Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.DOI

Bibliometrics & Citation Analysis of Pakistani Researchers in the Field of Data Science

ALI KHALID[1], MUHAMMAD AHMAD[2], MUHAMMAD TAYYAB[3], AND HAIDER ALI ADEEL[4]

[1]Information Technology University, Lahore, Pakistan (e-mail: masds21001@itu.edu.pk)
[2]Information Technology University, Lahore, Pakistan (e-mail: masds21021@itu.edu.pk)
[3]Information Technology University, Lahore, Pakistan (e-mail: masds21032@itu.edu.pk)
[4]Information Technology University, Lahore, Pakistan (e-mail: masds21044@itu.edu.pk)

ABSTRACT To be written.

INDEX TERMS Bibliometrics, Citation analysis, Research Impact, Data Science, Pakistan

I. INTRODUCTION

The scene of science and exploration is quickly developing. The number of people joining the field of research and innovation is increasing with the increase in population. Gone are the days when all individuals from a college office would praise the fruitful distribution of a colleague's paper [1]. Earlier, researchers would just consider the number of papers they had distributed as a proportion of their scholastic standing. The evolution of science has also led to the evolution of the quality of research work. Today, the center is progressively moving from whether an analyst has published a paper to where he/she has published it and the effect that piece of exploration has on established researchers and the world at large [2].

What criteria do you use to evaluate the quality of a research paper? How can you tell if your research is having an impact and is considered significant? An objective way is through citation analysis. Citation analysis is the process of evaluating and interpreting the citations that articles, scientists, universities, and countries receive to determine their scientific influence and productivity. Any research article should include a list of references that directs readers to previous relevant research. A reference, often known as a citation, is a type of recognition given by one research work to another. Scientists build on previous work to uncover new knowledge. Researchers read relevant published research and use it as a foundation for arguments made in their research papers to find gaps in existing research and choose a research topic.

Today's researchers are under increased pressure to publish their findings. Academic departments are required to publish a certain number of articles each year. Both individuals and institutions have a lot riding on the evaluation of research quality. As a result, governments, funding agencies, and tenure and promotion committees are looking for simple and objective approaches to evaluate growing research volumes in the shortest amount of time possible. To this purpose, they are increasingly resorting to citation analysis for objective impact assessment metrics. Data science is a hot research topic in this decade. Therefore, in this work, we performed a citation analysis of the publications in the field of data science by Pakistani authors. This analysis will help the researchers to identify trending topics, famous journals and productive research groups. Moreover, they will be in a better position to understand the reach of their work, identify the patterns in which their research work is used, benchmark themselves against their peers, and set objectives for themselves and their publications.

II. LITERATURE REVIEW

A citation index is similar to a bibliographic index in that it is a list of citations between publications that allows the user to quickly determine which later papers cite which older documents. There are different citation index databases. Some of them are proprietary while others are freely accessible to the public. The scope of these databases also varies. Most of them are focused on a single field of research and index material particular to that field. PubMed [6](medicine and the biomedical sciences), Chemical Abstracts [7], Mathematical Reviews [8], the ACM Digital Library [9] (computer sciences), and CiteSeer [10] (computer and information sciences), are some of the specialist databases that are available online and provide free access to bibliometric data in their particular field. Contrary to this, Web of Science [11](Clar-

ivate Analytics) and Scopus [12] (Elsevier) are multidisciplinary commercial citation index databases and most of the bibliometric studies are based on data covering these two databases. Google Scholar [15] is another free web search engine that indexes the full text or metadata of scholarly literature published in a variety of formats and fields. Researchers have also performed studies on the comparative analysis and coverage of databases [13], [14].

