


Please cite the Published Version

Sarwar, Raheem, Zia, Afifa, Nawaz, Raheel , Fayoumi, Ayman, Aljohani, Naif Radi and Hassan, Saeed-UI (2021) Webometrics: evolution of social media presence of universities. *Scientometrics*, 126 (2). pp. 951-967. ISSN 0138-9130

DOI: <https://doi.org/10.1007/s11192-020-03804-y>

Publisher: Springer Science and Business Media LLC

Version: Accepted Version

Downloaded from: <https://e-space.mmu.ac.uk/627086/>

Additional Information: This is an Author Accepted Manuscript of an article published in *Scientometrics*.

Enquiries:

If you have questions about this document, contact openresearch@mmu.ac.uk. Please include the URL of the record in e-space. If you believe that your, or a third party's rights have been compromised through this document please see our Take Down policy (available from <https://www.mmu.ac.uk/library/using-the-library/policies-and-guidelines>)



Webometrics: evolution of social media presence of universities

Raheem Sarwar¹ · Afifa Zia² · Raheel Nawaz³ · Ayman Fayoumi⁴ · Naif Radi Aljohani⁴ · Saeed-UI Hassan²

Received: 23 October 2019

© Akadémiai Kiadó, Budapest, Hungary 2021

Abstract

This paper aims at an important task of computing the *webometrics university ranking* and investigating if there exists a correlation between webometrics university ranking and the rankings provided by the world prominent university rankers such as QS world university ranking, for the time period of 2005–2016. However, the webometrics portal provides the required data for the recent years only, starting from 2012, which is insufficient for such an investigation. The rest of the required data can be obtained from the internet archive. However, the existing data extraction tools are incapable of extracting the required data from internet archive, due to unusual link structure that consists of web archive link, year, date, and target links. We developed an internet archive scrapper and extract the required data, for the time period of 2012–2016. After extracting the data, the webometrics indicators were quantified, and the universities were ranked accordingly. We used correlation coefficient to identify the relationship between webometrics university ranking computed by us and the original webometrics university ranking, using the spearman and pearson correlation measures. Our findings indicate a strong correlation between ours and the webometrics university rankings, which proves that the applied methodology can be used to compute the webometrics university ranking of those years for which the ranking is not available, i.e., from 2005 to 2011. We compute the webometrics ranking of the top 30 universities of North America, Europe and Asia for the time period of 2005–2016. Our findings indicate a positive correlation for North American and European universities, but weak correlation for Asian universities. This can be explained by the fact that Asian universities did not pay much attention to their websites as compared to the North American and European universities. The overall results reveal the fact that North American and European universities are higher in rank as compared to Asian universities. To the best of our knowledge, such an investigation has been executed for the very first time by us and no recorded work resembling this has been done before.

Keywords Webometrics university ranking · University rankers · Web impact indicators · Higher education · Internet archive

✉ Raheem Sarwar
R.Sarwar4@wlv.ac.uk

Extended author information available on the last page of the article

Introduction

Background

There are three world renowned university rankings systems: (1) Quacquarelli Symonds World University Ranking (QS), (2) Times Higher Education World University Ranking (THE), and (3) Shanghai Jiao Tong University also known as Academic Ranking of World Universities (ARWU). These ranking systems use different methodologies to rank the universities. For instance, the QS ranking is based upon university activity, academic peer review, employers review, faculty student ratios, citations per faculty and internationalization. THE ranking takes into account the opinion surveys, faculty per student ratio, citations per faculty, international faculty and international students (Thakur 2007). Meanwhile, ARWU covers wide range of alumni and staff winning Nobel prizes and medals, researchers having high citations, number of articles published in the journals, number of articles indexed in the science citation index (SCI) and social science citation index (SSCI), and per capita academic performance of an institution.

However none of the aforementioned renowned university ranking systems consider the web presence of the universities as a ranking factor. On the other hand several existing studies have shown the usefulness of ranking the universities based on their web presence (Aguillo et al. 2008; Ayu and Elgharabawy 2013; Patel and Parmar 2018; Dastani et al. 2019). A Spanish research group known as Cybermetrics Lab¹ initiated the webometrics university ranking.² Specifically, they measured the web presence of the universities using the web impact indicators (webometric features) such as web size, visibility, rich files and scholar articles and promoted the effectiveness of the web ranking of universities to their academic reputations. The academic websites in countries are the most important Internet communication tools. They introduce universities, their related institutes and departments, their resources and services, faculty members, students, and alumnae. The professors use academic websites to provide teaching materials, raw data, drafts, slides, software, bibliographic or link lists which inform about the commitment of professors to their students. The structure, composition, and all kinds of administrative information provided by the institution itself are very valuable. Nowadays, an important factor for the success of a university is its website and web accessibility and in particular its visibility on the web. When this information is made publicly available through the web, it speaks of the high academic level of the university (Sarwar et al. 2020a, b, c; Sarwar and Nutanong 2016; Nutanong et al. 2016). Therefore, it is important to evaluate their presence on the web as it is to compute their academic reputation/ranking. To validate this phenomenon we then compare it against the world-renowned university rankers (i.e., QS, THE, ARWU). Based on the aforementioned discussion we formulate the following important research question.

- *Research Question* Is it possible to entirely rely on the web of the universities to compute the university ranking?

¹ <http://internetlab.cchs.csic.es/>.

² <http://www.webometrics.info/en>.

The objectives of this investigation and challenges involved

To answer the aforementioned research question, we need to compute the webometrics university ranking of the top 30 universities of North America, Europe and Asia for the time period of 2005–2016, and compare it against the ranking provided by the world prominent university rankers such as QS, THE and ARWU. However, the webometrics portal³ provides the required data for the recent years only, starting from 2012, which is insufficient for such an investigation. The rest of the required data (i.e., from 2005 to 2011), which is not available by webometrics portal can be obtained from the internet archive.⁴ The internet archive is a nonprofit digital library that provides a platform to extract the previous versions of the websites (see Sect. 2.2 for more details). However, the existing available tools are incapable of extracting data from internet archive due to unusual link structure that consists of web archive link, year, date, and target link.

