Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 09:2023
DOI: 10.5281/zenodo.8343425

AN EXTENDED BESPOKE APPROACH TO EVALUATE THE QUALITY OF UK UNIVERSITIES RESEARCH USING MULTIPLE SOURCES

NOOR UL SABAH

Ph. D Candidate, Department of Computer Science, Government College University, Faisalabad, Pakistan. Email: n_nnoor@yahoo.com

MUHAMMAD MURAD KHAN*

Assistant Professor, Department of Computer Science, Government College University, Faisalabad, Pakistan. Corresponding Author Email: muradtariq.tk@gmail.com

RAMZAN TALIB

Professor and Chairman, Department of Computer Science, Government College University, Faisalabad, Pakistan. Email: ramzan.talib@gcuf.edu.pk

BUSHRA ZAFAR

Assistant Professor, Department of Computer Science, Government College University, Faisalabad, Pakistan. Email: bushrazafar@gcuf.edu.pk

MUHAMMAD KASHIF HANIF

Assistant Professor, Department of Computer Science, Government College University, Faisalabad, Pakistan. Email: mkashifhanif@gcuf.edu.pk

Abstract

People's reliance on the internet for acquiring institutional rankings to make comparisons has increased significantly in this day and age. Recently, researchers have begun to correlate institutional rank lists provided by government organizations such as the Research Excellence Framework (REF, UK) with ranked lists compiled by Google Scholar (GS). In this technological era, where well established tools Google Scholar and Microsoft Academic are complementing conventional sources Scopus and Web of Science. Unlike Google Scholar still provides the maximum coverage in contrast of other sources. To address this gap why is it so? An empirical investigation has been performed in comparison to Google Scholar, Web of Science, Scopus and Microsoft Academic. The data obtained from these sources is turned into structured data and subsequently used to perform statistical analyses on citations. These analyses are then used to determine rankings based on citation and publication metrics. Most popular of all is the bespoke algorithm that produces a correlation greater than 0.78 between google scholar and REF generated rank list. This study extends the bespoke algorithm by re-implementing it in python language; integrating exceptional handling and client-server architecture to make it scalable, and finally excluding duplicate author's profiles for refining consolidated citations before generating the rank list of the institutes using Google Scholar [1]. These methods are specialized in extracting the institutional citation but lack to capture feature dependencies effectively with other sources. After the experiment, the results compared to check whether the results are altered or the institutes ranking changed by altering the source. This study stood for the utilization of a predictive approach within the framework of correlational research. Moreover, the experimental evaluation yielded promising results when compared to the other multiple data extraction sources which implemented by the proposed taxonomy. This paper examines the existing evaluation limitations and proposes potential strategies for enhancing the research impact algorithm.

Index Terms: Google Scholar, Microsoft Academic, Scopus, Web of Science, Citation Data, Research Excellence Framework, Institutional Data, Impact Evaluation.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 09:2023
DOI: 10.5281/zenodo.8343425

1. INTRODUCTION

World Wide Web comprises information from a variety of areas such as social, educational, financial, entertainment, etc. Most users explore these areas using a search engine that indexes data available on the related websites. Web scraping is the main technique working behind indexing web pages [2]. Web scraping is used to extract information from unstructured data, store it and present it in a structured manner. The most popular example of web scraping is the Google search engine which extracts information from the web pages and indexes it for its users [3]. One specific area of the World Wide Web is academic publications available in popular document formats such as Microsoft Word files, portable document format (PDF), HTML file etc. Many companies have designed search engines for academic publications such as Google Scholar (GS), Microsoft Academic (MA), Scopus, and Web of Science (WoS). These search engines not only search through publications but also provide additional information such as the number of times an article has been cited, total citations an author has attained or total citations and organization have scored based on verified authors. Researchers have applied web scraping techniques to academic search engines for generating information not provided by the search engines. For example [4] extracts information from 146 scraped articles to identify a flaw in google scholar's h-index and proposes improvement under the name hIa-index. Similarly [5] utilizes data from 4,600 scraped articles the WoS and GS, for illustrating the difference in topic coverage between the two search engines. Whereas, in this paper, we focus on [6] which scrap publications data related to 130 UK universities from google scholar, WoS, Scopus and MA to compare it with UK Research Excellence Framework (REF) for ranking UK universities.

Ranking algorithms are used to order items in a dataset based on certain criteria. They are divided into deterministic and probabilistic algorithms. Search engines use these algorithms to rank webpages according to their relevance to the user's search query. Ranking, recommendation, and retrieval systems are employed in a variety of online and offline platforms, such as e-commerce, media streaming, admissions, gig platforms, and hiring. Recently, a large body of research has been developed to make these systems more efficient and beneficial for individual users, providers, and content [1]. This research typically defines equality for a single instance of retrieval, or as a cumulative measure for multiple instances of retrievals over time.

Multiple instances for retrievals refers to commercial databases such as WoS, Scopus, and search engines like GS apply scraping techniques on research articles, available in PDF or word format. These search engines extract the data and index it for searching purposes. Revealed that the search engine GS is considered the most reliable source for providing information more than other commercial databases. On the other hand, institutional repositories have less coverage than commercial databases. So, it is essential to choose the best source for extracting the information for further assessment and analysis, which can enhance the results and research quality [7]. Table 1 in literature review section showed a comparison between the various sources used to extract the citation and their different sample sizes. Exploration of numerous data samples was carried out by the researchers to rank them based on their quality factors of citation. An

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 09:2023
DOI: 10.5281/zenodo.8343425

analysis composed using data from multiple sources to mitigate the limitations according to their sources.

This research paper answers the following research question as the first three questions are already addressed in previous research [1] and now extended for more objective to achieve. The previous process is now mature enough to remove its discrepancies. Due to its abandoned database, there is a lack of built-in mechanisms to handle the larger data which is redundant also. For this purpose, existing code is upgraded to deal with the institutes having sample sizes from multiple sources.

RQ1. Extended the current algorithm utilized for the elimination of redundant profiles in order to enhance the generation of institutes output. [1]

RQ2. Extended the algorithm to incorporate the exclusion of theses and dissertations, resulting in the retrieval of authentic documentations. [1]

RQ3. The algorithm was extended to incorporate the exclusion of documents with falsified ownership from scholar profiles. [1]

RQ4. Extended the existing algorithm for utilizing multiple inputs sources for output generation.

This research contributes to answering the above research question pertains to the GS updated ranking algorithm methodology. The following outline constitutes the structure of this research paper: In Section 2, we will give a review of the relevant literature. In Section 3, you will find a description of the study's methodology. The discussion and the findings are provided in the next section. Section 5 contains the conclusion of this paper. The limitations of the study as well as suggestions for further research are discussed in Section.6.