Citation impact uses citation indexes to find the number of times an academic journal article, book, or author is cited by other articles, books, or writers [3]–[5]. Citation counts are used to determine the impact or influence of academic work, giving rise to the area of bibliometrics or scientometrics, which focuses on the study of academic impact patterns through citation analysis. Different citation index databases has introduced different metrics to measure citation impacts. The journal impact factor is a measure of the relevance of journals based on the two-year average ratio of citations to papers published. Academic institutions use it to make judgments about academic tenure, promotion, and employment, while authors use it to choose which journal to publish in. Clarivate's Web of Science's impact factor (IF) or journal impact factor (JIF) of an academic journal is a scientometric indicator that measures the yearly mean number of citations of papers published in the previous two years in a certain journal. It is widely employed as a proxy for a journal's relative relevance within its area as a journal-level indicator; journals with higher impact factor values are accorded the status of being more important, or carrying more prestige in their respective disciplines, than journals with lower values. While it is widely used by universities and funding organizations to make decisions about promotion and research proposals, it has recently been accused of distorting sound scientific standards [16], [17]. Journal rank is another metric which is extensively used in academic circles to assess the influence and quality of academic journals. The purpose of journal rankings is to represent a journal's position within its area, the relative difficulty of publishing in that journal, and the prestige associated with it. In a number of countries, they have been adopted as official research evaluation instruments [18]. CiteScore (CS) [19] is an academic publication's metric that reflects the yearly average number of citations to recent articles published in the journal. Elsevier introduced this journal rating statistic in December 2016 as an alternative to the commonly used JCR impact factors (calculated by Clarivate). CiteScore is based on citations recorded in the Scopus database rather than the JCR database, and those citations are collected for publications published in the previous four years rather than the previous two or five years. The SCImago Journal Rank (SJR) indicator is a measure of the scientific influence of scholarly journals that accounts for both the number of citations received by a journal and the importance or prestige of the journals where the citations come from. A journal's SJR indicator [20] is a numeric value representing the average number of weighted citations received during a selected year per document published in that journal during the previous three years, as indexed by Scopus. Higher SJR indicator values are meant to indicate greater journal prestige. SJR is developed by the Scimago Lab originated from a research group at University of Granada.

Iqbal et al. [21] examine the computer networking domain and present a study based on the content and metadata of four major computer networking journals; IEEE Communications Surveys and Tutorials (COMST), IEEE/ACM Transactions on Networking (TON), ACM Special Interest Group on Data Communications (SIGCOMM), and IEEE International Conference on Computer Communications (INFOCOM). They compared the publication trends in INFOCOM and SIGCOMM to the co-evolution of trends in the COMST and TON journals. They used metadata analysis, content-based analysis, and citation analysis to analyze the computer networking literature. They also identified noteworthy trends as well as the most influential authors, institutes, and nations based on the number of publications and article citations.

Through a survey, Ali et al. [22] conducted a bibliometric analysis of research published by Library and Information Sciences (LIS) researchers in Pakistan. The authors created a questionnaire, which they circulated to LIS scholars via email, Yahoo groups, and Facebook to representatives from all over Pakistan. The data from a total of 104 respondents were then analyzed using the SPSS version 21 programme. They identified relevant demographic information; gender and location, the extent of collaborative authorship, the extent of publishing based on geographical regions, the strength of the association between job title(seniority) and the number of publications, the strength of citation metrics for national outputs, and factors that may harm LIS scholars' ability to conduct research and/or publish it.

For the period 1980–2011, R. S. Bajwa et al. [23] examines the research trends in Pakistan in the field of biotechnology. They looked at the rise and fall of the annual growth rate, as well as the comparison of organizations that actively participate in biotechnology research using the publication rate, citation rate, average citation per paper, and numerous indexing methods. The top 100 articles in the discipline of software engineering were computed and classified by Garousi et al. [24]. Based on the number of citations and the average annual number of citations, they discovered the best publications. They established a GQM (Goal, Question, and Metric) system to determine their goal by formulating certain questions. They also compared the top papers to the top papers from other fields of study. They also determined which fields the highest cited papers belong to. They also pinpointed the sites where the best papers are delivered.