Our contributions

We develop a tool (archive scrapper) that can retrieve the required data from internet archive. Specifically, our archive scrapper retrieves the data from internet archive and extract the information regarding internal links, external links, self-links, word documents, PDF, PPT, audio and video files. Our archive scrapper is capable of handling several types of websites such as business, universities, hotels and online shopping etc., and outputs the results in a database. The peculiarity of crawler is its potency to crawl the links of internet archive which is not possible with existing tools. The data and our archive scrapper will be made publicly available.

We evaluated our archive scrapper on the data for the time period of 2012 to 2016. This is because the ground truth information regarding the webometrics university ranking is available for the recent years only (i.e., starting from 2012). Specifically, after preprocessing the extracted data, the webometrics indicators were quantified, and selected universities were ranked accordingly. We used correlation coefficient to establish the relationship between webometrics university ranking computed by us and the world webometrics university ranking using the Spearman correlation and Pearson correlation. Our findings indicate a strong correlation between ours and the webometrics university ranking which proves that the applied methodology can be used to find the rankings of those years for which the ranking is not available, i.e., from 2005 to 2011.

To summarize, the objectives of this research include calculating the webometric university rankings in previous years, comparing them with webometric university rankings of latest and archive websites, and correlating traditional and latest ranking system as well as studying the effects of web presence on ranking system.

Summary of our contributions

The principle contributions of this work can be summarized as follows.

³ https://web.archive.org/web/20120901000000/http://webometrics.info/en/Previous_editions.

⁴ <https://archive.org/>.

- *Re-computing Webometrics University Rankings* To the best of our knowledge, computing university rankings of top thirty universities of North America, Europe and Asia for the recent years (from 2005 to 2016), and evaluating correlation among the webometrics and our computed webometrics university rankings has been executed for the very first time by us and no recorded work resembling this has been done before.
- *Tool Development* We develop a tool that can extract the required data of universities from internet archive.
- *Comparison between Webometrics and World Prominent Ranking* We compared the webometrics university ranking with world prominent university rankings such as QS, THE and ARWU, of top thirty universities of North America, Europe and Asia for the recent years (from 2005 to 2016), also for the first time.

We organize the rest of the paper as follows. Section 2 reviews the existing studies. Section 3 explains the data collection process, our web crawler, and the used methodology. Section 4 presents the experimental results. Section 5 concludes this work and provides some future work directions.

Literature review

In this section we briefly review the webometrics, web impact indicators, internet archive and the web crawling tools.

A brief review of webometrics: an emerging scientific field

The webometrics is made up of two words i.e. web and metrics, where “web” is known as *world wide web* and *metrics* is the mathematical theory of measurement (Alsmadi and Taylor 2018; Ananiadou et al. 2013; Lorentzen 2014). Webometrics emerged from bibliometrics, scientometrics, infometrics and cybermetrics (Hassan et al. 2020; Almind and Ingwersen 1997; Sarwar et al. 2018a, b, c, 2019 Hassan et al. 2016; Sarwar and Hassan 2015; Sabah et al. 2019; Hassan et al. 2019). Almind and Ingwersen (Almind and Ingwersen 1997) proposed the webometrics for the first time and presented the idea of applying informetric methods to the world wide web. They used a case study of comparing web usage of Denmark with other Nordic countries to describe a method which can be used for webometric analyses. Later on Björneborn and Ingwersen (Björneborn and Ingwersen 2004) identified the webometrics resemblances to informetric and scientometrics methods, comparing web-links with citations, but with the noticeable difference that links can go either way. They argue that the webometric analyses of the nature, structures and content-properties of the websites and web-pages are important for understanding the web and its connections. The link structures of websites also have great importance for webometric analyses. They also acknowledged four main areas for webometrics which are (1) web page content analysis, (2) web link structure analysis, (3) web usage analysis and (4) web technology analysis (Thompson et al. 2013; Nawaz et al. 2010; Waheed et al. 2020; Björneborn and Ingwersen 2004; Shardlow et al. 2018; Jahangir et al. 2017; Thompson et al. 2017; Batista-Navarro et al. 2013; Hassan and Haddawy 2015).

The world wide web is a space which has become the part and parcel of life. The web allows documents to be connected through internet by hyperlinks so it has now become one of the significant informants on educational and research activities and is an exceptional

platform for evaluating the webometric activities. In webometric studies, the main focus is on the web impact of the academic institutes. Furthermore, these institutions have well developed websites. It is possible to build their web indicators which explain the academic reputation and activities. Universities rankings are one of the most visible consequences of the Webometrics based upon the websites and their online impact (Das et al. 2019; Stuart et al. 2017; Hande 2019; Aguillo 2018). Spanish Research group known as Cybermetrics Lab⁵ initiated the Webometrics university ranking. Specifically, they measured the web presence of institutions using a set of web indicators and reported that the web ranking of universities is a useful measure for university evaluations.

A brief review of internet archive

The internet archive is a Way Back Machine, the largest website which keeps the history of the evolving web. The internet archive currently contains more than 240 billion web pages with archives as far back as 1996, allowing the users to travel back in time to search archived versions of web pages through the Way back Machine and collects the portions of the web and saves them year by year (Chavez-Demoulin et al. 2000). The internet archive is an open access website where anyone can get data free of cost (Tofel 2007). The main objective of internet archive is to store a huge collection of digital information, save each page without checking the worth of the page, and keep the history of the web record. The internet archive is considered as one of the biggest web archive (Kenney et al. 2002).