2. LITERATURE REVIEW

A vast body of literature exists that compares the academic publication and/or citation coverage provided by various databases, namely the WoS, Scopus, and GS. Notable recent examples of such comparative studies include the works of [8] and [9]. In addition to these databases, there is a growing interest in examining the capabilities of Microsoft Academic, Scopus, WoS as evidenced by studies conducted by [10], [11], [12], [13] and [14]. In the second examination, both MA and GS had a similar impact on the data, with the exception of one cited institution. This institution provided data from 145 universities across five distinct fields. To mitigate the impact of MA, an alternative to WOS and Scopus is employed. In order to establish a conclusive evaluation of MA, it is imperative for scholars to undertake a comparative analysis of university rankings using diverse samples. This study examines the academic coverage of MA one year after its prelaunch [15], [16].

The GS search engine is dedicated to scholarly literature, which includes articles, theses, books, conference papers, and other academic resources. It offers access to full-text papers as well as citations from academic sources and enables users to search for

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 09:2023
DOI: 10.5281/zenodo.8343425

scholarly publications across a range of subjects. GS has implemented a ranking system for UK universities, which is based on their research performance and impact. This algorithm utilizes various metrics to determine the rankings, such as the h-index and the i10-index. The h-index measures the researcher's productivity and impact, while the i10-index tracks the number of articles published with at least ten citations. Moreover, the total number of citations received by a university's publications is also taken into account. This ranking system emphasizes the research output and impact of universities and provides a comprehensive and impartial assessment of their academic excellence [17], [18]. First, Institutional ranking using multiple sources like WOS, Scopus, GS and MA, as shown in table 1. The institutes' ranking can be any indicator like citation-based, h-index, and i10 based. Different sources do not show so much discrimination while ranking scholars' profiles or journal articles. However, ranking based on the indicators creates variation in results [19], [20]. It is also determined that the URL-based method of GS is an effective metric for evaluating university ranking in which derived the official URL of institutes.

In recent years, some new sources of data, such as PubMed, MA (2016) [21],[22] CrossRef (2017), and Dimension (2018), have been in the race to provide free citation coverage with GS, by complementing other two sources: WOS and Scopus [23],[24]. Crossref and Dimensions have better coverage of citations like WOS and Scopus but less than GS. However, MA received some consideration from the bibliometric community and retained citation and publication sites. A study compared these sources by their citations and concluded that MA has more coverage than WOS and Scopus but lower than GS [25]. However, GS has more citations among all the sources of bibliometric data like WOS, Scopus, and MA [26]. According to a recent study, GS interface has become a standard data retrieval process for generating an output for the Institute's ranking due to its open access and dynamicity [27].

The relative portion of the data source is shown in table 1 which is the most usable source of all. After analysis, various studies over the past years have shown that GS provides the best coverage for citing data and further classifies them by their institutional status, cited documents, and journal groups. As GS held the uncontrolled growth for its metrics, the h-index, and h-5 median of journals were analyzed and compared for average, minimum, and maximum values to check the impact on the correlation coefficient for the rankings. Based on the existing algorithm, it only used the GS data [28], [29]. GS depends on the academics that created a citation profile, while institute policy makes sure every scholar created its profile, which endorses the institute usage also [30].

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 09:2023
DOI: 10.5281/zenodo.8343425

Table 1: State-of-the-art discoveries of Multiple Data Extraction Sources

Study	Google Scholar	Scopus	Web of Science	Microsoft Academic	Sample Data	Limitations
[31]	✓ √√	×	×	×	64000 Documents	GS has detected more than one version for the documents.
[32]	×	√	×	×	146 institutes	hIa-index has enough potential for societal advancement.
[33]	√	×	×	×	1000 Journals	GS should update its ranking twice a year.
[34]	√	√	√	×	146 institutes	A large number of duplicate papers.
[35]	√	×	×	×	130 UK institutes	Paper with false ownership creates fake citations.
[36]	√	√	√	√	145 institutes	How Microsoft academic can be the best alternative source for citation analysis
[37]	×	×	×	√	118 institutes	Multiple sources can be used for output generation
[38]	√	×	×	×	1000 cited documents	Group citation based on university category.
[39]	√	√	√	×	34 Journal articles	Comparison of a range of different databases to analyze the gaps.
[40]	√	√	√	×	1 academic record	Few studies investigated the coverage for all citation sources.
[41]	√	√	√	×	100 Turkish institutes	Lack of institutes Commitment to open access.
[42]	√	×	×	×	Top 100 research articles	Central Tendency metric for evaluation.

2.1 Problem Statement

The purposed frame work used state-of the-art dataset describe in Table.2. The existing study tried to validate if different online citation sources can be used for ranking universities as effectively as REF. The Existing study used GS [43] as an online source of citations and the effectiveness was measured using the other multiple sources like WoS, Scopus and MA for correlational analysis [44]. The most crucial part that is pertinent to the sources, why it is important to focus on a single source for future study, and the GS source has the best publication and citation metrics.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 09:2023
DOI: 10.5281/zenodo.8343425

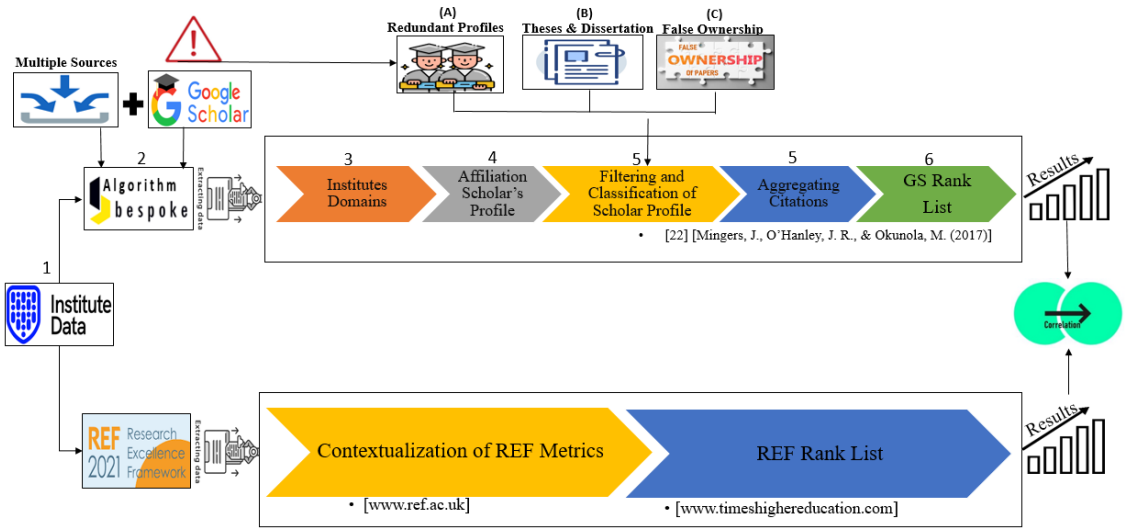


Figure 3: Multiple Source identification by Google Scholar University Ranking Algorithm

