Journal of Theoretical and Applied Information Technology
15th March 2022. Vol.100. No 5
2022 Little Lion Scientific
ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

BIG SCHOLARLY DATA TECHNIQUES, ISSUES, AND CHALLENGES SURVEY

NAGWA YASEEN HEGAZY [1], MOHAMED HELMY KHAFAGY [2], AYMAN ELSAYED KHDER[3]

[1]PHD Researcher at Faculty of Computers & Information, Fayoum University, Cairo, Egypt, Department of Information System.

[2]Professor Department of computer science, Faculty of Computers & Information, Fayoum University, Cairo, Egypt.

[3]Professor Department of information system, Faculty of computers and information Technology, Future University in Egypt (FUE).

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

ABSTRACT

Researchers around the world always request a research article relevant to their topic that satisfies their information need. The academic research environment generates an excessive amount of data called big scholarly data. Scholarly data usually includes millions of raw data represented in authors, papers, citations, and publication venues as well as author's information and affiliation. The enormous amount of valuable data generated by academic research has attracted researchers to explore this problem domain using different methodologies. Finding the most important articles in the field is considered a critical issue for researchers and journals as well as academic institutions. Ranking systems have become a very popular topic in the academic environment due to their importance in hiring, promotions, grants and award procedures. An accurate ranking system leads to an efficient recommendation system. This paper describes the background for big scholarly data and technologies. It also reviews the most important ranking systems and their algorithms. Recommendation systems approaches are presented for academic research, and the characteristics of highly cited papers are highlighted to help researchers improve their paper citations. Finally, this paper introduces an overview of big scholarly data visualization techniques and existing tools.

Keywords: *Big Scholarly Data (BSD), Ranking Systems, Recommendation System, Citation Network.*

1. INTRODUCTION

Big Scholarly Data (BSD) refers to the vast amount of scholarly growing data that includes millions of authors, papers, citations, co-authors as well as the citation networks and digital libraries [1]. Distributed file system (DFS) provides a suitable environment for processing the massive volume of data and enable replicates and store data as well as increasing the storage and resource consumption[2],[3]. As a result of growing development in technology more and more researches have been published and shared in digital libraries. Ranking systems have become a very popular topic in academic environment due to their importance in hiring, promotions, grant and award procedures. It is very difficult to find a manual assessment method to rank the scholarly data and authors. Researchers always used the internet and knowledge sharing platforms to obtain and share researches and results. Researchers efforts to find the most relevant and important papers to their topics considered a time consuming especially with the millions numbers of available articles on the internet or in digital libraries, so this requires to have a recommender system to recommend the most important paper in each topic and field[4]. This massive and convoluted volume of data generated every day requires special handling rather than a traditional database because it is not robust enough to manage the massive volume of data [5].

Big data analytics is a new important scientific field that can deal with a massive volume of data. We can say that the 5'Vs for big data (Volume, Velocity, Variety, Veracity, and the fifth V

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

is the Value) make the big scholarly data essential topic of study [6], [7]. Volume indicates the massive amount of data generated by academic research and publications every day. Velocity indicates that the fast data that are generated and transmitted to digital libraries and journals every day. Variety indicates the variety of relationships between big scholarly data that makes this area complex and harder to analyze. Veracity indicates the fact that the quality and accuracy of data may not be at the high levels especially when collecting data from different data sources it may inconstant and less controllable. The fifth V is the Value refers to the ability to convert big scholarly data into real values, which comprise the ability to collect data from different data sources. Feng xia [7] considers The 5'Vs for big scholarly data area with millions of authors, citations, figures and tables. The 5v's characteristics includes; Volume to indicate the huge volume of millions of available scholarly documents on the public web, academic social network, academic search engines, and digital libraries. Variety in big scholarly data represented in the various entities for big scholarly data such as researcher, paper, publication venue, and institution. Veracity illustrated in author name duplication and disambiguation that is related to the quality of data as well as its important impact in the analysis of data results. Velocity refer to the growing rate of generated scholarly data every year. According to [8] in year 2014 the growing average is 6.3% per year. Value characteristic implies in analyzing big scholarly data help to evaluate research publication, fund allocation, and impact evaluation [7].

According to a study [8] in 2014, it estimates that the number of scholarly documents in the public web is 114 million scholarly documents available on the public web. In the 2020 year, an academic social network such as ResearchGate [6] announced on their web site that they have more than 135 million publications available, more than 17 million authors and 700000 research projects available on their network. Microsoft academic [9] in 2020 have a 241,170,095 publication document, 244,552,188 authors available on their network. So analyzing this volume of scholarly big data provides a variety of information that enables us to build effective ranking and recommendation systems for academic research that widely used in academic journal's impact, academic research, authors, and quality of publications.