In his research, Muhammad Kamran et al. [25] presents a bibliometric analysis of articles in the Blockchain in the Internet of Things (BIoT) sector, encompassing papers published in prominent journals and conferences, and identifies research trends. It also looks at various study disciplines, the most prominent publications, the best publication venues, the best funding agencies, and the future of research. Our research will be quite similar to this, however our focus will

be on data science.Missen et al. [26] and others give a case study of Pakistan to examine the impact on international research from 2009 to 2018. The study looks at 2000 articles written by 50 research scholars from various areas. This research is carried out on three different levels: researcher, field, and domain. Readability scores, title formats, single and multiple authorships of papers, citation rates, publishing rates over time, the research contribution of both genders, and the impact of authors' Ph.D. institutions on research publications are all discussed in this work.

Fiala el al. [27] gives a bibliometric analysis of 1.9 million computer science papers indexed in Web of Science from 1945 to 2014. They examine the amount as well as the impact of these publications by document type, language, discipline, country, institution, and publication source. The most common keywords, cited references, and cited articles are also investigated, as well as the distribution of the number of references and citations per work and the age of cited references. They analyze the time and place of computer science conferences in terms of the most prolific months and locales because conference proceedings play such an important role in this scientific discipline. Finally, the production of journal articles and conference papers throughout the study, as well as the level of collaboration within different computer science disciplines, are examined. One of the most important findings is that "Artificial Intelligence" is the most popular field in computer science. Ding et al. [28] conducted a syntactic and semantic analysis of citations in computer science papers. The syntactic part entails identifying the location where the citations can be found (i.e. in which section of the article). Through a manual technique of predefined categorizations or an automated approach of NLP, semantic analysis determines the reason for citations. Citations are classified into categories defined by words or phrases in a decision tree in a predetermined classification.

III. DATA COLLECTION

The data on publications by Pakistani authors is the most essential part of our study. We gathered data from Web of Knowledge of Clarivate Analytics, commonly known as Web of Science. Web of Science is a subscription-based website that offers access to different citation databases. These databases include detailed citation information for a variety of academic areas. The Institute for Scientific Information created the Web of Sciences, which is now maintained by Clarivate Analytics. It has two search modes: basic search and advanced search. The advanced search option allows the users to build a query based on their needs and retrieve the required information. The data for publications by Pakistani researchers was generated using the query in table III.

The the abbreviated terms in table III represents **CU** as Country/Region, **SU** as Research Area, **SCI_EXPANDED** as Science Citation Index Expanded, **SSCI** as Social Science Citation Index, **A&HCI** as Arts & Humanities Citation Index, **ESCI** as Emerging Sources Citation Index. The query mentioned above returned a total of 8,807 publications in

Query
CU=*(PAKISTAN)* AND **SU**=*(Computer Science)* **Indexes** = *SCI-EXPANDED, SSCI, A&HCI, ESCI* **Timespan** = *(All years)*

TABLE 1. Query to get data from Web of Knowledge

English with a few exceptions in other languages. The results were extracted and stored in the form of text (.txt) file. The data-set contains 8807 rows where each row represents single publication and the associated 60 columns represents different information about the publication.

REFERENCES

[1] Dodson, M. V. "Research paper citation record keeping: It is not for wimps." Journal of animal science 86.10 (2008): 2795-2796.

[2] Thomson Reuters. History of citation indexing. Essay in Free Scientific Resources. http://thomsonreuters.com/products_services/science/free/essays/history_of_citation_indexing/.

[3] Garfield, Eugene. "Citation indexes for science. A new dimension in documentation through association of ideas." International journal of epidemiology 35.5 (2006): 1123-1127.

[4] Garfield, Eugene. "Citation frequency as a measure of research activity and performance." Essays of an Information Scientist 1.2 (1973): 406-408.

[5] Garfield, Eugene. "The use of journal impact factors and citation analysis for evaluation of science." 41st Annual Meeting of the Council of Biology Editors, Salt Lake City, UT. 1998.