2.2 Multiple Sources Coverage

2.2.1. Google Scholar

GS provides maximum coverage for data extraction because it is a search engine specifically designed to search scholarly literature and academic resources. It indexes a wide range of scholarly publications, including articles, conference papers, theses, books, and preprints, from a variety of academic publishers, universities, and other sources. GS uses a sophisticated algorithm to identify and rank scholarly publications based on a variety of factors, including citation counts, authorship, and publication date. It also includes advanced search features that enables users to effectively refine their search results by utilizing specific criteria but not limited to the author's name, journal title, or publication date. In addition, GS is freely accessible to anyone with an internet connection, making it a popular tool for researchers, academics, and students around the world. This popularity has led to a large and diverse pool of data being available for extraction. Overall, GS provides maximum coverage for data extraction because it offers a comprehensive and easy-to-use search engine for scholarly literature and academic resources, and its popularity has led to a vast collection of scholarly publications being indexed and available for analysis [45], [46].

2.2.2. Scopus

The bibliographic database and abstracting service "Scopus" offers thorough literature coverage in the social sciences, technology, and medicine. Researchers, academics, and professionals utilize it as a tool to access academic journals, conference papers, and other research resources. Users can track research trends, find key articles, and assess the influence of their research by using the data provided by Scopus, which includes information on citations, author profiles, and journal metrics.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 09:2023
DOI: 10.5281/zenodo.8343425

2.2.3. Web of Science

A research database and citation index called "Web of Science" offers thorough coverage of scientific publications from a variety of areas. Researchers, academics, and professionals utilize it as a tool to access academic journals, conference proceedings, and other research resources. With the help of WoS, users may track citations, find influential papers, and assess research trends by accessing data on author affiliations, publications, and citations. It is widely recognized as an important tool for performing bibliometric analysis and assessing the significance of research [47].

2.2.4. Microsoft Academic

Research database and academic search engine "Microsoft Academic" are services offered by Microsoft. It provides a thorough selection of academic works from numerous areas, including articles, conference papers, and other research resources. For scholars, academics, and professionals looking to study and find pertinent academic literature, MA offers extensive search options. It also has tools like author profiles, citation analysis, and cooperation networks that let users analyze research trends, assess the impact of that research, and find possible research partners.

GS, Web of Science, and Scopus differ in several aspects related to data extraction. GS provides coverage for indexed publications, author profiles, and ranking compared to other sources like Scopus, WoS, and Research Gate due to several reasons:

- Inclusion of diverse Coverage: GS aims to provide a comprehensive coverage of scholarly literature across various disciplines, including journal articles, conference papers, theses, and more. It indexes a wide range of sources, including both traditional publishers and non-traditional sources like institutional repositories and preprint servers. WoS and Scopus, on the other hand, primarily focus on indexed journals and conference proceedings, providing a more selective coverage of scholarly literature. This inclusiveness allows for a broader coverage of publications, including those that may not be indexed by other platforms.
- Data Sources: GS indexes content from a variety of sources, including publishers, universities, and individual researchers' websites. WoS is produced by Clarivate Analytics and primarily indexes content from major scholarly journals. Scopus, produced by Elsevier, also indexes scholarly journals but has a broader coverage that includes conference papers, book chapters, and patents.
- Freely accessible content: GS provides access to a significant amount of full-text content that is freely available on the web. This includes open access articles and content available on institutional repositories. This openness contributes to a broader coverage of scholarly materials that may not be accessible through other proprietary platforms.
- Data Extraction Platforms: While all three platforms offer search capabilities, they differ in their data extraction tools. GS provides a search engine that allows users to search for specific articles or keywords. It also provides the option to export search

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 09:2023
DOI: 10.5281/zenodo.8343425

results in various formats. WoS and Scopus offer more advanced search features by incorporating specific parameters, such as author, publication year, and journal, for result refinement.

- Citation Analysis: GS, WoS, and Scopus all provide citation analysis capabilities, but they differ in the sources and methodologies used. GS uses a broader range of sources to identify citations, including books, theses, and non-traditional sources. WoS and Scopus focus on indexed journals and conference proceedings for citation data. This comprehensive approach ensures a greater coverage of research outputs across various disciplines.
- Metrics and Rankings: Each platform offers its own set of metrics and rankings to evaluate the impact and influence of scholarly works. GS's ranking algorithm emphasizes citation counts, while WoS includes metrics like the Journal Impact Factor and the h-index. Scopus provides similar metrics like the Source Normalized Impact per Paper (SNIP) and the SCImago Journal Rank (SJR).
- Inclusion of gray literature: GS also indexes gray literature, which refers to research outputs that are not formally published or peer-reviewed, such as conference presentations, working papers, and reports. The inclusion of gray literature increases the coverage of research and allows for a more comprehensive understanding of the scholarly landscape.
- Author profile aggregation: GS allows researchers to create and manage their profiles, which automatically compile their publications and citations from various sources. This aggregation of author profiles provides a convenient way for researchers to showcase their work and facilitates comprehensive analysis of their research impact.
- Ranking based on citations: GS's ranking algorithm places a significant emphasis on citations, considering both the number of citations and the importance or relevance of the citing publications. This citation-based ranking system provides a useful metric for assessing the impact and influence of scholarly works.

While platforms like Scopus, WoS, and Research Gate also offer valuable resources for researchers, GS's approach to indexing, inclusiveness, freely accessible content, and author profile aggregation contribute to its broader coverage of indexed publications, author profiles, and ranking. Following table 1 figured out the parameters which multiple sources covered in data extraction sources.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 09:2023
DOI: 10.5281/zenodo.8343425

Table 2: Metrics Coverage of Multiple Data Extraction Sources [48]

Metrics	Sources			
	Google Scholar	Scopus	Web of Science	Microsoft Academic
Citation Count	●	●	●	●
Citation 5-year	●	●	●	●
h-index	●	●	●	○
h-index 5-year	●	●	●	○
i10-index	●	○	○	○
i10-index 5-year	●	○	○	○
Web content	●	●	●	●
hIa	●	●	●	○
Author Profiles	●	●	●	●
Co-Authorship Network	●	●	●	●
Field-Weighted Citation Impact	○	○	○	●
Journal Metrics	●	●	●	●
SCImago Journal Rank (SJR)	○	○	●	○
Institutional Affiliations	●	●	●	●
Source Normalized Impact per Paper (SNIP)	○	●	○	○
Bibliometric Analysis	●	●	●	●
Scholarly Output Count	○	○	○	●
Highly Cited Researchers	○	○	●	○
Journal Citation Reports (JCR)	○	○	●	○