Ranking researchers have become very important for various applications various hiring, promotions, grant or award procedures. Also considering citation for popular papers, most significant paper in a field and paper published in a prestigious venue should be regarded than less important paper and paper that published in the less famous venue. Creating a smart ranking system based big scholarly data analysis will be helpful for researches and academic institutions. The academic environment includes different entities with a large number of attributes rather than the complex relationship between these entities that make it difficult to analyze and understand as represented in figure-1. A researcher may be a student or author or

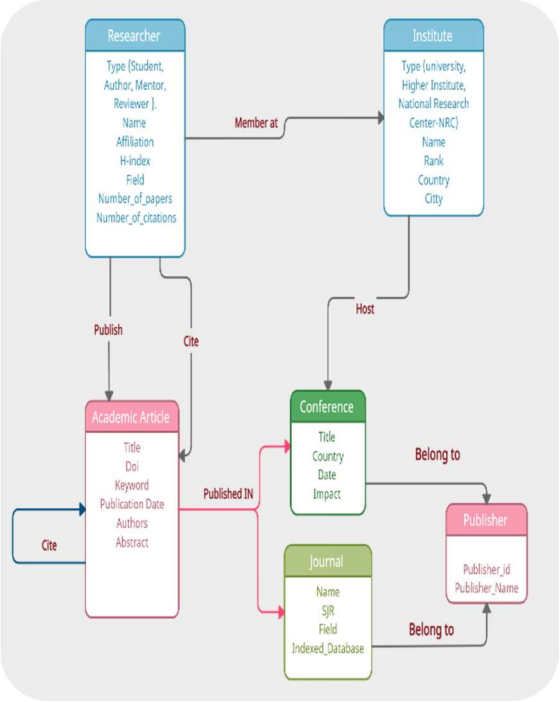

Figure-1: Major entities and their complex Relationships Associated to Big Scholarly Data in Academic Environment

mentor or reviewer. The researcher is a member at an institute and the institute may be a host for the academic conference. The institute has may be a university, higher institute, or a National research center all of them, help and support researchers to produce more academic articles. The researcher's productivity is an academic article that is published in a journal or conference also they cites articles.

2. CHARACTERISTICS OF HIGHLY CITED PAPER

Manuscripts Elgendi [10] proposed a novel technique to explore the characteristics of a highly cited paper using machine learning techniques. The proposed model goal is to analyze the publication

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

data for high and low cited paper data, to detect and uncover the pattern, features for highly cited papers that can help researchers and authors to improve their publication's citation rates. The model focused on eight factors these are number of tables, number of citations, number of views, number of characters with no space, number of figures, number of tables, number of equations, number of authors, and title length. The principal component (PC) algorithm has been used to uncover the features of highly cited papers to improve the citation rate. The results of the citation model demonstrated that the principal component algorithm has a good performance for detecting a complex dynamic between the publication features. Also, it proved that first; there is a significant positive correlation between the numbers of views, tables, and authors. Second, the number of citations is negatively correlated with the paper length. Third, the number of equations is not correlated with the number of citations [10]. According to Elgendi research authors may consider the following rules to produce a paper with a higher citation. The paper title length should be between 3 and 10 words, avoid using a dash or dot in the paper title recommended to use a colon instead, the highly cited papers have 6 or more authors. The minimum number of characters in the paper is 35.000 and the minimum number of tables are 2 tables also consider at least 6 figures in a paper.

Carlos et al. [11] 2021 proposed a new methodology to identify the highly cited paper in the Spanish public university system at the domestic level (HDP-DL). Carlos model aims to find the outstanding publication in the local context, exploring the impact among Spanish universities, determine universities that have the most significant role to determine papers that has a higher citation at the domestic level. The model was constructed based on the thematic classification in incites or essential science. The model results demonstrated that there is a preponderance of HCPS in the field of space science where the model has high visibility in the computer science field. Carlos's proposed methodology has a good performance in the local level of countries and it has complexity with the higher volume of data. Carlos's model does not use adequate data for a consecutive period to validate the accuracy and efficiency of the model. The model ignored important factors that effects the ranking of publications also model does not consider the time of the newly published paper.

According to previous literatures, we can say that the most important factors that produce a highly cited paper. These are the number of authors, quality of authors, and the prestigious journal that published the paper. Also selecting a meaningful title for the paper reflects the content of the paper considered a good indicator for a highly cited paper. Finally, the way of presenting and discussing the results part that is explained with figures and tables also gives a good impact for the paper to have higher citations.