Because of the increasingly use of web resources in teaching and academic research (Hickey et al. 2020; Galikyan and Admiraal 2019; Brown et al. 2019; Molinillo et al. 2018; Hassan et al. 2012, 2017; Waheed et al. 2018; Bonaccorsi et al. 2017b, a) internet archive is mostly used for scholarly communication. The particular distinguishing characteristic for this mission is that it retains all retrieved copies of web pages which are indexed by their URL so that variations in a page over time can be chased, and old pages that have been removed from the web can still be found. The resource has been found beneficial for several educational research projects. Search interface of internet archive provides easy access to the historical data. It has a web based search interface to access archived pages. Users submit their query by entering an URL in the search field and the archive returns a table of information having details of all the copies of pages along with their archive month and date (Koman 2002).

A brief review of the web crawling tools

The web crawlers defined as computer programs which retrieve data from the websites (Jalal et al. 2015). LexiURL and SocSciBot are the most powerful web crawlers, used to extract data from websites. The LexiUrl crawler automatically creates the required list of queries for a certain search engine from a simple list of domain names, and the SocSciBot crawler performs link analysis research for strengthening the webometric research. These web crawlers crawl the web pages according to the query, download them in a local machine and then tries to analyze them using integrated analytical software. But, specialized web crawlers are either having limited access or having limited features which were

⁵ <http://internetlab.cchs.csic.es/>.

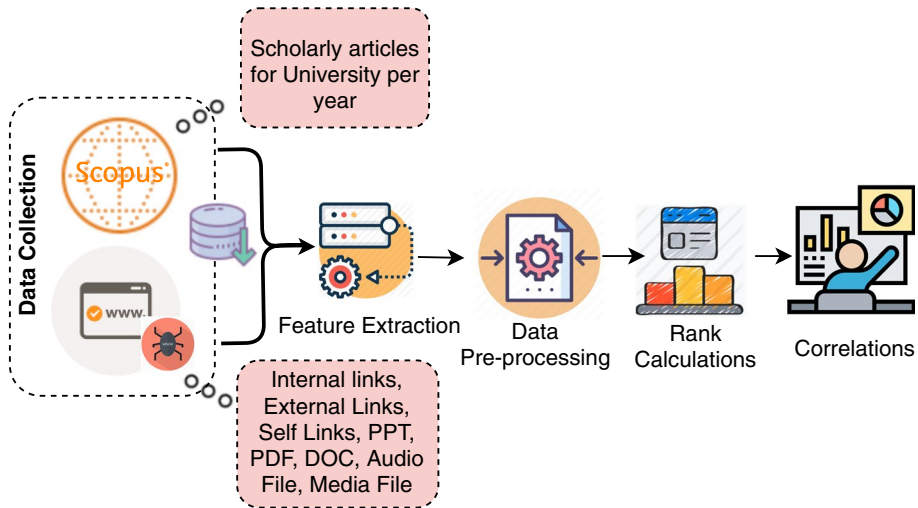


Fig. 1 System overview: our system consists of three: (1) data collection and feature extraction, (2) rank calculations, and (3) correlation calculations

developed for webometric research. In addition we consult these crawlers might be useful for other tasks, but in our case they are of no help.

Methodology

Overview

In this section, we discuss the methodologies to compute the webometrics university ranking and find the relation between the webometrics university ranking and world prominent university rankings. Figure 1 shows the complete scheme of the system. Our system consists of the following three processes:

1. *Data Collection and Features Extraction* The data collection and features extraction processes are responsible to retrieve the required data from the internet archive and extract the features (web indicators) from the retrieved data, respectively.
2. *Rank Calculations* The rank calculation process is responsible to compute the webometrics university ranking using the retrieved data by applying the webometrics university ranking methodology (Aguillo et al. 2008).
3. *Correlation Calculations* The correlation calculations process of our solution is responsible to compute the correlation between (1) the webometrics university ranking computed by us and the original webometrics university ranking for evaluation of our tool, and (2) the webometrics university ranking and the university rankings provided by the world prominent university rankers such as QS, THE and ARWUR.

Let us now explain each process in detail.

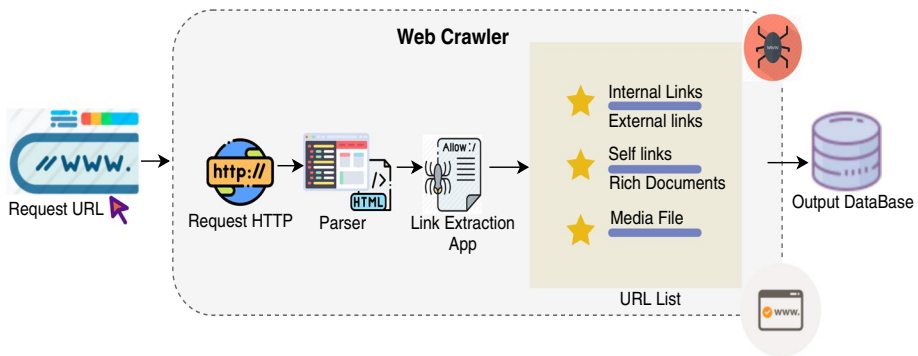


Fig. 2 Web crawler work flow: data extraction from internet archive using our archive scrapper

Data collection and features extraction

The data collection and features extraction processes are responsible to retrieve the required data from the internet archive and extract the features (web indicators) from the retrieved data, respectively. Since existing tools are incapable of retrieving the data from the internet archive and extracting the features from the retrieved data, we developed a tool (archive scrapper) to perform these processes. Our archive scrapper is explained latter in this section. We compute 8 features (web indicators) from the retrieved data. These features are explained in Table 1. We note that our features are language-independent.