3. METHODOLOGY

3.1 Data set

The present study utilizes an authentic UK-Universities dataset. This model combine the parameters that are common to all of these platforms in order to offer a comprehensive evaluation of research impact, productivity, and collaboration. We get a thorough understanding of elements like citation counts, h-index, publication records, and collaboration networks by combining these common metrics. The approach makes it possible to assess research output, identify key publications and contributors, and analyze cross-disciplinary research trends. The primary indicators included in this study were gathered from several sources and validated throughout the entire sample. It is additionally stated which datasets from the other numerous sources could be used most frequently. A thorough description of the selected dataset's essential metrics. It has been done in order to show comprehend dataset values in terms of its variables. Raw data filtration to provide useful data for the outcomes and the formula is developed to handle the values that were missing.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 09:2023
DOI: 10.5281/zenodo.8343425

Table 3: Dataset Description

Dataset Analysis	UK-REF	GS	MA	Scopus	WOS
Total Submissions	157	157	157	157	157
Ranked Institutions	130	130	130	137	137
Not Ranked Institutes	27	0	0	0	0
Institutes with no Domain Name	0	7	0	0	0
No match with multiple sources	3	3	3	3	3

3.2 Proposed Method

Existing studies have shown that bespoke code can effectively address issues such as redundant profiles, theses and dissertations, and falsely attributed citations by scholars. This enables the development of a reliable and efficient method for ranking institutes. The existing algorithm phases serve as a foundation for extracting data from Fig. 2. This study focuses on utilizing data collected through GS to perform various tasks for data preparation prior to further analysis.

ALGORITHM : **BESPOKE**

INPUT: **LIST OF UNIVERSITY DOMAINS**
OUTPUT: **CSV FILE CONTAINING CITATION MATRICS AGAINST EACH DOMAIN**

```
1   START
2     uni_list = [ list of university domains ]
3     FOR EACH uni IN uni_list:
4       uni_url = generate_url (uni)
5       Browser.open(uni_url)
6       authors = []
7       authors.append(Browser.getAuthorsURL())
8       FOR EACH author_url IN authors:
9         Browser.open(author_url)
10        citations.append(Browser.getCitations())
11        save_csv(citations)
12      END FOR
13      IF "Next" BUTTON EXISTS:
14        Browser.click("Next")
15        GOTO LINE 7
16      END IF
17    END FOR
18  STOP
```

Figure 4: Algorithm for proposed Google Scholar Updated Ranking Algorithm

This study investigated the approach of the proposed model for ranking educational institutions utilizing the multiple sources like GS, Scopus, WoS and MA data. In May 2022, UK-REF dataset has been published which is known as REF, in order to understand the process. The model was developed to address existing deficiencies as outlined by (Mingers, 2017). The proposed taxonomy for the Google Scholar University Ranking

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 09:2023
DOI: 10.5281/zenodo.8343425

Algorithm (GSURA), which could substitute for the present bespoke algorithm, is provided. The enhanced GS Ranking algorithm includes all the methodological changes that have been performed and implied in order to cover research target.

This model integrates the shared metrics across these platforms to provide a comprehensive and unified assessment of research impact, productivity, and collaboration. By fusing these common metrics, the GSURA Model enables researchers to evaluate the influence of publications, measure productivity, assess collaboration networks, and track research trends across disciplines. The extent of coverage of different sources with regard to data extraction determines which source has the most range and the most reliable data. This study analyzed the coverage of various sources in terms of data extraction and identified the best source that offers the most accurate data with the potential for the greatest degree of coverage. According to a recent study, GS is a popular tool for obtaining citations to create the ranking list for institutional rating. Ranking lists can also be created using data from multiple additional sources including MA, WoS, and Scopus.

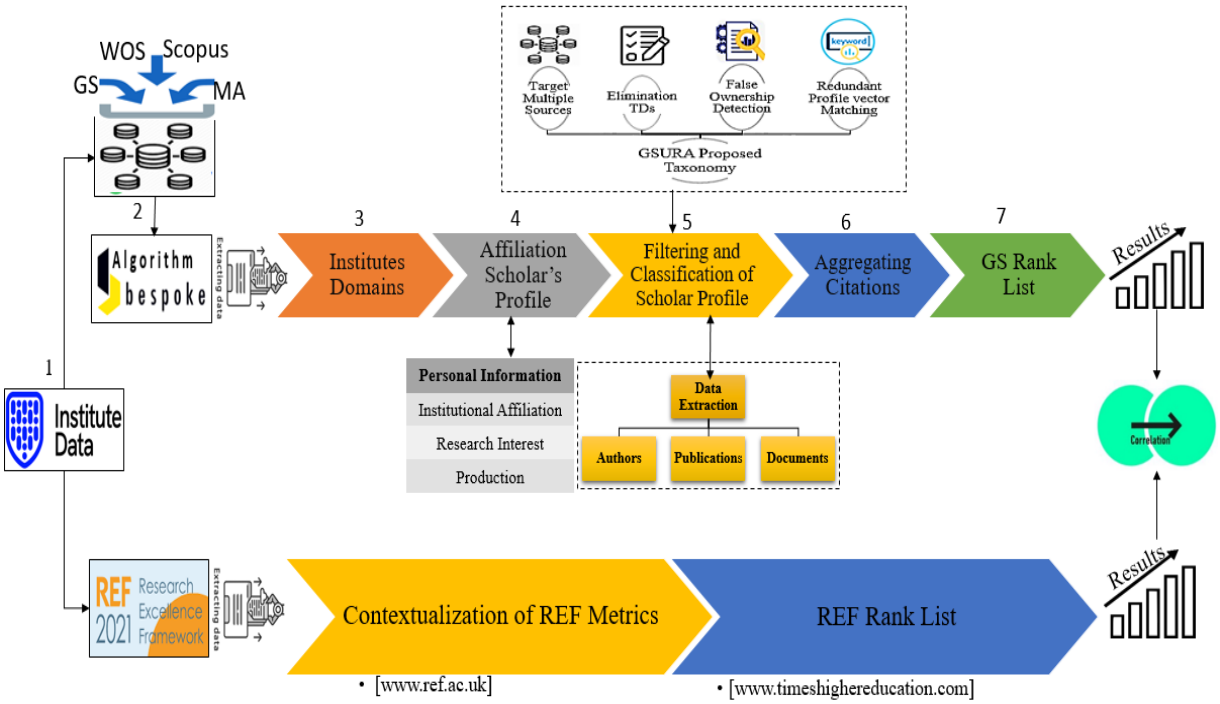


Figure 5: Proposed Multiple Source taxonomy by Google Scholar University Ranking Algorithm

To gain Unified Research Impact and Productivity accessed through multiple sources by incorporating the common metrics from GS, Scopus, WoS, and MA, following these steps:

- Define the common metrics: Determined which metrics which includes in this model. Common metrics can include citation counts, publication counts, and co-authorship data.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 09:2023
DOI: 10.5281/zenodo.8343425