3. RANKING SYSTEMS

Several approaches for ranking systems have been studied these approaches are based on the idea of search engine ranking techniques that includes different ranking algorithms such as Page Rank, CiteRank, YetRank, and NewRank [12] but, these studies have good performance with internet web pages but have lower performance when working with citation networks, some of these techniques does not consider important factors such as impact factor and h-index.

Several systems constructed based on PageRank algorithm [13], [14],[12],[15] used to rank webpages based on their importance or relevance. It rank papers based on the number of citations but does not consider the citation relationship between authors. It has been used to analyze the graph structure of the internet webpages based on computations that result in the probability distribution to represent a person likelihood to randomly clicking on links that will arrive at a certain web page. Google search engine recommendations are based on the PageRank algorithm [15][13]. Also, Google scholar used PageRank algorithm that generates recommendations for all related articles. It also ranks papers based on the citation number.

Dunaiski et al. [16] have proposed CiteRank algorithm that is constructed based on the idea of PageRank algorithm to overcome the problems of PageRank. Citeranke algorithm considers the aging effect in the citation network and the publication date of the paper but this algorithm has problems due to time and space complexity. Hwang et al. [17] developed yetRank algorithm to solve PageRank and CiteRank problems by considering the impact factor of publication venue. It gives a higher rank for the paper that published in the prestigious venue than the less famous venue but this algorithm have complexity to compute the impact factor for each venue per year. T. Abdel.et al. [18] proposed a new technique using Fair Paper Ranking algorithm (FPRT) that tries to solve problems with previous ranking techniques by considering seven important factors and developed

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

normalized impact factor to reduce the gap in impact factors between different scientific fields but this study does not consider the structural relationship in the citation network.

M.Rathor et al. [19] Constructed ranking system to rank researchers, journals using a modified version of page rank algorithm with consideration of new proposed impact factor that excludes the wrong and self-author/journal citations. Rathor system includes authors ranking to find the field experts using graph analysis, Hadoop eco. System and apache spark with graph X. Rathor model ignored the context of paper and the structural relationship between papers in the ranking system.

Chen et al. [20] 2019 devolved a novel citation context article influence ranking model to reduce the information redundancy in the semantic vector space and ameliorate the article retrieval. Chen uses the word2vector model and natural language processing technique to transfer article citation context to a word vector representation. The model goal is to allow uses to realize, compare article ranking results, and enable them to explore their desirable paper influence. Chen model ignored the influential pattern of the article that helps in evaluating the impact of paper also, other significant factors have not been included such as the authors, journal, and the impact factor of publication venue.

Table 1: Comparative Study of Ranking Systems

Author	Used Technique	Advantage	Disadvantage
Liua,Hasani, fiala[13] ,[15],[14]	PageRank algorithm	• Ranks web pages based on importance & relevance. • Used google scholar to generate a recommendation for all related articles. • Ranks papers based on the number of citations.	• Works well with internet web pages. • Has many limitations with citation networks. • Did not consider the citation network. • Recent papers have always given low rank.
Dunaiski, Bonchi [16] ,[21]	CiteRank Algorithm	• Based on the idea of PageRank algorithm but Considers the aging effect in the citation network. • Consider the publication date of paper.	• Time and space complexity. • More expensive.
Hwang, [17]	YetRank	• Consider the impact factor of the publication venue. • Gives higher rank for published paper in a prestigious venue than the less famous venue.	• Time and space complexity to compute the impact factors for each venue for each year.
Dunaiski, [16]	NewRank Algorithm	• Greatly improves the requirements of CiteRank algorithm because it is a combination of PageRank and CiteRank algorithms. • Normalize the initial value of paper based on its reference list.	• Does not consider the impact factor of the publication venue or author h-index. • The age of citing papers is not taken into consideration. • Citations from popular papers should be regarded as more important than citations from less important papers.
T.abdelatief [18]	FPRT Algorithm	• Consider the following seven factors; the number of authors, publication year, h-index of authors, citation score, journal impact factor, paper field, and the maximum value of impact factor in paper field. • It depends on three parameters Average h-index, Citations factor, and the Normalized impact factor.	• Algorithm works well with a large number of data but, does not apply the algorithm for enough size big scholarly data. • Does not consider the structural relationship for the citation network.
M.Rathor [19]	Modified version of page rank algorithm & proposed new impact factor	• Solve problems with the traditional PageRank algorithm by considering author ranking. • Help to the field experts. • Proposed new impact factor that excludes self-author citation and self-journal citations. • Provide the ability to avoid conflict of interest	• The context of paper has been neglected. • The quality of authors does not considered. • Ignored the structural relationship between citations.
Chen et al. [20]	Visual analysis VAIR for citation analysis & SPEAR model for ranking.	• Consider the context of article. • Improve article retrieval. • Reduce information redundancy.	• Ignored the influential pattern of article and the impact of paper. • Ignore the impact factor of the publication venue or author.