Archive Scrapper. This is a web-based tool developed in PHP (Hypertext preprocessing) for collecting the required data from the internet archive. Specifically, the archive scrapper retrieve the previous versions of the websites and extracts feature (such as internal links, external links, self-links, word documents, PDF, PPT, audio and video files) from websites and store them into the database. Figure 2 shows the working of web crawler in detail. The extraction of the required data with the help of a crawler requires copying the link of the website, pasting it in the internet archive search bar, and submitting it. Internet archive then shows the web pages harboring the data of the target website year by year. As one opens a link, copies and pastes it into web crawler search bar, the crawler along with its parser app begins to function. The pseudocode of parser app is as follows:

- *Step 1* Start;
- *Step 2* Read & clean URL (Uniform Resource Locator);
- *Step 3* Parser app creates objects;
- *Step 4* Load site;
- *Step 5* Create array of specific Tags;
- *Step 6* Create array of all the links of website;
- *Step 7* Return data as array;
- *Step 8* Output the results; and
- *Step 9* End.

To extract data from the target website, DOM (Document Object Model) parser Library is used. Our tool is capable of working with HTML (Hypertext markup language) and XML

(Extensible Markup Language), proceeds in a tree structure, and loads data on DOM objects before accessing it. It creates objects by loading a website with the help of curl library of PHP.

Scopus Database. As for the collection of scholarly data, university affiliations were recorded using Scopus, the largest database of literature consisting of scientific journals, books and conference proceedings.

Webometrics ranking calculations

Once we complete the data collection and features extraction processes, normalization has been applied to data because each feature has numeric values of a different range. We normalize all features, ranged from 0 to 1, by centering on the mean and scaling to unit variance. Specifically, we use the following equation for data normalization where x denotes the normalized value and x' denotes the original value.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

After the data normalization we compute the webometrics ranking (Aguillo et al. 2008) using the following model.

$$\text{Ranking} = 2 * \text{Web Size} + 4 * \text{Visibility} + 1 * \text{Research Articles} + 1 * \text{Rich Files},$$

where

$$\text{Web Size} = (\text{IL} + \text{SL} + \text{Media}) * 0.2$$

$$\text{Research Articles} = \text{RA} * 0.15$$

$$\text{Rich Files} = (\text{PPT} + \text{DOC} + \text{PDF}) * 0.15$$

$$\text{External Links} = \text{EL} * 0.5.$$

The aforementioned model accumulates four key aspects to be measured in the academic web (also given in Table 1):

1. *web size*, that is the volume of published information, i.e., number of web pages (IL and SL),
2. *visibility*, number of links to external websites (EL),
3. *scholar data*, amount of scholarly documents index by Scopus (RA), and
4. *rich files*, including PDFs, PPTs, Docs, and media files.

The used model categorizes the aforementioned four key aspects into activity (web size, scholar data, rich files) and impact (external links). A weight was assigned to better reflect the contribution of each aspect to the used model. The weight ratio proposed by the used model is 1:1, between activity and impact. That is, activity consists of 50% of the weight which includes total number of web pages, rich files from the website and research articles from Google scholar and visibility accounts for the other 50% which consists of external links received by repository.

Table 1 Webometrics features: all features are grouped into four categories such as web size, visibility, scholarly articles and rich files

Feature category	Features (web indicators)	Definitions
Web size	Internal links (IL)	Total number of internal links (IL): IL is a type of hyperlink on a web page that refers to any web page of the same website or domain
	Self links (SL)	Total Number of Self links (SL): SL is a type of hyperlink on a web document that refers to the same web document of a website
Visibility	External links (EL)	Total number of External Links (EL): EL is a type of hyperlink that refers to a page outside the same website or domain
Scholarly articles Rich files	Research articles (RA)	Number of scholarly documents indexed by Scopus
	PDF	Number of files in PDF format
	PPT	Number of files in PPT format
	Doc files (DOC)	Number of files in DOC format
	Media files	Number of video and audio files

Table 2 Correlation results of North American, European and Asian Universities

Method	2016	2015	2014	2013	2012
<i>North American universities correlation results</i>					
Spearman's (ρ)	0.97	0.92	0.98	0.98	0.96
Pearson (r)	0.96	0.92	0.98	0.98	0.96
<i>European universities correlation results</i>					
Spearman's (ρ)	0.93	0.93	0.95	0.96	0.96
Pearson (r)	0.90	0.92	0.94	0.95	0.95
<i>Asia universities correlation results</i>					
Spearman's (ρ)	0.91	0.92	0.91	0.88	0.86
Pearson (r)	0.70	0.86	0.86	0.88	0.81
<i>Top universities correlation results of three regions</i>					
Spearman's (ρ)	0.93	0.90	0.94	0.93	0.93
Pearson (r)	0.76	0.84	0.90	0.91	0.86

Correlation calculations between rankings

To validate the accuracy of obtained results, rank correlation methods such as Spearman (ρ) and Pearson (r) were used. The Spearman correlation (ρ) measures the correlation between two variables. The correlation between two variables is high when they have similar ranks and low with dissimilar values. The Pearson correlation (r) measures linear dependence between two variables. Its value is between $+1$ and -1 , where 1 denotes positive correlation, 0 denotes that there is no correlation, and -1 denotes the negative correlation. In next section we discuss our results in detail and also analyze the university rankings.

Experimental results

This section reports the findings obtained from our experimental studies. Based on the objectives of this investigation listed in Introduction, we performed the following two main studies. In the first study, we measure the correlation between original webometrics university ranking and the webometrics university ranking calculated by us for the time period of 2012–2016. In the second study, we measure the correlations between webometrics university ranking and the world prominent university rankers, i.e., QS, THE and ARWU. Third study investigates the correlation between webometrics and world prominent university rankers of top ten universities for the time period of 2005–2016. The experimental results from these studies are reported in the following subsections.