[6] Canese, Kathi, and Sarah Weis. "PubMed: the bibliographic database." The NCBI Handbook 2 (2013): 1.

[7] Dittmar, Paul G., Robert E. Stobaugh, and Charles E. Watson. "The chemical abstracts service chemical registry system. I. General design." Journal of Chemical Information and Computer Sciences 16.2 (1976): 111-121.

[8] TePaske-King, Bert, and Norman Richert. "The Identification of Authors in the Mathematical Reviews Database." Issues in Science and Technology Librarianship 31 (2001).

[9] Bergmark, Donna, Paradee Phempoonpanich, and Shumin Zhao. "Scraping the ACM digital library." ACM SIGIR Forum. Vol. 35. No. 2. New York, NY, USA: ACM, 2001.

[10] Caragea, Cornelia, et al. "Citeseer x: A scholarly big dataset." European Conference on Information Retrieval. Springer, Cham, 2014.

[11] Analytics, Clarivate. "Web of science." (2017).

[12] Burnham, Judy F. "Scopus database: a review." Biomedical digital libraries 3.1 (2006): 1-8.

[13] Mongeon, Philippe, and Adèle Paul-Hus. "The journal coverage of Web of Science and Scopus: a comparative analysis." Scientometrics 106.1 (2016): 213-228.

[14] Falagas, Matthew E., et al. "Comparison of PubMed, Scopus, web of science, and Google scholar: strengths and weaknesses." The FASEB journal 22.2 (2008): 338-342.

[15] Jacsó, Péter. "Google Scholar: the pros and the cons." Online information review (2005).

[16] Waltman, Ludo, and Vincent A. Traag. "Use of the journal impact factor for assessing individual articles need not be statistically wrong." F1000Research 9 (2020).

[17] Curry, Stephen. "Let's move beyond the rhetoric: it's time to change how we judge research." Nature 554.7690 (2018): 147-148.

[18] Lowry, Paul Benjamin, et al. "Evaluating journal quality and the association for information systems senior scholars' journal basket via bibliometric measures: Do expert journal assessments add value?." MIS quarterly (2013): 993-1012.

[19] da Silva, Jaime A. Teixeira, and Aamir Raoof Memon. "CiteScore: A cite for sore eyes, or a valuable, transparent metric?." Scientometrics 111.1 (2017): 553-556.

[20] Mañana-Rodríguez, Jorge. "A critical review of SCImago journal & country rank." Research evaluation 24.4 (2015): 343-354.

[21] Iqbal, Waleed, et al. "A bibliometric analysis of publications in computer networking research." Scientometrics 119.2 (2019): 1121-1155.

[22] Ali, Muhammad Yousuf, and Joanna Richardson. "Research publishing by library and information science scholars in Pakistan: A bibliometric analysis." Journal of Information Science Theory and Practice 4.1 (2016): 6-20.
[23] Bajwa, Rizwan S., and K. Yaldram. "Bibliometric analysis of biotechnology research in Pakistan." Scientometrics 95.2 (2013): 529-540.
[24] Garousi, Vahid, and João M. Fernandes. "Highly-cited papers in software engineering: The top-100." Information and Software Technology 71 (2016): 108-128.
[25] Kamran, Muhammad, et al. "Blockchain and Internet of Things: A bibliometric study." Computers & Electrical Engineering 81 (2020): 106525.
[26] Missen, Malik Muhammad Saad, et al. "Scientometric analysis of social science and science disciplines in a developing nation: a case study of Pakistan in the last decade." Scientometrics 123.1 (2020): 113-142.
[27] Fiala, Dalibor, and Gabriel Tutoky. "Computer science papers in Web of Science: A bibliometric analysis." Publications 5.4 (2017): 23.
[28] Ding, Ying, et al. "Content-based citation analysis: The next generation of citation analysis." Journal of the association for information science and technology 65.9 (2014): 1820-1833.