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

Table-1 represents a comparative study for ranking systems techniques, algorithms, advantages, and limitations for each technique that helps us to know that all previous ranking systems do not consider the quality of authors, influential pattern of an article, and the impact of academic paper. On another hand, all ranking system needs to consider the impact factor of the journal that publishes the academic article cause it affects the quality and the rank of the paper. Also, the most important challenge in big scholarly data is the complex nature that causes time and space complexity with most of the ranking algorithms..

4. RECOMMENDATION SYSTEMS

Nowadays all academic researchers deal with the increasing number of research publications, journals, conference proceedings white papers, etc. so, researchers cannot find their desired research paper related to their work easily due to information explosion that led to a waste of time. Recommendation systems for big scholarly data can help researchers to filter information and find their desired paper related to their academic research. Recommendation systems common techniques according to literatures [22], [23], [24] are content-based, collaborative filtering.

several studies have been conducted in recent years for finding similar related papers this is based on collaborative filtering that uses paper citation matrix for citation network to generate recommendations such as [25],[26]. Other studies recommend papers using content-based approach that considers the content of paper to find the relationship between papers such as [27]. Citation analysis techniques have been widely used to represent relationships between two papers that are cited together by other papers for paper recommendation, this is based on analyzing citation networks such as [28]. Depending on the previous three techniques for big scholarly data have difficulties in recommending highly qualified papers in the field especially when dealing with big data.

Several approaches were conducted based on text mining as a branch of data mining techniques to discover knowledge from big scholarly data that focuses on the content analysis of bibliographic data these studies also rely on information retrieval, information extraction such as [29],[30].

On the other hand, several theories have been proposed for document classification based textual pattern analysis for papers to classify papers for predefined topics such as ‘computer science’, ‘social science’ this done using k-nearest neighbors, naïve Bayes and support vector machine [31], [32]. Document clustering has been used to group similar documents into a labeled cluster [33].

Jieun son et al. [30] Proposed a new method called Multilevel Simultaneous Citation Network (MSCN) tends at constructing a recommendation system that tends to recommend informative and useful papers related to both the research topic and the academic theory. Jieun’s method combines citation analysis and content analysis as well as the implementation of content filtering and collaborative filtering. Citation analysis relies on the idea of analyzing the links that directly citing or cited by other papers for the multi-level network. Content analysis is implemented using a keyword matching process. Jieun’s method generates a multi-level citation network and then, selecting candidates papers after calculating the candidate score for each candidate paper. Jieun’s model is implemented on a limited volume of data and needs to be updated to deal with big scholarly data. Implementation of Content analysis for this model relies on a keyword matching process so it can not consider the semantic context.

Da.Zhang et al. [34] proposed a new system that relies on a distributed infrastructure for software and hardware to analyzing big scholarly data. The system goal where to discover the relationship between entities (papers/ authors) to recommend citations, discovering potential collaborator, recommend the paper to venue and expert recommendation. In order to perform the system goals Da.Zhang et al. proposed a mixed and weighted metapath (MWMP) to explore the relationship between entities. The proposed system was implemented using Apeach spark, Apeach Hadoop, Apeach HBase to evaluate the execution time and measure the efficiency of the proposed model. The limitations of Da.Zhang model does not consider the structural relationship between entities, excludes the self-citation and wrong citations and the neighbor information for each entity.

Jevin D et al. Constructing a recommender system based on Eigenfactor for citation-based method to improve scholarly navigation. The recommender system uses an algorithm based upon the hierarchical structure of scientific knowledge with a modified PageRank algorithm. The recommender system has been used to generate 300 million recommendations and founds that the system performs well when excluding citation overlap with co-citations and the system determines the most important paper in the field [35]. The model needs to

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

consider the context of paper and the structural relationship between papers.