Correlation between webometrics and our calculated university rankings for the time period of 2012–2016

In this section we discuss the findings of correlation between world webometrics university ranking and webometrics university ranking computed by us using the data obtained from internet archive for the time period of 2012–2016. Recall that the webometrics website

Table 3 Spearman correlation of North America, European and Asian universities

Year	QS	THE	ARWU
<i>Spearman correlation of North America universities</i>			
2016	0.76	0.79	0.78
2015	0.50	0.56	0.72
2014	0.55	0.68	0.57
2013	0.54	0.68	0.63
2012	0.37	0.40	0.42
<i>Spearman correlation of European universities</i>			
2016	0.93	0.31	0.46
2015	0.40	0.41	0.45
2014	0.44	0.37	0.57
2013	0.45	0.52	0.57
2012	0.44	0.36	0.50
<i>Spearman correlation of Asian universities</i>			
2016	0.56	0.63	0.62
2015	0.57	0.23	0.50
2014	0.43	0.19	0.59
2013	0.41	0.18	0.35
2012	0.46	0.20	0.46

provides the latest webometrics university ranking, for the years 2012–2016. After collecting the top thirty universities' webometrics ranking for years 2012–2016 from webometrics website, we correlate them with our calculated rankings.

The correlations [i.e., Spearman (ρ) and Pearson (r)] of our computed webometrics rankings and webometrics rankings of regions namely North America, Europe and Asia are given in the Table 2 for years 2012–2016. Table 2 depicts the strong positive correlation values for years 2012–2016 of North American universities, European universities and Asian universities. We extended our work to calculate correlation for all universities as a single group instead of region-based grouping and the results once again presented a strong correlation (see Table 2). The obtained results clearly show that there is a strong correlation between the webometrics university ranking and the webometrics university ranking computed by us for the time period of 2012–2016 which justifies our methodology and the effectiveness of our archive scraper.

Correlation between webometrics and world prominent rankings for the time period of 2012–2016

In this section we provide the experimental results regarding the correlation of webometrics university ranking with world prominent university rankers (i.e., QS, THE, and ARWU). The experimental results of North American Universities, European Universities and Asian universities are given in Table 3. The results for each region are explained below.

North American Universities. As for the North American University rankings, we can see that the Spearman correlation (ρ) values range between 0.3 and 0.8. Moreover, we can see the positive correlation between webometrics university ranking computed by us and the world prominent university rankings for all the years, but weak correlation

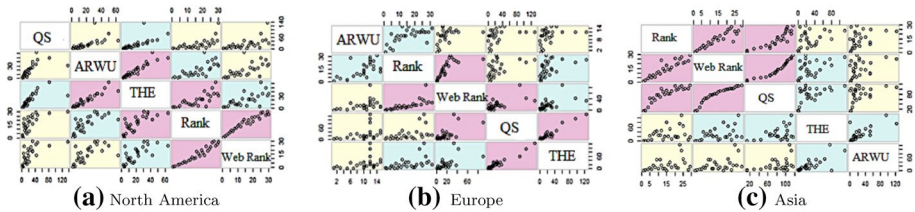


Fig. 3 Scatter plots of North America, Europe and Asia (rank represents our computed webometric rankings and Web Rank depicts webometric university rankings)

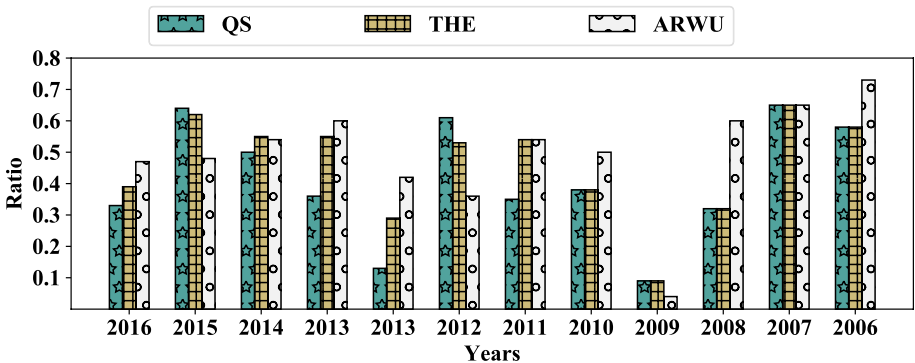


Fig. 4 Correlation between webometrics and world prominent ranking of North America

for the year 2012, that is below 0.5. We also note that there is a significant correlation value difference between the year 2016 and rest of the years.

European Universities. For European universities, the correlation between webometrics rankings and world-famous rankings is positive ranging from 0.3 to 0.9. A notable observation is that the correlation between webometrics ranking and QS World University Ranking is 0.93 for year 2016 but for the other years it is near 0.4. THE and QS rankings correlation with webometrics rankings is near 0.5 which reflects the lack of similarity in their case.

Asian Universities. Correlation between webometrics university rankings with world prominent university rankings for top 30 universities of Asia is positive but there is not much closeness between them because correlation coefficient is approximately 0.5. The correlation between ARWU, QS and Webometrics rankings is 0.4–0.6 for all the years but THE is close to 0 for years 2012–2015 and 0.6 for year 2016.

Figure 3a–c show the scatter plot matrix between our computed webometric rankings versus world prominent rankings for year 2016 of regions namely North America, Europe and Asia. In Fig. 3a–c, the rank represents our computed webometric rankings and web rank depicts webometric university rankings. Variables with highest correlations are closest to the principal diagonal and color of cells shows the size of correlations. These figures depict positive correlation between our computed rankings and webometric rankings and weak positive correlation with world prominent university rankings. The overall results declared the fact that North American and European