- Data collection: After attaining the common metrics among multiple source at the same level, gather the required metric data for each platform, ensuring that extracted the necessary data for the selected metrics. Retrieve the data from GS, Scopus, WoS, and MA using their respective APIs or by web scraping techniques for further processing.
- Data preprocessing: The collected data manipulated by handling missing values, outliers, and inconsistencies. Standardize the data if required, ensuring that the metrics are on a similar scale for fair comparison.
- Combine the metrics: Deciding on the weighting scheme or mathematical formula to combine the selected metrics into a single composite score for each researcher or institution. This step involves assigning weights to different metrics based on their relative importance or applying mathematical transformations to normalize the metrics.
- Calculate correlations: Once the composite scores for each researcher or institution is set, correlation coefficients calculates using a statistical method. This helps in measuring the strength and direction of the relationship between the scores of different researchers or institutions.
- Interpret the correlations: Analyzing the correlation coefficients to interpret the relationships between the multiple scores. A positive correlation indicates a direct relationship, while a negative correlation suggests an inverse relationship. The magnitude of the correlation coefficient represents the strength of the relationship.

3.3 Ranking Algorithm metrics Calculation

3.3.1 Mean Imputation (MI):

Mean imputation is a statistical technique used to fill in missing values in a dataset with the mean value of the non-missing values in the same variable. It is a simple and commonly used method for dealing with missing data in a dataset. The basic idea behind mean imputation is to replace the missing values with the average or mean value of the variable. This can be done for a single missing value or for multiple missing values in a column of data. The mean value is calculated based on the non-missing values in the column. The imputed values are then used in further analysis or modeling. The formula for mean imputation can be expressed mathematically as follows in Eq.1.

$$x_{i,j(\text{imputed})} = x_{m,j} \quad (1)$$

In this equation, xi represent data extracted by source and j (imputed) means to find out those variables who do not carry any data against variable and impute them in missing values spaces.

3.3.2 Correlational Coefficients (CC)

Collect the data: Gather the corresponding data from GS and your institutional data for the variables of interest (e.g., publication counts, citation counts, or impact factors.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 09:2023
DOI: 10.5281/zenodo.8343425

Calculate the means: Calculate the mean (average) of each variable.

Calculate the differences: Calculate the differences between each data point and its respective mean for both variables.

Calculate the products: Multiply the differences for each data point of the two variables.

Calculate the sum of the products: Sum up the products obtained in the previous step.

Calculate the standard deviations: Calculate the standard deviation for each variable

Multiply the standard deviations: Multiply the standard deviations of the two variables

Calculate the correlation coefficient: Divide the sum of the products by the multiplication of the standard deviations. This gives you the correlation coefficient (r).

Interpret the correlation coefficient: The correlation coefficient (r) is a statistical measure that quantifies the strength and direction of the linear relationship between two variables. It is a scalar value that ranges from -1 to 1, inclusive. A positive numerical value denotes a positive correlation between two variables, while a negative numerical value signifies a negative correlation. Conversely, a numerical value that is in close proximity to zero suggests a weak or negligible correlation between the variables under consideration. The correlation coefficient's magnitude serves as an indicator of the strength of the correlation, with values that are closer to 1 or -1 suggesting a more robust correlation.

$$\mathrm{r} = \frac{n(\Sigma x\, y) - (\Sigma\, x)\,(\Sigma y)}{\sqrt{[\mathrm{n}\,\Sigma\,\mathrm{x}^{2} - (\Sigma\,\mathrm{x})^{2}]\;[\mathrm{n}\,\Sigma\, y^{2} - (\Sigma\,\mathrm{y})^{2}]}} \qquad (2)$$

4. RESULTS AND DISCUSSION

This study examines various sources utilized at the institutional level, namely WoS, Scopus, GS and MA. Various indices, comprising citation-based, h-index, and i10-based, can be employed for institutional ranking. The bias in comparing scholars' profiles or journal article rankings is not evident across various sources. The implementation of indicators for ranking purposes yields diverse results [48]. The analysis carried out GS as the most dependable source for data coverage compared to Scopus, WoS, and MA.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 09:2023
DOI: 10.5281/zenodo.8343425

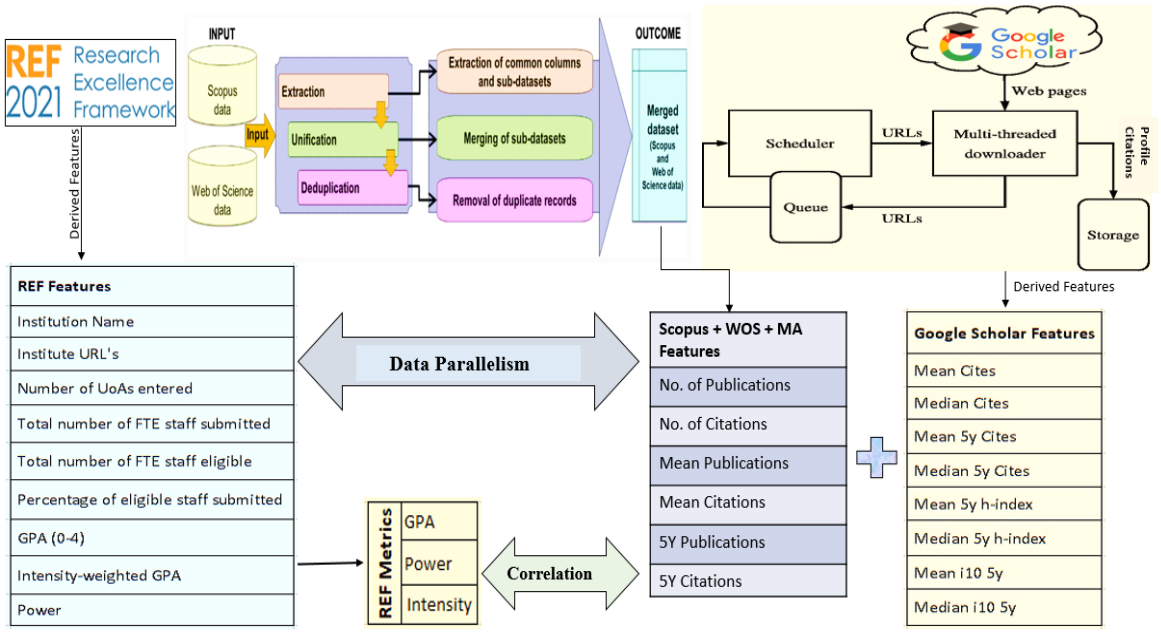


Figure 6: Integration of Proposed Multiple Source taxonomy for data extraction factors

In recent research, various bibliometric sources have been compared in a cross-sectional manner. In its attempt of identifying higher citation coverage, MA conducted a search and determined that it encompasses a greater number of citations compared to WoS and Scopus. Nevertheless, it is worth noting that GS remains inclusive, as evidenced in the following table.4. In this study, the corrections factors are higher due to its relevance to the existing results. These calculations are based on samples and Mean, Median which is supposed to be a where the bespoke code was modified to make it more scalable and reliable which consumed fewer resources and provided the best coverage for collecting the citation data against institutes and Scholar's profile.

In this study, where the corrections factor was not higher due to its change of source and relevancy to the benchmark study. GS and MA are correlated with a reasonable correlation among them. Despite its change of platform, this procedure results contribute to supporting GS citations and publications as a useful metric for an institute's rank and profile.

The correlation results based on publications are lower in Scopus and the WoS compared to GS and MA. But in 5 year correlation in terms of publication GS and Scopus are greater in average. The correlation results based on citations are lower in Scopus and the WoS compared to GS and MA. But in 5 year correlation in terms of publication GS and Scopus are greater in average.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 09:2023
DOI: 10.5281/zenodo.8343425

Table 4: Multiple Source results comparison in terms of Correlation

Publications Correlation					
GS:MA	GS:SCO	GS:WOS	MA:SCO	MA:WOS	SCO:WOS
0.96	0.89	0.91	0.88	0.82	0.90

Mean Publications					
GS:MA	GS:SCO	GS:WOS	MA:SCO	MA:WOS	SCO:WOS
0.82	0.94	0.89	0.91	0.93	0.88

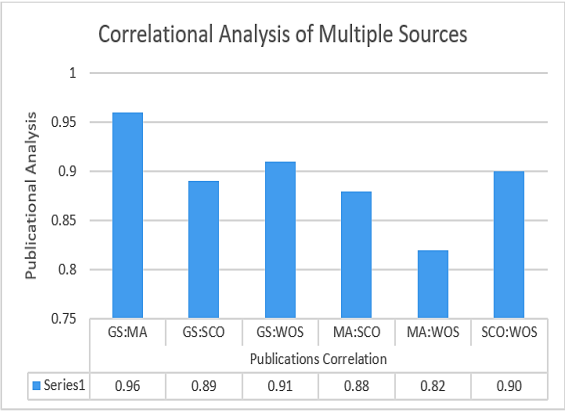
Publications 5year Correlation					
GS:MA	GS:SCO	GS:WOS	MA:SCO	MA:WOS	SCO:WOS
0.88	0.94	0.86	0.81	0.76	0.71

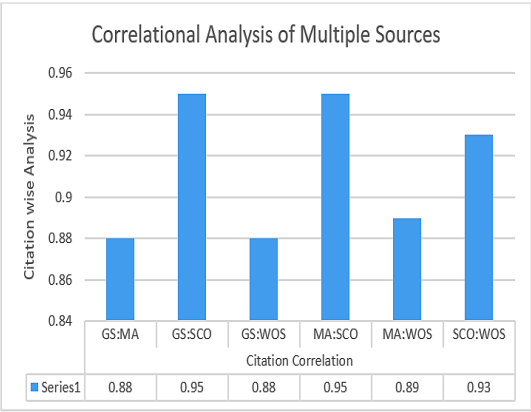
Citation Correlation					
GS:MA	GS:SCO	GS:WOS	MA:SCO	MA:WOS	SCO:WOS
0.88	0.95	0.88	0.95	0.89	0.93

Citation 5year Correlation					
GS:MA	GS:SCO	GS:WOS	MA:SCO	MA:WOS	SCO:WOS
0.87	0.94	0.90	0.94	0.86	0.93

Mean Citation Correlation					
GS:MA	GS:SCO	GS:WOS	MA:SCO	MA:WOS	SCO:WOS
0.76	0.94	0.86	0.90	0.93	0.89

All these combinations showed consistent growth while yielding diverse outcomes upon each time the data is extracted from the search engine GS, MA, Scopus and WoS in terms of the number of published papers which increased the citations in a scholar's profile. As the primary metrics are not well researched by all the citation sources. However, Mean or Median is mainly used as a central tendency for better results.





Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 09:2023
DOI: 10.5281/zenodo.8343425

Figure 5: Multiple Source results comparison in terms of Correlation

5. CONCLUSION

In today's era, people rely on the information acquired from the web system. This study explored the REF (Resource excellence framework) tool for the institute's ranking using their cited data extracted from GS. The other data sources: MA, Scopus, and WoS also analyzed to choose the utmost reliable source for data extraction. For this purpose, the Bespoke program is modified, as the already published code is in R language, and changed to python after removing its discrepancies which in resultant increased the scalability of the code. After conducting a comprehensive examination, it has been established that GS exhibits exceptional qualities in terms of consistency, validity, and certainty. This is primarily attributed to its ability to offer a substantial number of highly cited documents and generate a maximum output. The data is extracted from GS, Scopus, WoS and MA for the UK-REF universities in order to achieve more substantial improvements in its outcomes after excluding redundant profiles, domain names with no profiles, and profiles with no citations which ranked an institute falsely even if the institute doesn't secure that position. After the experiment, it apparently shows the promising result by obtaining 0.96% correlational coefficient among GS and MA in terms of

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 09:2023
DOI: 10.5281/zenodo.8343425

publications, 0.95% correlational coefficient for Scopus and GS in terms of citations, and 0.95% for Scopus and MA. One important point to note that the specific calculation methodologies and weighting schemes depends on research goals, preferences, and the availability of data. It is essential to carefully consider the limitations and assumptions of correlation analysis based on your specific requirements.

6. FUTURE WORK AND LIMITATIONS

First and foremost, it is essential to acknowledge that the incorporation of publications holds significant importance publications in an academic profile should not exclude those that were published under different affiliations, such as the scholar's previous university. Furthermore, the current algorithm lacks the capability to generate group citations according to university categories such as Technology University, Science University, social sciences, and so on. In addition, it is worth noting that comparisons can be made among other international ranking ministries instead of REF and African organizations that operate independently from their corresponding ministries. Additionally, it is crucial to acknowledge that each modification implemented in the procedure has the potential to impact the consolidated outcomes and modify the rankings of institutions, thereby resulting in diverse correlation results. Hence, it is important to enhance the algorithm in order to produce precise outcomes.

Acknowledgment

The authors express gratitude to Government College University for provision of resources in support of this research.

References

1) N. U. Sabah, M. M. Khan, R. Talib, M. Anwar, M. S. Arshad Malik, & P. N. Ellyza Nohuddin, “Google Scholar University Ranking Algorithm to Evaluate the Quality of Institutional Research,” *Computers, Materials & Continua*, vol. *75, no.* 3, pp. 4955-4972, Jun. 2023, doi: 10.32604/cmc.2023.037436.

2) D. S. Sirisuriya, “A Comparative Study on Web Scraping,” *Proceedings of 8th International Research Conference*, pp. 135-140, Nov. 2015, [Online]. Available: http://ir.kdu.ac.lk/handle/345/1051.

3) R. Lawson, “Web Scraping with Python,” Birmingham, UK: Packt Publishing Ltd, pp. 1-372, Oct. 2015.

4) A.-W. Harzing, S. Alakangas, and D. Adams, “hIa: An individual Annual h-index to Accommodate Disciplinary and career length differences,” *Scientometrics*, vol. 99, no. 3, pp. 811–821, Dec. 2014, doi: 10.1007/s11192-013-1208-0.

5) J. Mingers and E. Lipitakis, “Counting the citations: A comparison of web of science and Google Scholar in the field of business and management,” *Scientometrics*, vol. 85, no. 2, pp. 613–625, Jul. 2010, doi: 10.1007/s11192-010-0270-0.

6) J. Mingers, J. R. O’Hanley, and M. Okunola, “Using Google Scholar institutional level data to evaluate the quality of university research,” *Scientometrics*, vol. 113, no. 3, pp. 1627–1643, Oct. 2017, doi: 10.1007/s11192-017-2532-6.

7) M. y. Tsay, Y.-w. Tseng, and T.-l. Wu, “Comprehensiveness and uniqueness of commercial databases and open access systems,” *Scientometrics*, vol. 121, no. 3, pp. 1323–1338, Oct. 2019, doi: 10.1007/s11192-019-03252-3.

8) A. W. Harzing, and S. Alakangas, “Google Scholar, Scopus and the Web of Science: A longitudinal

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 09:2023
DOI: 10.5281/zenodo.8343425

and cross-disciplinary comparison," *Scientometrics*, vol. 106, no. 2, pp. 787–804, Feb. 2016, doi: 10.1007/s11192-015-1798-9.

9) A. Martín-Martín, E. Orduna-Malea and E. D. López-Cózar, "A novel method for depicting academic disciplines through Google Scholar citations: The case of bibliometrics," *Scientometrics*, vol. 114, no. 3, pp. 1251–1273, 2018.

10) A. W. Harzing, & S. Alakangas, "Microsoft Academic: is the phoenix getting wings?." *Scientometrics*, vol. 110, no. 1, pp. 371-383, Jan. 2017, doi: 10.1007/s11192-016-2185-x.

11) A. W. Harzing and S. Alakangas, "Microsoft Academic is one year old: The Phoenix is ready to leave the nest," *Scientometrics*, vol. 112, no. 3, pp. 1887–1894, Sep. 2017, doi: 10.1007/s11192-017-2454-3.

12) S. E. Hug, & M. P. Brändle, "The coverage of Microsoft Academic: Analyzing the publication output of a university," *Scientometrics*, vol. 113, no. 1, pp. 1551-1571, Dec. 2017, doi: 10.1007/s11192-017-2535-3.

13) S. E. Hug, M. Ochsner, & M. P. Brändle, "Citation analysis with microsoft academic," *Scientometrics*, vol. 111, no.1, pp.371-378, Oct. 2017. Doi: 10.1007/s11192-017-2535-3.

14) M. Thelwall, "Microsoft Academic: A multidisciplinary comparison of citation counts with Scopus and Mendeley for 29 journals," *Journal of Informetrics*, vol. *11, no.* 4, pp. 1201-1212, Nov. 2017, doi: 10.1016/j.joi.2017.10.006.

15) J. M. Iqbal, M. W. Iqbal, M. Anwar, M. M. Khan, J. A. Nazimi, and N. M. Ahmad, "Brain Tumor Segmentation in Multimodal MRI Using U-Net Layered Structure," *CMC-COMPUTERS MATERIALS & CONTINUA*, vol. 74, no. 3, pp. 5267-5281, Sep. 2023, doi: 0.32604/cmc.2023.033024.

16) M. Thelwall, "Microsoft Academic automatic document searches: Accuracy for journal articles and suitability for citation analysis," *Journal of Informetrics*, vol. 12, no. 1, pp. 1-9, Feb. 2018, doi: 10.1016/j.joi.2017.11.001.

17) E. Orduna-Malea, S. Aytac, and C. Y. Tran, "Universities through the eyes of bibliographic databases: a retroactive growth comparison of Google Scholar, Scopus and Web of Science," *Scientometrics*, vol. 121, no. 1, pp. 433–450, Aug. 2019, doi: 10.1007/s11192-019-03208-7.

18) A. Cheema, M. Tariq, A. Hafiz, M. M. Khan, F. Ahmad and M. Anwar, "Prevention Techniques against Distributed Denial of Service Attacks in Heterogeneous Networks: A Systematic Review," Security and Communication Networks, pp. 1-15, May. 2022, doi: 10.1155/2022/8379532.

19) M. Franceschet, "A comparison of bibliometric indicators for computer science scholars and journals on Web of Science and Google Scholar," *Scientometrics*, vol. 83, no. 1, pp. 243–258, Apr. 2010, doi: 10.1007/s11192-009-0021-2.

20) Y. A Sheikh, M. U. Maqbool, A. D. Butt, A. R. Bhatti, A. B. Awan, K. N. Paracha, and M. M. Khan, "Impact of rooftop photovoltaic on energy demand of a building in a hot semi-arid climate," *Journal of Renewable and Sustainable Energy*, vol. 13, no. 6, pp. 065101, Nov. 2021, doi: 10.1063/5.0063044.

21) A.-W. Harzing, "Microsoft Academic (Search): a Phoenix arisen from the ashes?," *Scientometrics*, vol. 108, no. 3, pp. 1637-1647, Jun. 2016, doi: 10.1007/s11192-016-2026-y.

22) M. M. Khan, M. Bakhtiari, and S. Bakhtiari, "An HTTPS approach to resist man in the middle attack in secure SMS using ECC and RSA," *In 2013 13th International Conference on Intellient Systems Design and Applications,* pp. 115-120, Dec. 2013, IEEE, doi: 10.1109/ISDA.2013.6920718.

23) A. Martín-Martín, E. Orduna-Malea and E. D. López-Cózar, "A novel method for depicting academic disciplines through Google Scholar citations: The case of bibliometrics," *Scientometrics*, vol. 114, no. 3, pp. 1251–1273, Sep, 2018, doi: 10.1007/s11192-020-03690-4.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 09:2023
DOI: 10.5281/zenodo.8343425

24) M. S. Aliero, I. Ghani, S. Zainudden, M. M. Khan, and M. Bello, “Review on SQL injection protection methods and tools,” *Jurnal Teknologi*, vol. 77, no. 13, pp. 49-66, Nov. 2015, doi: 10.11113/jt.v77.6359.

25) G. Etxebarria and M. Gomez-Uranga, “Use of Scopus and Google Scholar to measure social sciences production in four major Spanish universities,” *Scientometrics*, vol. 82, no. 2, pp. 333–349, Jun. 2010, doi: 10.1007/s11192-009-0043-9.

26) M. M. Khan, M. Bakhtiari, and S. Bakhtiari, “An HTTPS approach to resist Man in the Middle attack in secure SMS,” *Journal of Information Assurance and Security*, vol. 9, no. 3, pp. 157-166, 2014.

27) M. K. Merga, S. Mat Roni, and S. Mason, “Should Google Scholar be used for benchmarking against the professoriate in education?,” *Scientometrics*, vol. 125 no. 1, pp. 2505-2522, Sep. 2020, doi: 10.1007/s11192-020-03691-3.

28) S. Bangani, “The impact of electronic theses and dissertations: a study of the institutional repository of a university in South Africa,” *Scientometrics*, vol. 115, no. 1, pp. 131–151, Jan. 2018, doi: 10.1007/s11192-018-2657-2.

29) B. Hameed, M. M. Khan, A. Noman, M. J. Ahmad, M. R. Talib, F. Ashfaq, H. Usman and M. Yousaf, “A review of Blockchain based educational projects,” *International Journal of Advanced Computer Science and Applications,* vol. 10, no. 10, pp. 491-499, 2019.

30) A.-W. Harzing, “Running the REF on a rainy Sunday afternoon: Can we exchange peer review for metrics?,” *in STI 2018 Conference Proceedings*, 2017: Centre for Science and Technology Studies (CWTS). pp. 339-345, 2018, [Online]. Available: https://hdl.handle.net/1887/64521.

31) A. Martín-Martín, E. Orduña-Malea, J. M. Ayllón, and E. D. López-Cózar, “Does Google Scholar contain all highly cited documents,” *Research Gate*, vol. 1.2, pp. 1-97, Nov. 2014, doi: https://10.48550/arxiv.1410.8464.

32) A. W. Harzing, S. Alakangas, and D. Adams, “hIa: An individual annual h-index to accommodate disciplinary and career length differences,” *Scientometrics*, vol. 99, no. 3, pp. 811–821, Dec. 2014, doi: 10.1007/s11192-013-1208-0.

33) E. Orduña-Malea and E. D. López-Cózar, “Google Scholar Metrics evolution: an analysis according to languages,” *Scientometrics*, vol. 98, no. 3, pp. 2353–2367, Oct. 2014, doi: 10.1007/s11192-013-1164-8.

34) A. W. Harzing and S. Alakangas, “Google Scholar, Scopus and the Web of Science: a longitudinal and cross-disciplinary comparison,” *Scientometrics*, vol. 106, no. 2, pp. 787–804, Feb. 2016, doi: 10.1007/s11192-015-1798-9.

35) J. Mingers, J. R. O’Hanley, and M. Okunola, “Using Google Scholar institutional level data to evaluate the quality of university research,” *Scientometrics*, vol. 113, no. 3, pp. 1627–1643, Oct. 2017, doi: 10.1007/s11192-017-2532-6.

36) A. W. Harzing, & S. Alakangas, “Microsoft Academic: is the phoenix getting wings?” *Scientometrics*, vol. *110, no.* 1, pp. 371-383, Jan. 2017, doi: 10.1007/s11192-016-2185-x.

37) A. W. Harzing and S. Alakangas, “Microsoft Academic is one year old: The Phoenix is ready to leave the nest,” *Scientometrics*, vol. 112, no. 3, pp. 1887–1894, Sep. 2017, doi: 10.1007/s11192-017-2454-3.

38) A. Martín-Martín, E. Orduna-Malea, and E. D. López-Cózar, “A novel method for depicting academic disciplines through Google Scholar Citations: The case of Bibliometrics,” *Scientometrics*, vol. 114, no. 3, pp. 1251–1273, Nov. 2018, doi: 10.1007/s11192-017-2587-4.

39) M.-y. Tsay, Y.-w. Tseng, and T.-l. Wu, “Comprehensiveness and uniqueness of commercial databases and open access systems,” Scientometrics, vol. 121, no. 3, pp. 1323–1338, Oct. 2019, doi: 10.1007/s11192-019-03252-3.

Tianjin Daxue Xuebao (Ziran Kexue yu Gongcheng Jishu Ban)/
Journal of Tianjin University Science and Technology
ISSN (Online):0493-2137
E-Publication: Online Open Access
Vol: 56 Issue: 09:2023
DOI: 10.5281/zenodo.8343425

40) A.-W. Harzing, “Two new kids on the block: How do Crossref and Dimensions compare with Google Scholar, Microsoft Academic, Scopus and the Web of Science?,” *Scientometrics*, vol. 120, no. 1, pp. 341–349, May. 2019, doi: 10.1007/s11192-019-03114-y.

41) E. Orduna-Malea, S. Aytac, and C. Y. Tran, “Universities through the eyes of bibliographic databases: a retroactive growth comparison of Google Scholar, Scopus, and Web of Science,” *Scientometrics*, vol. 121, no. 1, pp. 433–450, Aug. 2019, doi: 10.1007/s11192-019-03208-7.

42) A. Furnham, “What I have learned from my Google Scholar and H index,” *Scientometrics*, vol. 122, no. 2, pp. 1249–1254, Dec. 2020, doi: 10.1007/s11192-019-03316-4.

43) M. W. Bramer, “Variation in number of hits for complex searches in Google Scholar,” *Journal of the Medical Library Association*, vol. 104, no. 2, pp. 143–145. Nov. 2016, doi: 10.3163/1536-5050.104.2.009.

44) R. N. Khan, J. C. Thompson, R. D. Taylor, S. K. Gabrick, F. A. Choudhri, R. F. Boop, et al., “Part II: Should the h-index be modified? An analysis of the m-quotient, contemporary h-index, authorship value, and impact factor,” *World Neurosurgery*, vol. 80, no. 6, pp. 766–774, Dec. 2013, doi: 10.1016/j.wneu.2013.07.011.

45) G. Nasreen, K. Haneef, M. Tamoor and A. Irshad, “A Comparative study of state-of-the-art skin image segmentation techniques with CNN. *Multimedia Tools and Applications*, vol. 82, no. 7, pp. 10921-10942, Sep. 2023, doi: 10.1007/s11042-022-13756-5.

46) S. da Teixeira, J. A, “The Google Scholar h-index: useful but burdensome metric,” *Scientometrics*, vol. 117, no. 1, pp. 631-635, Jul. 2018, doi: 10.1007/s11192-018-2859-7.

47) S. Mikki, “Google Scholar compared to Web of Science. A literature review,” *Nordic Journal of Information Literacy in Higher Education*, vol. 1, no. 1, pp. 41–51, Mar. 2009, doi: 10.15845/noril.v1i1.10.

48) N. S. Dhamdhere, “Cumulative citations index, h-index and i10-index (research metrics) of an educational institute: A case study,” *International Journal of Library and Information Science*, vol. *10, no.* 1, pp. 1-9, Jan. 2018, doi: 10.5897/IJLIS2017.0797.