Musa et al. proposed a system to recommend the academic reviewer and potential collaborators for the academic article. The proposed model was constructed based on the deep learning and pattern mining process by taking into consideration the author's H-index and the citation count of each paper. The proposed model compare between eight different algorithms and founded that EFIM (Efficient high-utility Itemset Mining) algorithm have a good performance in term of the run time and memory usage [36]. Mousa's system needs to apply using enough size of data that can be considered big scholarly data also the system

Table 2: Comparative Study of Recommendation Systems

Author	***Used Technique***	***Advantage***	***Disadvantage***
Jieun son et al.[30]	• Multi-level Simultaneous citation network (MSCN). • Combine content-based filtering and collaborative filtering methods	• The model has the ability to recommend informative useful papers related to the research topic. • The model recommends paper based on the similarity between papers. • The combination between content based filtering and collaborative filtering method led to higher performance for the model.	• The model needs to analyze the structural relationship between papers. • The quality of model relies on the highest rate of paper lists.
Da.Zhang et al.[34]	• Mixed weighted metapath (MWMP)	• Improve the relationship mining accuracy. • Reduce the running time. • Help to discover potential collaborators for research and recommend paper to venue.	• The model need to improve the length of metapath. • The model need to consider the neighbor information for each paper. • The model need to consider the structural relationship between entities. • The model did not exclude the self-citation and wrong citation.
Jevin et al.[35]	• Jevin's algorithm based hierarchical structure of scientific knowledge. • Eigenfactor citation-based method. • Modified PageRank algorithm.	• Improve the scholarly navigation. • Excluding citation overlap with co-citation. • The model has the ability to determine the most important paper in the field.	• The model ignored the structural relationship between papers. • The model need to consider the citation count and the impact factor.
Musa et al. [36]	• Deep learning algorithm FHM, UP-Growth, EFIM. • Mining algorithms RGP & RSP	• The system consider the citation count and author H-index. • The system has the ability to recommend academic reviewer and potential collaborators. • The system provide a good performance with run time and memory usage.	• The system used data cannot be fully relied on to prove the validity of the system, it need to be enough to considered as a big scholarly data that analyzed with a big data platforms. • The system need to consider the structural relationship between papers.
Magara et al. [37]	• Altmetric based technique to use the Altmetric form of paper. • Research paper ontology.	• The model enhanced the performance of recommending paper	• The model need to consider journal impact factor. • The established model need to compare its data with statistical sources. • The model need to analyze the structural relationship between papers.
Akhil et al. [38]	• Similarity analysis using binary code. • SABED algorithm	• The model provide better accuracy against other similar models (UBCF & IBCF).	• The model focused only on the content of paper and ignored the context of paper. • The model need to consider the impact factor and citation relationship.

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

ignored the structural relationship between papers and authors.

Magara et al. 2017 [37] established an Altemrtric based technique model to produce an improved recommendation system for a research paper. Magara model was based on the research paper ontology to enhance the performance of recommendation systems. Magara provides speed and the real-time impact that help in better recommendation but the model shout compared with other existing systems to evaluate the model efficiency.

Akhil et al. 2020 [38] proposed a model based on the similarity analysis using binary encoded data SABED algorithm that converts the article data into a binary code and stores it in a database. The model uses a binary code as a query to recommend citation. The model focused on the author name, paper DOI, keyword, abstract, and content of paper. The SABED algorithm provides better accuracy compared to other similarity analysis algorithms. The proposed model focused only on the content of article and ignored the article context. Table-2 represents a comparative overview of the advantages and limitations of big scholarly data recommendation systems that demonstrate, there was a need to consider the structural relationship between papers and does not reply on the highest paper rate in determining the quality of the recommended paper. The recommendation system should not consider only the content and context of paper it need a hybrid approach that considers the citation matrix, paper context, and relationship between papers that are cited together by other papers. Comparing data with its statistical source will be a new issue.

5. POPULAR DATASET

The first process for analyzing and visualizing big scholarly data is to collect data as raw data then extract information from collected data such as extracting author information, citation information. The popular datasets for scholarly data are DBLP, APS, MAG, and ORC as listed in table-3. DBLP is a computer science bibliographic dataset that provides information for bibliographic computer science journals, conferences, and publications [2dplb6]. DBLP contains more than 3.8 million publications that contain around more than 1.9 million authors. It has been indexed in 31,000 conferences or workshop proceedings and 32,000 journal volume [11]. APS dataset [30] for American Physical Society that provides reviews on modern physics that contains 450,000 articles science 1983. MAG dataset is a heterogeneous academic graph for scientific publication and citation relationship [31] it can be easily accessed through Microsoft Cognitive Services Academic Knowledge. ORC is a dataset that provided by semantic scholar project it contains more than 7 million paper for computer science and neuroscience fields that includes paper title, abstract, keywords, author name, paper URL, citation publication, publication date, and publication venue.

6. BIG SCHOLARLY DATA VISUALIZATION

Big scholarly data includes the massive size of row data that involves many attributes such as authors, paper title, keywords, citations, publication venues, and the citation network. Due to the rapid growth in technology and using digital publications.

visualization techniques for big scholarly data is a challenging area that paid heel to easily visualize big scholarly data to present the structure of the dataset to uncover the hidden relationship and patterns in the scholarly data [39]. Visualizing authors considers the best way to reflect the collaboration network as a two-dimensional network. Various types of networks used to visualizing big scholarly data are termed as Bibliographic network [40].