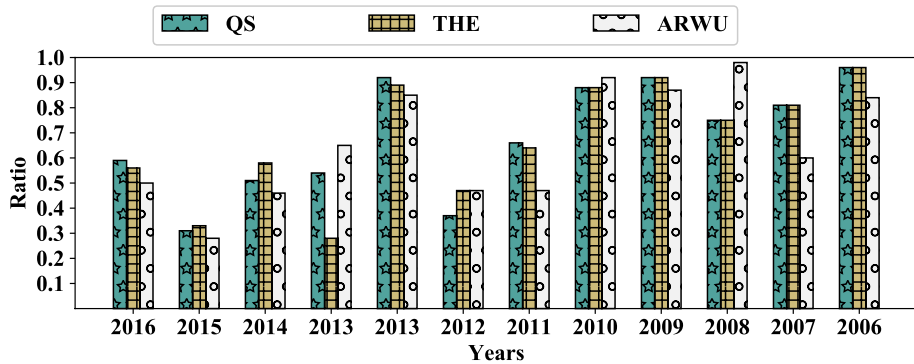


Fig. 5 Correlation between webometrics and world prominent ranking of Europe

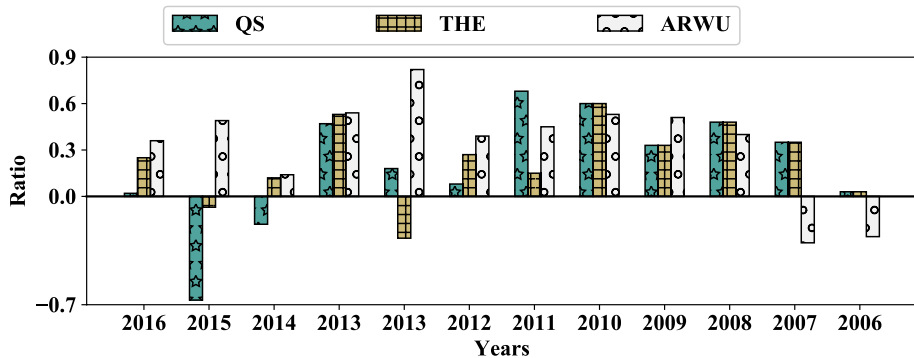


Fig. 6 Correlation between webometrics and world prominent ranking of Asia

universities are higher in rank as compared to Asian universities which implies that the university websites have immense effect on the university rankings.

Correlation between webometrics and world prominent ranking of top ten universities for the time period of 2005–2016

This study was conducted on top ten universities selected according to obtained results. The university ranking data of the three prominent ranking organizations such as ARWU, QS and THE rankings was collected from their respective websites for required years. After data collection, the correlation was calculated between our computed webometrics rankings and world prominent rankings for Asian, European and North American universities for years 2005–2016.

Figures 4, 5, and 6 show the correlation, X-axis shows the years, Y-axis represents correlation and bar colors signifies the world prominent rankings (QS, THE, ARWU). Figure 4 shows correlation between our computed webometrics rankings versus world prominent rankings of North America. The North American universities' results show positive correlation roughly ranging from 0.3 to 0.5 for QS and THE rankings and 0.4–0.7 for ARWU. In Fig. 5, correlation between our computed webometrics rankings

versus world prominent rankings of European countries is shown. The correlation results for European universities hint at much closeness as they range from 0.59 to 0.96. Figure 6 shows the correlation between our computed webometrics rankings and world prominent rankings of Asia for the top ten years. Asian universities show negative correlation with THE, QS, and ARWU for some years and weak positive correlation for other years. It depicts that they did not pay much attention to their websites.

Unlike the existing study (Aguillo et al. 2008) we show that if there is a strong correlation between an adequate web presence and quality of the institution, the contrary is not true and there are prestigious universities underperforming in the webometrics arena due to insufficient motivations regarding their web policy.

Conclusions

This paper aims at answering the following important research question: is it possible to entirely rely on the web of the universities to compute the university rankings? To answer this questions, we compute the webometrics university ranking of the top 30 universities of North America, Europe and Asia for the time period of 2005–2016 and compare it against the ranking provided by the world prominent university rankers such as QS, THE and ARWU.

Our findings obtained from the analysis performed on the dataset from 2005 to 2016 indicate a positive correlation for North American and European universities, but weak correlation for Asian universities. This can be explained by the fact that top Asian universities (according to the world prominent university rankers) did not pay much attention to their websites as compared to North American and European universities. It implies that in order to rank the universities we cannot entirely rely on the web of the universities. Asian universities webmasters should pay more attention to the universities web design and content to make them more attractive and usable not only for their own students and staff, but for all Asian and non-Asian users of the Internet. If there is a strong correlation between an adequate web presence and quality of the institution, the contrary is not true and there are prestigious universities underperforming in the webometrics arena due to insufficient motivations regarding their web policy.

There are specific challenges in computing the webometrics ranking such as the changes of domains which affects the visibility of the institution's web presence. In addition to this, as mentioned earlier if there is a strong correlation between an adequate web presence and quality of the institution, the contrary is not true and there are prestigious universities underperforming in the webometrics arena due to insufficient motivations regarding their web policy.

For the future works, instead of calculating the results manually using MS Excel, our tool can be enhanced to automatically preprocess the data, do further calculations, calculate the universities rankings, correlate the webometric ranking with other ranking organizations around the world and plot the rankings and analysis on graphs for a bigger picture. Moreover the text mining techniques can be applied to analyze and compare results to possibly improve the evaluation process.

