. In this step, we have used author co-author dataset to explore the collaborator citation for 4,258,615 collaboration relationships.

Nowadays, many authors cite their collaborators to increase the citation count, so we explore author's collaborators and remove all incoming citations from collaborators to get a fair citation count for each author. For authors that are collaborators more than three times, all cations for each other are considered a fraudulent citations. We have analyzed collaborator's citation data, and we have observed that there are many authors that are co-authors with each other more than three times to 56 times, so we have considered co-authors who are collaborators more than three times a fake citation and we have removed it. We refer to collaborators who cited each other more than 3 times as fake citations.

The number of removed cations for collaborators is 139396, which indicates a higher number from the citation count.

**C. PageRank for Author Ranking**

In this step, we have used the author's dataset after eliminating author-self citations and collaborator citations from the author's data to rank authors using PageRank algorithm to get a fair ranking value for each author. Identifying Top authors is the output of this step. The outcome of this step also includes identifying Experts for each scientific field.

**D. Identify the Most Important Papers for Top Authors**

The initial procedure of this step implementation was employing two datasets for the most important papers based on citation count and centrality measures and the top author's dataset. The second procedure involves removing all papers for authors with lower ranks from the dataset to produce a dataset that only contains the most important papers for qualified authors. Finally, the merging procedure for the most significant paper file with top author's data files using the Pyspark SQL for joining two data files.

### 4.3 Proposed Model Third Phase Rank Papers According Weight Score.

This phase consists of several processes. The first step is the data merging with the journal dataset, the second step involves computing the similarity measures between papers to determine their relevancy, and the third step is calculating a score for each paper and then ranking papers according to the new weight.

#### 4.3.1 Data Merging

This step involves joining the Scopus publication venue dataset with the top author's dataset to combine the most significant papers for those publications. The Scopus dataset includes Journal id, Title, Citation Count, SJR, and Quartile.

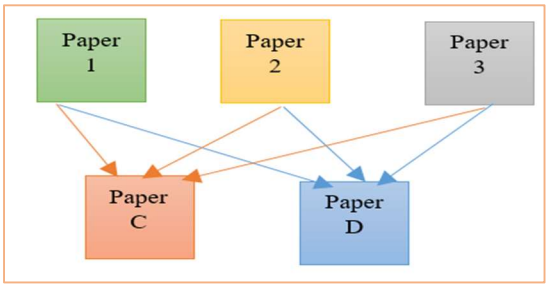


*Figure 7: Bibliographic Coupling Analysis*

Important papers for qualified authors are produced as a result of this stage, together with venue information that is important in the paper evaluation and ranking process, as applied in all academic institutions.

#### 4.3.2 Calculating Paper Similarity: Bibliographic Coupling, Co-Citations and Distance

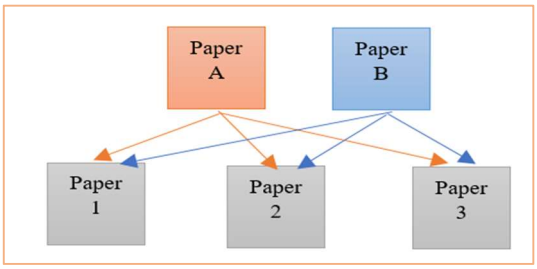


*Figure 8: Co-citation Analysis*

**A. Bibliographic coupling**

Bibliographic coupling and co-citation are two similarity measures that are used to measure the similarity relationships between papers based on the literature [43],[44],[45]. Therefore, all of the relationships stated above are anticipated to reflect relationships in the similarity between scientific papers, transmitting semantic interdependence between papers.

Bibliographic coupling considers cite-out; cite-out means two papers cite other three papers together (cite-out, in graph theory out-degree). For example, in figure 7, paper A and B both cite-out other three papers these are paper-1, paper-2 and paper-3, so the bibliographic coupling strength for paper A =3 & Paper B=3. Bibliographic coupling can be calculated using equation 1 [32]. That represents that the bibliographic coupling score for

paper A and paper B is one if they are citing paper i and zero otherwise.

$$B.C.score(A,B) = \sum_{i=1}^{n} B.C(A_i, B_i)$$

$$B.C(A_i, B_i) = \begin{cases} 1, if\ paper\ A\ and\ paper\ B\ cites\ paper\ i \\ 0, else \end{cases} \quad (1)$$

Bibliographic Coupling (BC) has calculated in our proposed model RBSD by getting the out neighbors for paper A and paper B, then checking the intersection between two lists and counting the bibliographic coupling score for each paper.

**B. Co-Citation**

Co-citation focuses on the cite-in; cite-in means that two papers are cited by other papers for example in Figure 8, three papers (in graph theory in-degree for the node). Figure 8 represent the co-citation strength for papers C & D cited together by papers 1, 2, and 3, so that the co-citation strength for paper C = 3 and paper D=3. The calculations of the co-citation score are represented in equation 2 [32] as the co-citation score is one if papers C and D have been cited by paper i, 0 otherwise.

$$Co.c.score(C,D) = \sum_{i=1}^{n} Co.C(C_i, D_i)$$

$$Co.C(C_i, D_i) = \begin{cases} 1, if\ both\ papers\ C\ and\ D\ cited\ by\ paper\ i \\ 0, else \end{cases} \quad (2)$$

**C. Distance**

In graph theory, distance is the traversing cost between pairs of vertices; it represents the number of hops between nodes or the value of the weighted relationship. Distance between papers has been calculated using graph analysis theory applied by Apache Spark Graphx. In graph analysis, we get the distance between paper vertices for the most important papers determined by PageRank as a centrality algorithm for a graph using the adjacency matrix. Distance calculation has been made using the shortest path algorithm to calculate the distance between a vertex and all other vertices in the citation graph. The shortest path algorithm is a graph pathfinding algorithm that determines the shortest path distance between two nodes represented in the number of nodes or the weighted relationship value. "Hops express the number of relationships between two nodes" [46]. For distance values, after finding the distance between papers using the shortest path algorithm, we have calculated the total distance for each paper in the graph.

The higher distance value between papers is considered an indicator that; these two papers are in different research fields, so it means that they are not related to the same research topic. On the other hand, the lower distance value between papers implies that these two papers are related and belong to the same scientific research field.

### 4.3.3 Calculating Paper Weight

To calculate the weight score for each paper, we combine paper information, author information, paper similarity information, and publication venue information. In the first section of this step, we represent publication venue information that has been used in the proposed ranking approach; the second section explains paper weight score calculation.

**A. Publication Venue Information Used from Scopus Dataset**

For publication venue information, we have considered journal SJR and journal Quartile because they are essential factors that greatly influence the paper's evolution and ranking according to mentioned in the literature journal citation report (JCR) mainly depends on SJR and quartile. Journal SJR is a metric for evaluating the scientific impact of academic publications that takes into consideration both the volume of citations a journal receives and the standing or significance of the journals the citations are from. The SJR indicator for a journal is a numerical representation that reflects the typical weighted number of citations obtained each year for articles published in that journal over the previous three years, as indexed by Scopus. Higher SJR indicator values are intended to signify more prestigious journals. SJR is used for ranking academic journals based on citation weighting schemes and eigenvector centrality to be utilized in complicated and heterogeneous citation networks like Scopus; there is a tool called the SJR indicator[47],[48].

The Journal quartile is an important feature that affects the journal rank and determines; how prestigious a journal or less important journal is. There are many studies conducted by massy university teams that work on the importance of publication venue quartile analysis and proved that article ranking increased by higher journal quartile that publishing the article. Journal quartile is already used now in most universities to rank academic papers; also, it is used by Scopus and web of science. The quartile is a measuring parameter that evaluates the journal rank, where quartile-1 contains the highest ranked journals and quartile-4 is a category of the lowest ranked journals. Thomson Reuters journal citation report (JCR) also includes journal ranking, the top 25 %of journals in the category of Q1, the next 25% placed in Q2, the third 25% placed

in Q3, and the last category includes the fourth 25% in Q4 [49].

### B. paper weight score

In this step, we have calculated paper weight according to our proposed RBSD model that can consider the citation count, the structural relationship between papers, the author citation count, h-index, Paper similarity or relevancy, and Publication venue information. Publication venue information includes SJR and journal quartile, which affects the quality and rank of the paper. In order to include author's information in our weight score for each paper, the author H-index has been employed in the calculation [50]. Author's H-index refers to the Hirsch index as the evaluation method that is used to measure an author's productivity and citation's impact on authors or academic members' levels. H-index has been used for evaluations by the three most popular bibliometric databases; these are Web of Science (Thomson Reuters), Google Scholar, and Scopus Elsevier [50], [51].

Paper weight has been calculated according to equation-3. The numerator includes the total similarity of paper which is represents the value of bibliographic coupling and co-citation, author, and journal information. Paper similarity based on citation network analysis these are bibliographic coupling and co-citation. Author information considered in paper weight is the author H-index and journal information is the journal SJR and Quartile. Paper that has many authors, we get a normalized H-Index represented as an average h-index. The denominator is the distance between papers on the network to determine the relevancy and similarity of papers. In our proposed model, we calculated the weight score for 858797 unique papers after filtering papers missing journal information in the Scopus dataset. For the journal quartile, we have given each quartile a value; quartile-1 will receive a score of 4, quartile-2 will receive a score of 3, quartile-3 will receive a score of 2, and quartile-4 will receive a score of 1.

$$Paper\ weight = \frac{B.B + CO.C + J.SJR + J.Quartile + aut\ .H.index}{Distance} \quad (3)$$

Where (B.B) is the bibliographic coupling score, (Co.c) is the co-citation value for each paper, J.SJR is the SJR value for journal.

## 5. EXCREMENTAL STUDY

### 5.1 Experimental Dataset Description

In this research, we apply our proposed model using the Aminer [52] computer science citation dataset for an academic, social network that is extracted from DBLP [53], ACM [54], and other data sources.

*Table 2: Dataset Description*

| Ser. | Dataset name | Description | Attributes | Size |
|---|---|---|---|---|
| **Aminer Academic social media dataset (DBLP, ACM)** | | | | |
| *1* | Paper dataset | Paper information with its citations data | Paper index, authors, affiliations, year, publication venue, reference id, and abstract. | 2,092,356 papers with 8,024,869 citations |
| 2 | Authors dataset | Author's information who published the paper. | Author index, author name, affiliation, no.of.papers, h-index, p-index and research interests. | 1,712,433 authors |
| *3* | Author co-author citation relationship. | authors collaborators | Index of authors with the number of collaboration between them | 4,258,615 collaboration relationship |
| *4* | Author2 paper | Author with it papers id's | The relation between author id and paper id | 1048576 |
| **SCOPUS Elsevier Dataset (Publication venue information)** | | | | |
| 5 | Scopus Elsevier | Publication venue information | Journal_id, Tittle, citation count, SJR, Quartile | 59345 |

The dataset is divided into four files as represented in Table 2, these are paper information data, author's data, co-author relationship information, and the last file saves the relationship between papers and authors. The first data file holds paper information data for 2,092,356 papers with 8,024,869 citation relationships, which are paper id, paper title, Authors, Author's Affiliation, publication venue, publication year, reference id, and abstract for the paper. Also we have 1000 record collected manually form the internet for academic paper data and added to our dataset. The second file for author's data includes 1,712,433 records for authors, represented as author id, author name, Affiliation, Publication Count, Citation Count, author H-Index, P-index for each author with equal A-index, P-index

with unequal A-index of this author, Research Interests for each author.

The third file comprises the author collaborators relationship that is represented as an index for each author with the count number of collaborations between them for 4,712,615 collaboration relationships. Finally, the fourth file holds the relationship between authors with their published papers with 1,048,576 records. The features for the fourth data file are author-id, paper-id, and author position in each paper. Author position means if the paper has two authors, X and

Y; if author x is written first and author Y second in the paper, the position of author X=1 and author Y=2 and vice versa).

Also, for publication venue information, we have used the Scopus Elsevier dataset [55] for computer science and information systems journals which includes 59345 journal information.

### 5.2 Experimental Data Preprocessing

Several preprocessing steps has performed to transform our dataset from an unstructured to a structured form, as represented in figure 9 for raw data and figure 3 for structured form. All duplicates in our dataset have been removed from papers and authors data files to work on unique and qualified data. Space removal for each record value in all dataset files has been made because it causes errors in reading or transforming data. All special characters or symbols that are represented in figure 3 are removed from dataset files represented as (#, %, @,*, #o, #t ). We have used dataframe and SQL for apache spark to read the data file and transform data into a structured format to be suitable for analysis. In the papers dataset, papers with no reference have been removed because it does not serve our analysis task for creating a citation network.

```
#* Space-Time Trade-Offs for Banded Matrix Problems
#@ John E. Savage
#o Brown Umversity, Providence, Rhode Island
#t 1984
#c Journal of the ACM (JACM)
#% 289024
#% 408638
#% 600828
#% 688897

#index 13
#* The VLSI Complexity of Selected Graph Problems
#@ Joseph Já Já
#o The Pennsylvania State Univ., University Park, PA
#t 1984
#c Journal of the ACM (JACM)
#% 769
#% 289452
#% 319821
#% 408343
#% 598672
#% 598673
#% 598675
#% 600547
#% 600560
#% 600811
```

*Figure 9: Paper Data before Preprocessing*

There are several Data preprocessing steps, including building a data pipeline that includes many tasks; each task has an output for each step that is used as input for the next stage. ETL process for Extract, transforming, and loading our data includes extracting data using Spark, loading data to Hive and transforming it using Spark then loading it to HDFS to be ready for our analysis tasks.

## 5. EXPERIMENTAL RESULTS AND DISCUSSION

This section outlines the findings of an experiment conducted to rank big scholarly data using the RBSD model. This section consists of three parts; the first part describes the results of central Papers in the Citation Network-Based PageRank algorithm. The second section explain results of the author's rank, and the citation analysis for authors includes; self and collaborator citation analysis. Finally, the third section describes the results of big scholarly data ranking using RBSD.

### 6.1 Central Papers in the Citation Network-Based PageRank Algorithm

The identification of the central paper has made using PageRank algorithm a centrality algorithm for graph analytics. The result of this step is a list of central and significant papers. Figure 10 illustrate a sub-view of visualization for the paper citation network, demonstrating that the article with id 1016299 as a central paper has a higher number of incoming citations as an In-degree for graph analytics.

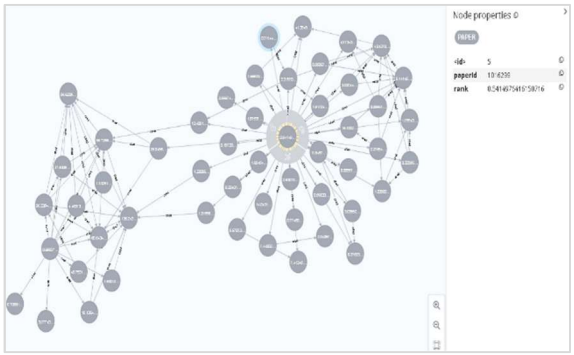


*Figure 10: Screenshot for Central Paper from paper citation Network using Neo4j*

### 6.2 Results for Authors Citation Analysis (Self-Citation &Collaborator Citation).

Our experimental results show that self-citation and collaborative citation rates in our dataset are 56.3% and 13.3 %, respectively. This indicates that most authors rely on self-citation more than collaborative citation to increase their rank, as shown in figure 11. Collaborator citations used by authors in our dataset

are 590221 citation relationship and 139396 for the self-citation relationship.

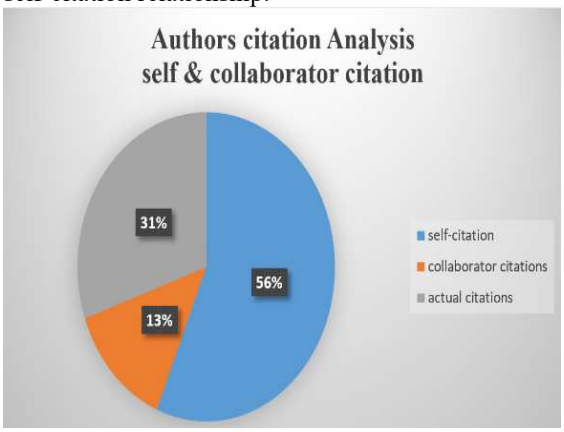


*Figure 11: Self-citation and collaborator citation Rate*

According to the results, ranking authors without removing self and collaborative citations consider an inaccurate rank for authors. The results demonstrated that author ranking before removing self and collaborator citations, as shown in figure 12, differs from author rank after removing self and collaborative cations.

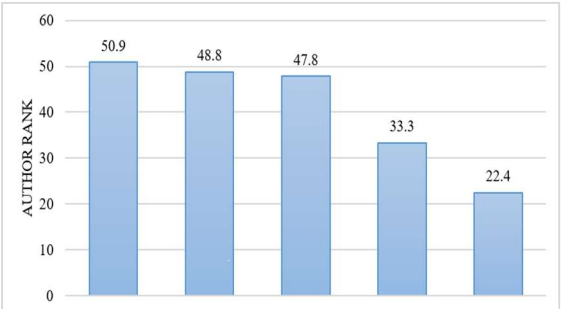


*Figure 12: Authors Rank Filtering Self-Citations And Collaborator Citations.*

Therefore some top-qualified authors have ranked low due to using self and collaborator citations, such as an author with id 779043; before removing self and collaborator citations was given the third rank, but when removing them, he got the first rank. Figure 13 represents the author's rank after excluding self- and collaborator citations. By comparing different ranks from figure 12 and figure 13, we noticed that author 532138 ranks as the first top author, however, when we removed self and collaborator citations the same author gets the fourth rank as a fair rank for authors. Also, author 795174 got the fifth rank before filtering self and collaborator citations, but after excluding them he gets the third rank. This proved that our proposed model has a great influence on fair rank for authors that affects the award procedures for authors.

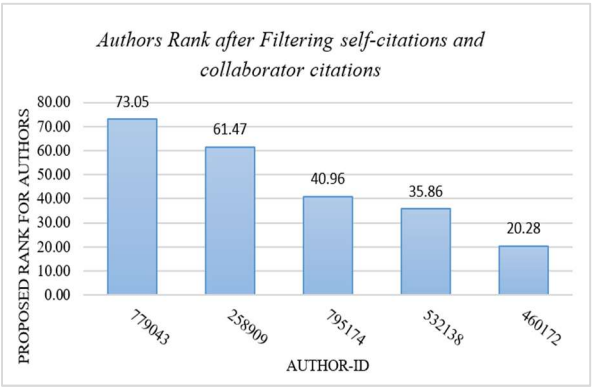


*Figure 13: Authors Rank after Filtering self-citations and collaborator citations.*

### 6.3 Results of Ranking System for Big Scholarly Data (RBSD)

Different experiments have been implemented to evaluate our proposed model (RBSD) capability. The proposed ranking system (RBSD) for big scholarly data has been compared with the traditional ranking systems to measure the validity of our proposed model. Other ranking procedure was applied to our dataset to rank papers based their models, including self and collaborator citations with all the paper dataset which caused implementation complexities. In performance evaluation, we have used statistical methods that are common in information retrieval to measure the performance of ranking algorithms; these methods are normalized discounted cumulative gain (NDCG) and mean reciprocal rank (MRR)[15],[56]. Normalized discounted cumulative gain is used to measure the quality of ranking system performance evaluation in information retrieval methods that are used to measure the effectiveness of web search engine ranking algorithms and related applications [57]. Two underlying assumptions govern how mean reciprocal rank operates; First, highly relevant items are more helpful when appearing in search results.

Second, articles with a high degree of relevance are more helpful than those with a low Two underlying assumptions govern how mean reciprocal rank operates; First, highly relevant items are more helpful when appearing in search results. Second, articles with a high degree of relevance are more helpful than those with a low degree of relevance, which is more helpful than those with no relevance.

$$NDCG_{\mathrm{p}} = \frac{DCG_{\mathrm{p}}}{IDCG} \quad (4)$$

$$DCGp = \sum_{i=1}^{p} \frac{(2^{r}-1)}{\log(1+i)} \quad (5)$$

$$IDCGp = \sum_{i=1}^{p} \frac{1}{\log(1+i)} \quad (6)$$

The total number of normalized discounted cumulative gains, obtained at rank p is represented by NDCGp as in equation-2. DCGp is the cumulative gain at a specific rank p. The recommended item's relevance value is at ranges from 0 to 1. If the system rank has efficient performance, the value will be set to 1, and 0 value indicates that the ranking system has poor performance. If the value is 1 so we have an ideal ranking. As represented in Table 3, our experimental results for NDCG values validate our proposed RBSD model and indicate a good performance of our model than google scholar ranking system relying mainly on citation count. We have compared our model with paper ranking FRPT algorithm [2], and Google scholar ranking algorithm [59], [60], [61]. The NDCG results for ranking-based google scholar ranking procedure are 0.596, FPRT 0.381, and RBSD gained 0.758 as represented in figure 14, demonstrating that our proposed RBSD model outperforms google scholar and FPRT ranking systems. Mean Reciprocal Rank MRR is commonly used in information retrieval for measuring the performance of ranking and algorithms to measure; if the system will rank the most important and relevant paper at the top rank. Equation-5 represents the calculation of MRR, where n is the number of users, and it represents the item rank.

$$MRR = \frac{1}{n} \sum_{i=1}^{n} \frac{1}{rank\,i} \quad (7)$$

*Table 3: Performance Evaluation Comparison For RBSD Proposed Model, Google Scholar, And FPRT Ranking Systems.*

| System | MRR | NDCG |
|---|---|---|
| Google scholar Ranking | 0.623 | 0.596 |
| FPRT | 0.425 | 0.381 |
| The Proposed Model (RBSD) | 0.823 | 0.758 |

Our experimental findings showed that our suggested RBSD model has a higher value than other ranking systems, as illustrated in figure 14, demonstrating its greater acceptance and superior performance. The classic ranking system's performance achieved lower values than our model due to issues with recent papers; these papers consistently received low rankings, even though they were written by eminent authors and published in reputable journals.

MRR results for the proposed model are 0.823, FPRT ranking 0.425, and google scholar rank that depends on the incoming number of citations achieved 0.623. By comparing the performance of other systems, we have found that FPRT system has many complexities and does not have the ability to deal with the massive dataset.

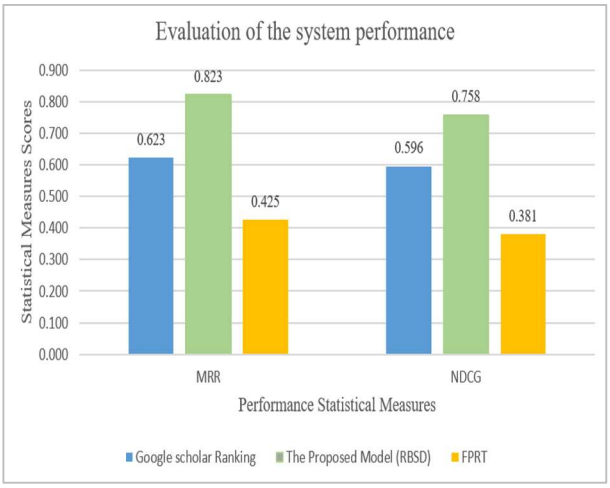


*Figure 14: Performance Evaluation comparison for RBSD Proposed Model, Google Scholar, and FPRT Ranking systems*

Our suggested model performs well because it considers the structural relationships between papers, the relevancy of each paper, and the capability to take into account the papers, authors, and journal information used to score each paper. Figure 14 represents the higher values of MRR and NDCG for our proposed model.

## 6. CONCLUSION AND FUTURE WORK

The vast amount of big scholarly data that is now available on the internet as a result of digital transformation and academic and social networks presents a significant challenge for ranking algorithms in academic institutions. The earlier methods relied on the traditional PageRank algorithm, which only considers the number of citations made to academic articles, ignoring the quality of more recent studies and resulting in consistently poor rankings. Ranking systems constructed using the traditional PageRank algorithm have implementation complexities and are conducted on a limited number of dataset records. Furthermore, the state-of-art ranking systems do not consider all relationships between big scholarly data environment entities. In spite of the fact that PageRank algorithm ranks papers based on their number of citations rather than their quality, many excellent papers are neglected since they are new and do not yet have any citations. The proposed RBSD model overcomes previous approaches' limitations

and considers the papers' authority. Authority of the paper means the credibility of the paper source and the authoritative source of authors and information.

The proposed model was conducted by considering citation network analysis, the author's network analysis using graph theory, and the similarity of papers. In addition, publication venue information from the Scopus dataset has been considered for 59,345 journals to investigate the influence of the prestigious or low-ranked venue and found that it has a greater impact on paper ranking value. To do that, we have joined the Scopus dataset for publication venue information with DPLB and ACM paper and author datasets.

We have used a modified PageRank algorithm for parallel computing using apache spark graphX to identify the most central and significant papers in paper citation network. The modified version can handle massive amounts of data, reduce intensive computations and overcome the limitations of the traditional PageRank algorithm. Our proposed model was conducted on different datasets paper, author, co-author, and journal datasets to explore the relationship between them for 2,092,356 papers, with 8,024,869 citations.

The proposed model has suggested a fair rank for authors by excluding incoming self and collaborators author's citations. We have found that PageRank algorithm performed well in author ranking because it is more suitable for the author's data than paper ranking due to its complexities and interrelated entities that affect the paper rank and is not considered by PageRank algorithm. In addition, the rate of self and collaborators citation has been explored in the used dataset, which found 56% and 13%, respectively, indicating that authors used self-citation rather than collaborator citations to increase their citation count.

Similarity for each paper has been analysed by calculating bibliographic coupling, co-citations for each paper, and the distance between papers in the network. RBSD model ranks papers according to a weighted score for a paper that considers all the required information that affects the paper's rank.

We have compared our RBSD model with other ranking systems, and the experimental results proved that our proposed model outperforms the google scholar ranking procedure and FPRT ranking technique. The proposed model also achieves more qualified results by analyzing data from different perspectives. Paper weight has been calculated for 858,797 unique papers with their authors and journal features. The proposed model helps determine the most important papers, key authors, and field expertise. It also considers the similarity of papers in ranking and explores the effect of publication venue information in academic paper ranking. For future work, we can incorporate journal self-citations in our analysis and use other author indicators, such as the UPI index. We can also consider the analysis of co-work network between papers as well as the semantics of the paper in future studies.

## REFERENCES:

[1] D. Fiala and G. Tutoky, "PageRank-based prediction of award-winning researchers and the impact of citations," J. Informetr., vol. 11, no. 4, pp. 1044–1068, 2017, doi: 10.1016/j.joi.2017.09.008.

[2] T. Abdel and L. Ali, "FPRT : Fair Paper Ranking Technique," vol. 15, no. 6, pp. 136–143, 2017.

[3] M. Dunaiski and W. Visser, "Comparing paper ranking algorithms," in ACM International Conference Proceeding Series, 2012, pp. 21–30, doi: 10.1145/2389836.2389840.

[4] Y.-R. Lin, H. Tong, J. Tang, and K. S. Candan, "Guest Editorial: Big Scholar Data Discovery and Collaboration," IEEE Trans. Big Data, vol. 2, no. 1, pp. 1–2, 2016, doi: 10.1109/tbdata.2016.2562840.

[5] F. Xia, W. Wang, T. M. Bekele, and H. Liu, "Big Scholarly Data: A Survey," IEEE Trans. Big Data, vol. 3, no. 1, pp. 18–35, 2017, doi: 10.1109/tbdata.2016.2641460.

[6] A. M. Idrees, Y. Helmy, and A. E. Khedr, "Credibility aspects' perceptions of social networks, a survey," Soc. Netw. Anal. Min., vol. 12, no. 1, p. 98, 2022, doi: 10.1007/s13278-022-00924-6.

[7] M. Khabsa and C. L. Giles, "The number of scholarly documents on the public web," PLoS One, vol. 9, no. 5, 2014, doi: 10.1371/journal.pone.0093949.

[8] Https://academic.microsoft.com/home.

[9]https://www.semanticscholar.org/topic/Bibliographic-database/108261.

[10] S. Misra and S. Bera, "Introduction to Big Data Analytics," in Smart Grid Technology, Cambridge University Press, 2018, pp. 38–48.

[11] S. Z. A. Elhady, N. I. Ghali, A. Abo-Elfetoh, and A. M. Idrees, "Exploratory Big Data Statistical Analysis the Impact of People Life'S Characteristics on Their Educational Level," J.

Theor. Appl. Inf. Technol., vol. 100, no. 5, pp. 1495–1509, 2022.

[12] T. A. L. Ali, M. H. Khafagy, and M. H. Farrag, "Big Data Challenges: Preserving Techniques for Privacy Violations," J. Theor. Appl. Inf. Technol., vol. 100, no. 8, pp. 2505–2517, 2022.

[13] T. L. Nguyen, "A Framework for Five Big V's of Big Data and Organizational Culture in Firms," in Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018, 2019, pp. 5411–5413, doi: 10.1109/BigData.2018.8622377.

[14] N. Y. Hegazy, M. H. Khafagy, and A. E. Khder, "Big Scholarly Data Techniques, Issues, and Challenges Survey," J. Theor. Appl. Inf. Technol., vol. 100, no. 5, pp. 1236–1246, 2022.

[15] R. R. Larson, Introduction to Information Retrieval, Cambridge University Press, ISBN:978-0-521-86571-5, 2009.

[16] A. Dode and S. Hasani, "PageRank Algorithm," IOSR J. Comput. Eng., vol. 19, no. 01, pp. 01–07, 2017, doi: 10.9790/0661-1901030107.

[17] M. Attia, M. A. Abdel-Fattah, and A. E. Khedr, "A proposed multi criteria indexing and ranking model for documents and web pages on large scale data," J. King Saud Univ. - Comput. Inf. Sci., no. xxxx, 2021, doi: 10.1016/j.jksuci.2021.10.009.

[18] R. Kleminski, P. Kazienko, and T. Kajdanowicz, "Analysis of direct citation, co-citation and bibliographic coupling in scientific topic identification," J. Inf. Sci., vol. 48, no. 3, pp. 349–373, 2022, doi: 10.1177/0165551520962775.

[19] E. Yan and Y. Ding, "Scholarly network similarities: How bibliographic coupling networks, citation networks, cocitation networks, topical networks, coauthorship networks, and coword networks relate to each other," J. Am. Soc. Inf. Sci. Technol., vol. 63, no. 7, pp. 1313–1326, 2012, doi: 10.1002/asi.22680.

[20] V. I. Voloshin, Introduction to graph theory, Fourth Edi. Addison Wesley,1996, 2009.

[21] D. K. Singh, P. K. Dutta Pramanik, and P. Choudhury, "Big Graph Analytics: Techniques, Tools, Challenges, and Applications," in Data Analytics, no. September, 2019, pp. 171–197.

[22] B. Chambers and M. Zaharia, Spark: The Definitive Guide Big Data Processing Made Simple. 2018.

[23] J. E. Gonzalez, R. S. Xin, A. Dave, D. Crankshaw, M. J. Franklin, and I. Stoica, "GraphX: Graph processing in a distributed dataflow framework," Proc. 11th USENIX Symp. Oper. Syst. Des. Implementation, OSDI 2014, pp. 599–613, 2014.

[24] M. Zaharia et al., "Resilient distributed datasets: A fault-tolerant abstraction for in-memory cluster computing," Proc. NSDI 2012 9th USENIX Symp. Networked Syst. Des. Implement. pp. 15–28, 2012.

[25] "Graphx spark 3.3.0 documentation." https://spark.apache.org/docs/latest/graphx-programming-guide.html.

[26] https://docs.databricks.com/integrations/graphframes/index.html.

[27] M. S. Malak and R. East, Spark GraphX in Action. Manning Publications Co., 2016.

[28] M. Dunaiski and W. Visser, "Comparing paper ranking algorithms," ACM Int. Conf. Proceeding Ser., pp. 21–30, 2012, doi: 10.1145/2389836.2389840.

[29] X. Liu, "PageRank for Ranking Authors in Co-citation Networks Ying," J. Am. Soc. Inf. Sci. Technol., vol. 64, no. July, pp. 1852–1863, 2009, doi: 10.1002/asi.

[30] D. Zhang and M. R. Kabuka, "Distributed Relationship Mining over Big Scholar Data," IEEE Trans. Emerg. Top. Comput. vol. 9, no. 1, pp. 354–365, 2021, doi: 10.1109/TETC.2018.2829772.

[31] A. M. Nair, J. P. George, and S. M. H. Gaikwad, "Similarity Analysis for Citation Recommendation System using Binary Encoded Data," 2nd Int. Conf. Electr. Commun. Comput. Eng. ICECCE 2020, no. June, pp. 12–13, 2020, doi: 10.1109/ICECCE49384.2020.9179380.

[32] J. Son and S. B. Kim, "Academic paper recommender system using multilevel simultaneous citation networks," Decis. Support Syst. Sci. Direct, 2017, doi: 10.1016/j.dss.2017.10.011.

[33] F. Majeed et al., "Self-Citation Analysis on Google Scholar Dataset for H-Index Corrections," IEEE Access, vol. 7, pp. 126025–126036, 2019, doi: 10.1109/ACCESS.2019.2938657.

[34] N. Yurko, I. Styfanyshyn, and U. Protsenko, "Self-Citation: the Risks and Benefits," Грааль Науки, vol. 1, no. 1, pp. 280–283, 2021, doi: 10.36074/grail-of-science.19.02.2021.057.

[35] E. Roberts and K. Schroeder, "The Google PageRank Algorithm," 2016, [Online]. Available: https://web.stanford.edu/class/cs54n/handouts/24-GooglePageRankAlgorithm.pdf.

[36] J. Beel, "Google Scholar ' s Ranking Algorithm : The Impact of Citation Counts ( An Empirical Study ),"  no. April, 2009, doi: 10.1109/RCIS.2009.5089308.

[37] J. Beel, "Google Scholar " s Ranking Algorithm : An Introductory Overview," no. July 2009, 2014.

[38] W. S. Hwang, S. M. Chae, S. W. Kim, and G. Woo, "Yet another paper ranking algorithm advocating recent publications," Proc. 19th Int. Conf. World Wide Web, WWW '10, no. July 2014, pp. 1117–1118, 2010, doi: 10.1145/1772690.1772832.

[39] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates, "FA∗IR: A fair top-k ranking algorithm," Int. Conf. Inf. Knowl. Manag. Proc., vol. Part F1318, pp. 1569–1578, 2017, doi: 10.1145/3132847.3132938.

[40] M. M. U. Rathore et al., "Multilevel Graph-Based Decision Making in Big Scholarly Data: An Approach to Identify Expert Reviewer, Finding Quality Impact Factor, Ranking Journals and Researchers," IEEE Trans. Emerg. Top. Comput., vol. 9, no. 1, pp. 280–292, 2021, doi: 10.1109/TETC.2018.2869458.

[41] C. Shi, H. Wang, B. Chen, Y. Liu, and Z. Zhou, "VAIR: A Novel Visualization System for Article Influence Ranking based on Citation Context," IEEE Access, vol. 7, pp. 113853–113866, 2019, doi: 10.1109/ACCESS.2019.2932051.

[42] Z. Taşkın, G. Doğan, E. Kulczycki, and A. A. Zuccala, "Self-Citation Patterns of Journals Indexed in the Journal Citation Reports," J. Informetr., vol. 15, no. 4, 2021, doi: 10.1016/j.joi.2021.101221.

[43] K. W. Boyack and R. Klavans, "Co-citation analysis, bibliographic coupling, and direct citation: Which citation approach represents the research front most accurately?," J. Am. Soc. Inf. Sci. Technol., vol. 61, no. 12, pp. 2389–2404, 2010, doi: 10.1002/asi.21419.

[44] X. Liu, "Co-Citation Analysis, Bibliographic Coupling, and Direct Citation: Which Citation Approach Represents the Research Front Most Accurately?," J. Am. Soc. Inf. Sci. Technol., vol. 64, no. July, pp. 1852–1863, 2013, doi: 10.1002/asi.

[45] J. Yun, "Generalization of bibliographic coupling and co-citation using the node split network," J. Informetr., vol. 16, no. 2, pp. 1–14, 2022, doi: 10.1016/j.joi.2022.101291.

[46] M. A. AMANU, Graph Algorithms (Example in Spark & Neo4j), First edit. O'Reilly Media; 1st edition (May 26, 2019), 2019.

[47] B. González-Pereira, V. P. Guerrero-Bote, and F. Moya-Anegón, "A new approach to the metric of journals scientific prestige: The SJR indicator," J. Informetr., vol. 4, no. 3, pp. 379–391, 2010, doi: 10.1016/j.joi.2010.03.002.

[48] SCImago, "Description of Scimago Journal Rank Indicator," pp. 1–4, 2007, [Online]. Available: http://bit.ly/1tNwvj6.

[49] https://www.massey.ac.nz/study/library/researcher-support/publish-and-share-your-research/journal-ranking-and-impact/.

[50] P. Khurana and K. Sharma, "Impact of h-index on authors ranking: A comparative analysis of Scopus and WoS," arXiv Prepr. arXiv, p. 2102.06964, 2021.

[51] Y. Chen and Z. Liu, "The H-index and First-author H-index of Chinese Scholars in LIS," Proc. 2016 6th Int. Conf. Mechatronics, Comput. Educ. Informationization (MCEI 2016), vol. 130, no. Mcei, pp. 1107–1111, 2017, doi: 10.2991/mcei-16.2016.233.

[52] https://www.aminer.cn/data/?nav=openData#Citation.

[53] https://paperswithcode.com/dataset/dblp.

[54] https://paperswithcode.com/dataset/acm.

[55] https://www.scopus.com/home.uri.

[56] Q. Le and A. Smola, "Direct Optimization of Ranking Measures," vol. 1, no. 2999, pp. 1–29, 2007, [Online]. Available: http://arxiv.org/abs/0704.3359.

[57] P. Boldi and B. Ribeiro-Neto, "Proceedings of the 2nd ACM International Conference on Web Search and Data Mining, WSDM'09: Preface from the program chairs," Proc. 2nd ACM Int. Conf. Web Search Data Mining, WSDM'09, 2009.

[58] D. Walker, H. Xie, K. K. Yan, and S. Maslov, "Ranking scientific publications using a model of network traffic," J. Stat. Mech. Theory Exp., no. 6, pp. 1–5, 2007, doi: 10.1088/1742-5468/2007/06/P06010.

[59] J. Beel and B. Gipp, "Google scholar's ranking algorithm: The impact of citation counts (an empirical study)," in Proceedings of the 2009 3rd International Conference on Research Challenges in Information Science, RCIS 2009, 2009, no. April, pp. 439–446, doi: 10.1109/RCIS.2009.5089308.

[60] C. Rovira, F. Guerrero-Solé, and L. Codina, "Received citations as a main seo factor of google scholar results ranking," Prof. la Inf., vol. 27, no. 3, pp. 559–569, 2018, doi: 10.3145/epi.2018.may.09.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[61] C. Rovira, L. Codina, F. Guerrero-Solé, and C. Lopezosa, "Ranking by relevance and citation counts, a comparative study: Google Scholar, Microsoft Academic,WoS and Scopus," Futur. Internet, vol. 11, no. 9, 2019, doi: 10.3390/fi11090202.

[62] M. S. Shanoda, S. A. Senbel and M. H. Khafagy, "JOMR: Multi-join optimizer technique to enhance map-reduce job," 2014 9th International Conference on Informatics and Systems, Cairo, Egypt, 2014, pp. PDC-80-PDC-87, doi: 10.1109/INFOS.2014.7036682.

[63] Mahmoud, Hadeer and Thabet, Mostafa and Khafagy, Mohamed H. and Omara, Fatma A, An Efficient Load Balancing Technique for Task Scheduling in Heterogeneous Cloud Environment, 2021, Kluwer Academic Publishers, vol. 24, no. 4, 2021, doi: 10.1007/s10586-021-03334-z.