6.1 Visualization Tools

Visualization tools are include processing data and visual analysis for this data. Visual analysis facilitates the analysis of data than the raw data. Jiaying et al. [29] divide Visualization tools into two main categories; these are visualization tools without a programming language and visualization tools based on programming languages.

6.1.1 Visualization Tools without Programming Languages

a-Tableau: is an analytical tool for business intelligence that links data files on both local and server as well as its ability to deal with a variety of data formats such as XLS, CSV, and Text. It can provide importing data from online servers such as Oracle and MySQL[29],[32].

b-ICHART: is a business intelligence analytical tool for data visualization that provides an integration for multiple database platforms such as the official optimized API connector for NetSuite, Salesforce, and Google Cloud Platform. It can be linked for an automatically updated database as a real-time BI tool. ICHART provides various types of charts that

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

are suitable for different types of data as users' needs [29], [33].

c-INFOGRAM: [34] is a web-based application tool for data visualization charts. It only requires user registration then, it enables users to upload their data with different file formats as well as add the data file into google drive, one drive, or drobox. This application help in improving the sharing function of data so, users can embed their visualized charts on a web page using code that can automatically be generated or shared by URL or Email this make it an easy tool for users to visualize their data[35],[29].

Other Tools that need zero cod or do not need a programming language such as NODEBOX, GGPLOT, and JGRAPH.

Table 3: big scholarly data Popular Dataset with its size

Ser.	Dataset	Size
1	DBLP	3,800,000 publications
		1,900,000 authors
2	APS	450,000 articles
3	ORC	7 million papers

d-NODEBOX: is a program used for big scholarly data visualization based on python programming code that is a free tool building on MAC operating system. The program can provide two-dimensional visualization for data through the web either (static, animated, and interactive). It enables users to combine different types of functions through writing python scripts. It also supports many document formats and exports the visualization image into a PDF file also animations can be exported as quick-time movies [44], [39].

E-Ggplot2: is a visualization tools based on R that enable users for creating the statistical data and editing the plotting than the basic R graphs. The input data file format that supported by Ggplot2 are R and API. The features provided by Ggplot2 is the plotting process is based on layers and the graph composed of layers [45], [39], [41].

F- JPgraph: is a library based on object-oriented that helps users for creating graphs through visualizing their data [46]. The library is based on PHP5 and PHP7 that compatible with any PHP script [47],[39],[48].

6.1.2 Visualization tools based programming languages

Visualization tools that do not need a programming language or zero coding tools are easy for users to visualize their data into graphs or charts. It provides a flexible way for users to design their row data into graphs and charts. Some visualization tools are combined with JavaScript, others depend on other programming language such as R, Python, Java, and PHP. Visualization tools can be divided into two categories, there are tools based on JavaScript others based on other programming languages [39], [41]. There are common tools that are based on JavaScript these are D3.js, Charts.js, Fusioncharts, and Zingchart.

A-D3.js: is an open-source program based on JavaScript library that used HTML and CSS techniques. D3.js visualize data and import data in .SVG file format. The program can visualize and run data as HTML code in the browser platform under the user environment. This program provides various examples on their website for (graphs, charts, and source code) that encourage users to visualize their data using D3.js[42]. The program supports JASON, CSV, XML input file format and the output chart will be in HTML, Cavans, SVG, CSS. It has a powerful gallery with multiple charts, graphs, and maps that includes the world map and US map[41],[39].

b-Chart.js: is an open-source JavaScript program that uses canvas based on HTML5 so, it performs well with all current browsers[39], [42]. It has the ability to visualize data into various chart types based on the determined script language and the official chart library. Using chart.js users can incorporate the chart library into their coding file by code after that use the API from the library to set the parameters. [41]. The input file format for chart.js is the JavaScript API and the produced chart format is HTML5 and canvas. The program has 8 chart types that involve over 23 charts and graphs.

c-Fusioncharts: is a program that integrates JavaScript and action script 3.0 technologies considered easy to use platform that enables the program to run on different devices and browsers. Fusioncharts includes over 90 types of charts and more than 1000 maps. The file type that supported by fusionchart is an .XML and .Jason and the generated chart file type is .JPG, .PNG AND .PDF files[42], [39], [43]. It enables users to incorporate the generated interactive charts to user's applications with different wrappers in officially offered plugins such as PHP, JSP, JQuery, and Django charts.

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

7. CONCLUSION

Big scholarly data analysis provides a variety of data for the academic environment that help authors and journals to evaluate their work. The explosion of big scholarly data is due to the digitization of academic research and the increased number of digital documents published every day. Analyzing this data can help researchers determine the most important paper in their research field, Expert finding, research recommendation systems, and ranking systems. Ranking systems are required in the academic environment to evaluate the academic research, journals, and the quality of publications used in award and hiring procedures. Recommendation systems for academic research are used to recommend related articles to researchers according to their interests. We introduced different approaches for big scholarly data challenges for ranking and recommendation systems, and gave authors a useful overview of the characteristics of highly cited papers. A comprehensive representation for ranking systems in the academic articles has been provided, this helps uncover the vital attributes used for ranking the academic work. This study can open new insights for researchers in the era of big scholarly data analysis, provide the advantages and limitations of the current systems for both recommendation and ranking systems. Finally, the paper demonstrated big scholarly data visualization techniques and tools that help developers, researchers to gain knowledge and scope of existing programs and approaches for visualizing scholarly data. We introduce the limitations of the previous studies that give researchers the starting point of the open research issues. These limitations demonstrate that there is a need for a qualified ranking system that considers the quality of the research paper. The recommendation system needs to consider the structural relationship between papers, improve the length of the meta path, and consider the neighbor information for each paper. Additionally, we supposed that analyzing the semantic context of papers helps in determining the similarity of papers that can cluster papers according to their fields and extracting the important complex features of big scholarly data.

REFERENCES:

[1] 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.

[2] M. R. Kaseb, M. H. Khafagy, I. A. Ali, and E. S. M. Saad, "An improved technique for increasing availability in Big Data replication," Futur. Gener. Comput. Syst., vol. 91, pp. 493–505, 2019, doi: 10.1016/j.future.2018.08.015.

[3] R. Sahal, M. Nihad, M. H. Khafagy, and F. A. Omara, "iHOME: Index-Based JOIN Query Optimization for Limited Big Data Storage," J. Grid Comput., vol. 16, no. 2, pp. 345–380, 2018, doi: 10.1007/s10723-018-9431-9.

[4] F. Yang, J. Zhu, J. Lun, Z. Zheng, Y. Tang, and J. Wu, "A Keyword-based Scholar Recommendation Framework for Biomedical Literature," 2018 IEEE 22nd Int. Conf. Comput. Support. Coop. Work Des., pp. 247–252, 2018.

[5] R. D. Kim H. Pries, BIG DATA ANALYTICS, 1st editio. Auerbach Publications.

[6] T. L. Nguyen, "A Framework for Five Big V ' s of Big Data and Organizational Culture in Firms," 2018 IEEE Int. Conf. Big Data (Big Data), pp. 5411–5413, 2018, doi: 10.1109/BigData.2018.8622377.

[7] F. Xia, S. Member, W. Wang, T. M. Bekele, and H. Liu, "Big Scholarly Data : A Survey," vol. 00, no. 00, pp. 1–19, 2016, doi: 10.1109/TBDATA.2016.2641460.

[8] M. Khabsa and C. L. Giles, "The Number of Scholarly Documents on the Public Web," vol. 9, no. 5, 2014, doi: 10.1371/journal.pone.0093949.

[9] Https://academic.microsoft.com/home, "microsof academic.pdf." .

[10] M. Elgendi and S. Member, "Characteristics of a Highly Cited Article : A Machine Learning Perspective," IEEE Access, vol. 7, no. Ml, pp. 87977–87986, 2019, doi: 10.1109/ACCESS.2019.2925965.

[11]C. García-zorita, S. Marugán, D. De Filippo, and E. Sanz-casado, "Highly Cited Papers at the Spanish Domestic Level," vol. 6, no. April, pp. 1–14, 2021, doi: 10.3389/frma.2021.651991.

[12] M. Dunaiski and W. Visser, "Comparing Paper Ranking Algorithms," pp. 21–30, 2012.

[13] 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.

[14] 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.

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

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

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

[17] 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.

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

[19] 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.

[20] 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.

[21] 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.

[22] M. H. Mohamed, M. Khafagy, and M. H. Ibrahim, "Recommender Systems Challenges and Solutions Survey," no. February, 2019, doi: 10.1109/ITCE.2019.8646645.

[23] K. Elmenshawy and H. R. Fadlallah, "MUSIC RECOMMENDATION SYSTEM USED EMOTIONS TO TRACK AND CHANGE NEGATIVE USERS ' MOOD," J. Theor. Appl. Inf. Technol. Sept., vol. 99, no. 17, pp. 4358–4376, 2021.

[24] M. H. Mohamed, M. H. Khafagy, H. Elbeh, and A. M. Abdalla, "Sparsity and cold start recommendation system challenges solved by hybrid feedback," Int. J. Eng. Res. Technol., vol. 12, no. 12, pp. 2735–2742, 2019.

[25] H. Liu, X. Kong, X. Bai, and W. E. I. Wang, "Context-Based Collaborative Filtering for Citation Recommendation," IEEE J. rapid open access publisshing, vol. 3, pp. 1695–1703, 2015.

[26] K. Haruna, M. A. Ismail, D. Damiasih, and J. Sutopo, "A collaborative approach for research paper recommender system," pp. 1–17, 2017.

[27] Z. Taşkın and U. Al, "A Content-based Citation Analysis Study based on Text Categorization," reserch gate, no. November 2017, 2018, doi: 10.1007/s11192-017-2560-2.

[28] N. J. Van Eck, L. Waltman, and T. Studies, "Citation - based clustering of publications using CitNetExplorer and VOSviewer," pp. 1–25.

[29] M. Song and T. Chambers, "Text Mining with the Stanford CoreNLP, Measuring Scholarly Impact," Springer Int. Publ. Switz., pp. 215–234, 2014, doi: 10.1007/978-3-319-10377-8.

[30] 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.

[31] D. Munteanu and S. Bumbaru, "Classification Process in a Text Document Recommender System," 2005.

[32] imaduddin amin et al. Widyantoro, "citaions senrence identifiction and classification for related work," vol. 5, pp. 291–296, 2014.

[33] L. Bolelli, S. Ertekin, and C. L. Giles, "Graph Analysis," pp. 30–41, 2006.

[34] 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.

[35] J. D. West, I. Wesley-Smith, and C. T. Bergstrom, "A Recommendation System Based on Hierarchical Clustering of an Article-Level Citation Network," IEEE Trans. Big Data, vol. 2, no. 2, pp. 113–123, 2016, doi: 10.1109/tbdata.2016.2541167.

[36] M. I. M. Ishag, K. H. Park, J. Y. Lee, and K. H. Ryu, "A Pattern-Based Academic Reviewer Recommendation Combining Author-Paper and Diversity Metrics," IEEE Access, vol. 7, no. February, pp. 16460–16475, 2019, doi: 10.1109/ACCESS.2019.2894680.

[37] T. Zuva, "Toward Altmetric-driven Research-paper Recommender System Framework," 2017, doi: 10.1109/SITIS.2017.21.

[38] 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–

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

13, 2020, doi: 10.1109/ICECCE49384.2020.9179380.

[39] J. Liu, T. Tang, W. Wang, B. Xu, X. Kong, and F. Xia, "A Survey of Scholarly Data Visualization," IEEE Access, vol. 6, pp. 19205–19221, 2018, doi: 10.1109/ACCESS.2018.2815030.

[40] N. J. van Eck and L. Waltman, Visualizing Bibliometric Networks. 2014.

[41] S. K. A. Fahad and A. E. Yahya, "Big Data Visualization: Allotting by R and Python with GUI Tools," 2018 Int. Conf. Smart Comput. Electron. Enterp. ICSCEE 2018, no. October, 2018, doi: 10.1109/ICSCEE.2018.8538413.

[42] A. P. S, K. M. Khule, S. Karthika, and N. K. T, "Data Visualization Tools and Techniques For Datasets In Big Data," Int. Res. J. Eng. Technol., vol. 4, no. 8, 2017, [Online]. Available: https://irjet.net/archives/V4/i8/IRJET-V4I8296.pdf.

[43] "S. Nadhani and P. Nadhani, FusionCharts Beginner's Guide: The Of_cial Guide for FusionCharts Suite . Birmingham, U.K.: Packt, 2012.," p. 2012, 2012.

[44] T. De Smedt, L. Lechat, and W. Daelemans, "Generative art inspired by nature, using NodeBox," Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics), vol. 6625 LNCS, no. PART 2, pp. 264–272, 2011, doi: 10.1007/978-3-642-20520-0_27.

[45] V. Gómez-Rubio, "ggplot2 - Elegant Graphics for Data Analysis (2nd Edition) ," J. Stat. Softw., vol. 77, no. Book Review 2, pp. 3–5, 2017, doi: 10.18637/jss.v077.b02.

[46] "jpgraph," https://jpgraph.net/, 30-9-2021.

[47] A. Corporation, "JpGraph Manual," https://jpgraph.net/download/manuals/chunkhtml/index.html, 2016. .

[48] L. Pasteur and R. Koch, "JpGraph Manual," vol. 74, no. 1934, pp. 535–546, 1941.