References

- Aguillo, I. F. (2018). Altmetrics of the open access institutional repositories: A webometrics approach. In *23rd international conference on science and technology indicators (STI 2018)*, September 12–14, 2018, Leiden, The Netherlands, Centre for Science and Technology Studies (CWTS), (2018). Centre for Science and Technology Studies (CWTS).
- Aguillo, I. F., Ortega, J. L., & Fernández, M. (2008). Webometric ranking of world universities: Introduction, methodology, and future developments. *Higher Education in Europe*, 33(2–3), 233–244.
- Almind, T. C., & Ingwersen, P. (1997). Informetric analyses on the world wide web: Methodological approaches to 'webometrics'. *Journal of Documentation*, 53(4), 404–426.
- Alsmadi, I., & Taylor, Z. (2018). Examining university ranking metrics: Articulating issues of size and web dependency. In *Proceedings of the 2018 international conference on computing and big data* (pp. 73–77). ACM.
- Ananiadou, S., Thompson, P., & Nawaz, R. (2013). Enhancing search: Events and their discourse context. In *International conference on intelligent text processing and computational linguistics* (pp. 318–334). Springer.
- Ayu, M. A., & Elgharabawy, M. A. (2013). Effects of web accessibility on search engines and webometrics ranking. *IJCMC*, 5(1), 69–94.
- Batista-Navarro, R.T., Kontonatsios, G., Mihăilă, C., Thompson, P., Rak, R., Nawaz, R., Korkontzelos, I., & Ananiadou, S. (2013). Facilitating the analysis of discourse phenomena in an interoperable NLP platform. In *International conference on intelligent text processing and computational linguistics* (pp. 559–571). Springer.
- Björneborn, L., & Ingwersen, P. (2004). Toward a basic framework for webometrics. *Journal of the American society for Information Science and Technology*, 55(14), 1216–1227.
- Bonaccorsi, A., Cicero, T., Haddawy, P., & Hassan, S.-U. (2017a). Explaining the transatlantic gap in research excellence. *Scientometrics*, 110(1), 217–241.
- Bonaccorsi, A., Haddawy, P., Cicero, T., & Hassan, S.-U. (2017b). The solitude of stars: An analysis of the distributed excellence model of european universities. *Journal of Informetrics*, 11(2), 435–454.
- Brown, M. G., Schiltz, J., Derry, H., & Holman, C. (2019). Implementing online personalized social comparison nudges in a web-enabled coaching system. *The Internet and Higher Education*, 43, 100691.
- Chavez-Demoulin, V.C., Roehrl, A.S., Roehrl, R.A., & Weinberg, A. (2000). The WEB archives: A time-machine in your pocket! In *Proceedings of The Internet archive colloquium 2000*.
- Das, S. S., Balasubramanian, P., & Chowdhury, A. R. (2019). Webometrics ranking (WR) of world universities and national institutional ranking framework (NIRF): A comparative study. *SRELS Journal of Information Management*, 56(3), 154–158.
- Dastani, M., Panahi, S., Sattari, M., et al. (2019). Webometrics analysis of Iranian universities about medical sciences' websites between september 2016 and March 2017. *Acta Informatica Malaysia (AIM)*, 3(1), 7–12.
- Galikyan, I., & Admiraal, W. (2019). Students' engagement in asynchronous online discussion: The relationship between cognitive presence, learner prominence, and academic performance. *The Internet and Higher Education*, 43, 100692.
- Hande, N. H. (2019). Websites of IITs, IIMs and NITs: A webometrics study. *Journal of Advancements in Library Sciences*, 6(1), 351–357.
- Hassan, S.-U., Aljohani, N. R., Idrees, N., Sarwar, R., Nawaz, R., Martínez-Cámara, E., et al. (2019). Predicting literature's early impact with sentiment analysis in Twitter. *Knowledge-Based Systems*, 192, 105383.
- Hassan, S.-U., Akram, A., & Haddawy, P. (2017). Identifying important citations using contextual information from full text. In *2017 ACM/IEEE joint conference on digital libraries (JCDL)* (pp. 1–8). IEEE.
- Hassan, S.-U., & Haddawy, P. (2015). Analyzing knowledge flows of scientific literature through semantic links: A case study in the field of energy. *Scientometrics*, 103(1), 33–46.
- Hassan, S.-U., Haddawy, P., Kuinkel, P., Degelsegger, A., & Blasy, C. (2012). A bibliometric study of research activity in asean related to the eu in fp7 priority areas. *Scientometrics*, 91(3), 1035–1051.
- Hassan, S.-U., Sarwar, R., & Muazzam, A. (2016). Tapping into intra-and international collaborations of the organization of islamic cooperation states across science and technology disciplines. *Science and Public Policy*, 43(5), 690–701.
- Hassan, S. U., Aljohani, N. R., Shabbir, M., Ali, U., Iqbal, S., Sarwar, R., Martínez-Cámara, E., Ventura, S., & Herrera, F. (2020). Tweet Coupling: a social media methodology for clusteringscientific publications. *Scientometrics*, 1–19. <https://doi.org/10.1007/s11192-020-03804-y>.
- Hickey, D. T., Robinson, J., Fiorini, S., & Feng, Y. (2020). Internet-based alternatives for equitable preparation, access, and success in gateway courses. *The Internet and Higher Education*, 44, 100693.

- Jahangir, M., Afzal, H., Ahmed, M., Khurshid, K., & Nawaz, R. (2017). An expert system for diabetes prediction using auto tuned multi-layer perceptron. In *2017 Intelligent systems conference (IntelliSys)* (pp. 722–728). IEEE.
- Jalal, S. K., Sutradhar, B., Sahu, K., Mukhopadhyay, P., & Biswas, S. C. (2015). Search engines and alternative data sources in webometric research: An exploratory study. *DESIDOC Journal of Library & Information Technology*, 35(6), 427–435.
- Kenney, A. R., McGovern, N. Y., Botticelli, P., Entlich, R., Lagoze, C., & Payette, S. (2002). Preservation risk management for web resources. *Information Management Journal-Prairie Village*, 36(5), 52–61.
- Koman, R. (2002). How the wayback machine works. *XML.com Jan*, 21, 6.
- Lorentzen, D. G. (2014). Webometrics benefitting from web mining? an investigation of methods and applications of two research fields. *Scientometrics*, 99(2), 409–445.
- Molinillo, S., Anaya-Sánchez, R., Aguilar-Illescas, R., & Vallespín-Arán, M. (2018). Social media-based collaborative learning: Exploring antecedents of attitude. *Internet and Higher Education*, 38(1), 18–27.
- Nawaz, R., Thompson, P., McNaught, J., & Ananiadou, S. (2010). Meta-Knowledge annotation of bio-events. In *LREC* (Vol. 17, pp. 2498–2507).
- Nutanong, S., Yu, C., Sarwar, R., Xu, P., & Chow, D. (2016, December). A scalable framework for stylistic analysis query processing. In *2016 IEEE 16th International Conference on Data Mining (ICDM)* (pp. 1125–1130). IEEE.
- Patel, H. J., & Parmar, S. D. (2018). Webometrics study of all india institutes of medical sciences. *Journal of Advancements in Library Sciences*, 2(2), 12–17.
- Sabah, F., Hassan, S.-U., Muazzam, A., Iqbal, S., Soroya, S. H., & Sarwar, R. (2019). Scientific collaboration networks in pakistan and their impact on institutional research performance: A case study based on scopus publications. *Library Hi Tech*, 37(1), 19–29.
- Sarwar, R., & Hassan, S.-U. (2015). A bibliometric assessment of scientific productivity and international collaboration of the Islamic world in science and technology (s&t) areas. *Scientometrics*, 105(2), 1059–1077.
- Sarwar, R., Li, Q., Rakthanmanon, T., & Nutanong, S. (2018a). A scalable framework for cross-lingual authorship identification. *Information Sciences*, 465, 323–339.
- Sarwar, R., Porthavepong, T., Rutherford, A., Rakthanmanon, T., & Nutanong, S. (2020a). StyloThai: A scalable framework for stylometric authorship identification of thai documents. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(3), 1–15.
- Sarwar, R., Rutherford, A. T., Hassan, S. U., Rakthanmanon, T., & Nutanong, S. (2020b). Native Language Identification of Fluent and Advanced Non-Native Writers. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 19(4), 1–19.
- Sarwar, R., Soroya, S.H., Muazzam, A., Sabah, F., Iqbal, S., & Hassan, S.-U. (2019). A bibliometric perspective on technology-driven innovation in the Gulf Cooperation Council (GCC) countries in relation to its transformative impact on international business. In *Technology-driven innovation in Gulf Cooperation Council (GCC) countries: Emerging research and opportunities* (pp. 49–66). IGI Global.
- Sarwar, R., Uraileertprasert, N., Vannaboot, N., Yu, C., Rakthanmanon, T., Chuangsuwanich, E., & Nutanong, S. (2020c). CAG: Stylometric authorship attribution of multi-author documents using a co-authorship graph. *IEEE Access*, 8, 18374–18393.
- Sarwar, R., Yu, C., Nutanong, S., Uraileertprasert, N., Vannaboot, N., & Rakthanmanon, T. (2018c). A scalable framework for stylometric analysis of multi-author documents. In *International Conference on Database Systems for Advanced Applications* (pp. 813–829). Cham: Springer.
- Sarwar, R., Yu, C., Tungare, N., Chitavisutthivong, K., Sriratanawilai, S., Xu, Y., & Nutanong, S. (2018b). An effective and scalable framework for authorship attribution query processing. *IEEE Access*, 6, 50030–50048.
- Sarwar, R., & Nutanong, S. (2016). The key factors and their influence in authorship attribution. *Research in Computer Science*, 110, 139–150.
- Shardlow, M., Batista-Navarro, R., Thompson, P., Nawaz, R., McNaught, J., & Ananiadou, S. (2018). Identification of research hypotheses and new knowledge from scientific literature. *BMC Medical Informatics and Decision Making*, 18(1), 46.
- Stuart, E., Stuart, D., & Thelwall, M. (2017). An investigation of the online presence of UK universities on instagram. *Online Information Review*, 41(5), 582–597.
- Thakur, M. (2007). The impact of ranking systems on higher education and its stakeholders. *Journal of Institutional Research*, 13(1), 83–96.
- Thompson, P., Nawaz, R., Korkontzelos, I., Black, W., McNaught, J., & Ananiadou, S. (2013, October). News search using discourse analytics. In *2013 Digital Heritage International Congress (Digital Heritage)* (Vol. 1, pp. 597–604). IEEE.

- Thompson, P., Nawaz, R., McNaught, J., & Ananiadou, S. (2017). Enriching news events with meta-knowledge information. *Language Resources and Evaluation*, 51(2), 409–438.
- Tofel, B. (2007). Wayback' for accessing web archives. In *Proceedings of the 7th international web archiving workshop* (pp. 27–37).
- Waheed, H., Hassan, S.-U., Aljohani, N. R., & Wasif, M. (2018). A bibliometric perspective of learning analytics research landscape. *Behaviour & Information Technology*, 37(10–11), 941–957.
- Waheed, H., Hassan, S. U., Aljohani, N. R., Hardman, J., Alelyani, S., & Nawaz, R. (2020). Predicting academic performance of students from VLE big data using deep learning models. *Computers in Human Behavior*, 104, 106189.

Affiliations

Raheem Sarwar¹  · Afifa Zia² · Raheel Nawaz³ · Ayman Fayoumi⁴ · Naif Radi Aljohani⁴ · Saeed-UI Hassan²

Afifa Zia
mcs14019@itu.edu.pk

Raheel Nawaz
r.nawaz@mmu.ac.uk

Ayman Fayoumi
afayoumi@kau.edu.sa

Naif Radi Aljohani
nr.aljohani@kau.edu.sa

Saeed-UI Hassan
saeed-ul-hassan@itu.edu.pk

¹ Research Group in Computational Linguistics, University of Wolverhampton, Wolverhampton, UK

² Information Technology University, 346-B, Ferozepur Road, Lahore, Pakistan

³ Department of Operations, Technology, Events and Hospitality Management, Manchester Metropolitan University, Manchester, UK

⁴ Faculty of Computing and Information Technology, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia