ISSN:0975-9646

Empirical Validation of Webometrics based Ranking of World Universities

R. K. Pandey
University Institute of Computer Science and Applications (UICSA)
R. D. University,Jabalpur

Abstract— **Webometrics is an emerging discipline out of the growth of WWW and publication of the scientific research using the WWW as a vehicle for disseminating, propagating and publishing by the individuals and organizations. Webometrics data has been used to rank the world universities on the web serving as indicators of their academic performance. This paper makes an attempt to compare the webometrics based rankings of World Universities with the rankings done using the conventional parameters (non-webometrics) like quality of education, quality of faculty, research output etc. which have been used over the years for the purpose before the web came into existence. The purpose of this study is to understand the correlation between the two rankings done using the different approaches. This paper also discusses the vulnerability and manipulability aspects of webometrics based rankings of the entities like universities or hospitals or something else.**

Keywords- backlinks, bibliometrics, search engine, Search Engine Optimization (SEO),

I. INTRODUCTION

The advent and popularity of the World Wide Web (WWW) has given birth to a new discipline webometrics. Webometrics, is the quantitative study of Web-related phenomena, emerged from the realization that methods originally designed for bibliometrics analysis of scientific journal article citation patterns could be applied to the web, with commercial search engines providing the raw data [3]. At the heart of webometrics studies is the information provided by the large-scale search engines, such as Yahoo ! (more suitable) or Google, about the structure of the web like total number of pages in a web site and the total number of back-links to the web site etc. This information and other attributes of this information which have been termed as webometrics can serve as indicators to predict the status and performance attributes of these entities which are responsible for generating this information on the web. This webometrics data has been used to rank the World Universities to assess their Web based performance which in turn can be interpreted as an indicator of their academic performance as well. The Universities and institutions of higher learning have been traditionally ranked using the parameters like the quality of faculty, the number of patents registered, the number of research publications, quality of publications etc. even before the web came into existence. The overall purpose of this paper is to study the closeness or dispersion in the rankings assigned to universities using these two different methods. As web became more and more popular amongst the commercial organizations to promote their commercial activities, so came into existence the Search Engine Marketing [35] companies and Search engine optimization (SEO) techniques [36,37] whose only job was to devise mechanisms to boost the rank of a web page. Unethical SEO practices called black hat tricks like link farming, spamdexing etc. [38,39] with over all a single task in the agenda how to manipulate the search engine ranking to rank first with their competitors on the web. Another purpose of this paper is to analyze the effect of this unethical SEO practice [40] on the ranking of Universities

II. DEFINITIONS

"the study of the quantitative aspects of the construction and use of information resources, structures and technologies on the Web drawing on bibiliometric and informetric approaches." The term *webometrics* was first coined by [1]. Another definition of webometrics has also been introduced by [2], which is "the study of web-based content with primarily quantitative methods for social science research goals using techniques that are not specific to one field of study", which emphasises a small subset of relatively applied methods for use in the wider social sciences

III. WEBOMETRICS VS BIBLIOMETRICS

Historically the development of quantitative analysis of academic publishing (bibliometrics) was the creation of the Institute for Scientific Information (ISI, now Thomson Reuter) citation database, which started operating since 1962 [4,5] was a major step. Another development for bibliometrics was the web publishing of research related documents, from articles to e-mail discussion lists, allowing the creation of a range of new metrics relating to their access and use [6].

A Bibliometrics

Bibliometrics refers to the measurement of “properties of documents, and document-related processes” [7]. Bibliometric techniques include word frequency analysis [8], citation analysis [9], co-word analysis [10] and simple document counting, such as the number of publication by an author or research-group. In practice however, bibliometrics has been primarily applied to science documents and hence has considerable overlap with scientometrics, the science measurement field [6]. The emergence of bibliometrics as a scientific field was triggered by the development of Institute for Scientific Information (ISI) Science Citation Index (SCI) by Eugene Garfield [5]. The SCI was created as a database of the references made by the authors in their articles to the articles published earlier in the top scientific journals [6]. Since then ISI’s SCI served as the main instrument to assess the impact of scholarly work to evaluate or compare the relative scientific contributions of two or more individuals or groups.

B Webometrics

Webometrics is the quantitative analysis of web phenomena, drawing upon informetric methods [11] and typically addressing problems related to bibliometrics. Webometrics was triggered by the realization that the web is an enormous document repository with many of these documents being academic-related [12]. Moreover, the web has its own citation index in the form of commercial search engines. Some of the search engines are

automated thereby enabling the researchers to carry out large-scale investigations [13]. Ranking of world universities [14] which is also the focus of this paper, based upon their web sites and online impact is an excellent example of webometrics application.

Webometrics includes link analysis, web citation analysis, search engine evaluation and purely descriptive studies of the web [6].

- *Link Analysis*

Link analysis is the quantitative study of hyperlinks between web pages [6]. The use of links in bibliometrics was triggered by Web Impact Factor (WIF) [15] analogous to Journal Impact Factor (JIF), on the assumption that hyperlinks might be usable by bibliometrician in a similar way as citations [16]. The standard WIF measures the average number of links per page to a web space from external pages [15]. The idea underlying link analysis was that the number of links targeting an academic web site might be proportional to the research productivity of the organization at the level of university [17], departments [18], research groups [19], or individual scientist [20].

- *Web Citation Analysis*

As scientific publication moves to the web, and novel approaches to scholarly communication and peer review establish themselves, new methods of citation and link analysis have emerged to capture often liminal expressions of peer esteem, influence and approbation. The web thus affords bibliometricians rich opportunities to apply and adapt their techniques to new contexts and content [28].

The hypertextual character of the web means that the principles of citation indexing can be applied much more widely than at present. On the web, scholars do more than publish, or post, their working papers and finished articles: they 'seed ideas, discuss issues and debate positions, in ways which, occasionally deviate from, and challenge, established norms' [21]. Furthermore, they recommend their own work, and the work of selected others, to their peers. A number of studies have revealed that the results of web-based citation counting correlates significantly with ISI citation count across a range of disciplines, with web citations being typically more numerous [22, 23-25].

- *Search Engines*

Search engines have been the main portal to the web for most users since their inception. Search engines are at the heart of webometrics studies. Two main topics of webometrics research have been the extent of coverage of the web and accuracy of the reported results. Studies of the main search engines have revealed that none covered more than 17.5 % of the indexable web and that the overlap between search engines was surprisingly low [26]. The issue of accuracy of search engines results is multifaceted, relating to the extent to which a search engine correctly reports its own knowledge of the web. Studies [27] have shown that search engines are not internally consistent in the way they report results to users. In the background of the above knowledge we carry out our study on the ranking of world universities in the following manner.

IV. METHODOLOGY

Comparisons made in this paper between the ranking orders of universities may be just a representative as it is not possible to make an exhaustive comparison between the universities because of number of reasons and difficulties. The list of the ranked universities done by different organizations Table-1, does not include the same set of universities, which causes a difficulty in the comparison process. For the purpose of comparison we selected four organizations two of which have used webometrics based parameters to rank the universities and two have used traditional parameters, details are given in Table-5. As the number of universities ranked is also not same in the case of all these four organizations, we picked up a sample of first fifty universities from the list of each organisation and is shown in Table-1 and Table-2. Out of these fifty universities in four lists we retained the universities which are present in all four lists Table-4. We also made some shifting adjustment in the rank of universities in each of these four lists for eliminating the absent universities. Once we have the relative rankings for the four lists available, we compute r_s Spearman rank correlation coefficient to analyze the closeness or dispersion in the rankings so assigned. We take webometrics [30] ranking as the benchmark for webomnetrics-based rankings and then compare this ranking with the rankings assigned using the traditional parameters. In the following sections we present some of the details of these organisations for further details reader is referred to their respective web sites.

Table-1

Organisation	Method
Cybermetrics Lab (WEBO) http://www.webometrics.info/about.html	Webometrics
Academic Ranking of World Universities (ARWU) http://www.arwu.org/aboutARWU.jsp	Conventional (Non-Webometric)
Times Higher Education (TIMES) http://www.timeshighereducation.co.uk/	Conventional (Non- Webometric)
4 International Colleges and Universities (4ICU) http://www.4icu.org/top200/	Webometrics

WEBOMETRICS [30] RANKING IS BASED UPON THE FOLLOWING FOUR METRICS. [33]

- **Size (S)**. Number of pages recovered from four engines: Google, Yahoo, Live Search and Exalead. For each engine, results are log-normalised to 1 for the highest value. Then for each domain, maximum and minimum results are excluded and every institution is assigned a rank according to the combined sum.
- **Visibility (V).** The total number of unique external links received (inlinks) by a site can be only confidently obtained from Yahoo Search. Results are log-normalised to 1 for the highest value and then combined to generate the rank.
- **Rich Files (R)**. After evaluation of their relevance to academic and publication activities and considering the volume of the different file formats, the following were selected: Adobe Acrobat (*.pdf*), Adobe PostScript (*.ps*), Microsoft Word (*.doc*) and Microsoft Powerpoint (*.ppt*). These data were extracted using Google and merging the results for each filetype after log-normalising in the same way as described before.
- **Scholar (Sc)**. Google Scholar provides the number of papers and citations for each academic domain.

These results from the Scholar database represent papers, reports and other academic items.

The four ranks were combined according to a formula where each one has a different weight

VI. ACADEMIC RANKING OF WORLD UNIVERSITIES (ARWU) [31]

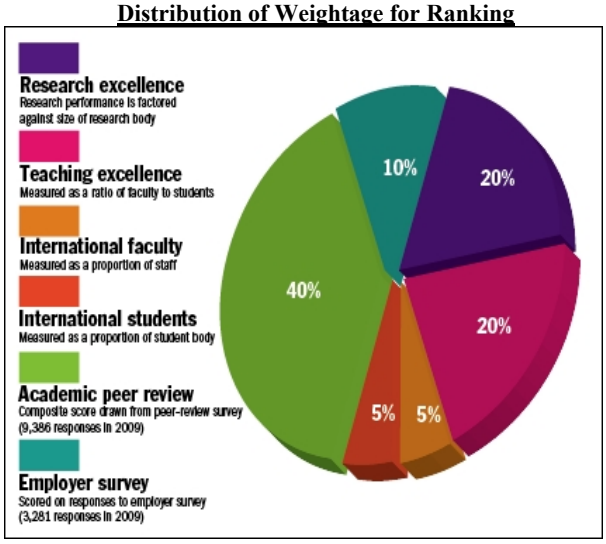
The ARWU, first published in June 2003 by the Center for World-Class Universities and the Institute of Higher Education of Shanghai Jiao Tong University, China, and then updated on an annual basis. ARWU uses six objective indicators to rank world universities, including the number of alumni and staff winning Nobel Prizes and Fields Medals, number of highly cited researchers selected by Thomson Scientific, number of articles published in journals of *Nature* and *Science*, number of articles indexed in Science Citation Index - Expanded and Social Sciences Citation Index, and per capita performance with respect to the size of an institution. More than 1000 universities are actually ranked by ARWU every year and the best 500 are published on the web.

Table 2 Indicators and Weights for ARWU

Criteria	Indicator	Code	Weight
Quality of Education	Alumni of an institution winning Nobel Prizes and Fields Medals	Alumni	10%
Quality of Faculty	Staff of an institution winning Nobel Prizes and Fields Medals	Award	20%
	Highly cited researchers in 21 broad subject categories	HiCi	20%
Research Output	Papers published in Nature and Science*	N&S	20%
	Papers indexed in Science Citation Index-expanded and Social Science Citation Index	PUB	20%
Per Capita Performance	Per capita academic performance of an institution	PCP	10%
	Total		100%

VII. TIMES HIGHER EDUCATION (TIMES) [32]

Distribution of Weightage for Ranking



VIII. 4INTERNATIONAL COLLEGES AND UNIVERSITIES (4ICU)

The 4ICU ranking is based upon an algorithm including three unbiased and independent web metrics extracted from three different search engines: [29]

1. Google Page Rank
2. Yahoo Inbound Links
3. Alexa Traffic Rank

Table-3

S.No.	Universities	d_1 (WEBO)	d_2 (ARWU)	d_3 (TIMES)	WEBO VS ARWU $d_J=(d_1-d_2)^2$	WEBO VS TIMES $d_K=(d_1-d_3)^2$
		Rank	Rank	Rank		
1	Harvard University	1	1	1	0	0
2	MIT	2	5	7	9	25
3	Stanford University	3	2	12	1	81
4	University of California	4	3	18	1	196
5	Berkeley Cornell University	5	11	11	36	36
6	Johns Hopkins University	6	13	9	49	9
7	California Institute of Technology	7	6	8	1	1
8	Carnegie Mellon University	8	19	15	121	49
9	University of California LA	9	12	17	9	64
10	University of Cambridge	10	4	2	36	64
11	Yale University	11	10	3	1	64
12	New York University	12	17	19	25	49
13	Duke University	13	16	10	9	9
14	University of Toronto	14	15	16	1	4
15	University of Oxford	15	9	4	36	121
16	University of Tokyo	16	14	14	4	4
17	Princeton University	17	7	6	100	121
18	University of Chicago	18	8	5	100	169
19	University of Edinburgh	19	18	13	1	36
	N=19				**Total 540**	**1102**

http://www.webometrics.info/=WEBO
http://www.arwu.org/=ARWU
http://www.timeshighereducation.co.uk/=TIMES
http://www.4icu.org/top200/=4ICU

IX. RESULTS

Spearman Rank Correlation

$$r_s = 1 - \frac{6\sum_i d_i^2}{n(n2-1)}$$

Rank Correlation -The value of rank correlation coefficient rs lies between -1 and 1

Table -4

S.No.	Comparison of Ranks	rs
1	Webometrics VS ARWU	.5277
2	Webometrics VS TIMES	.033334
3	Webometrics VS 4ICU	.3333334

Table-5

S. No.	University	d_1 (WEBO) Rank	d_4 4ICU Rank	$d_j=(d_1-d_4)$	d_j^2
1	Harvard University	1	3	-2	4
2	MIT	2	1	1	1
3	Stanford University	3	2	1	1
4	University of California Berkeley	4	4	0	0
5	Cornell University	5	5	0	0
6	California Institute of Technology	7	10	-3	9
7	University of Cambridge	10	7	3	9
8	Yale University	11	6	5	25
9	Duke University	13	8	5	25
10	University of Oxford	15	9	6	36

n=10 **Total= 110**

Table-6

WEBOhttp://www.webometrics.info/	**Weightage**
Visibility(external links)	50%
Rich files (web pages)	20%
Size	15%
Scholar	15%
ARWU =http://www.arwu.org/	**Weightage**
Quality of Education	10%
Quality of Faculty	40%
Research Output	40%
Per Capita Performance	10%
TIMES =http://www.timeshighereducation.co.uk/	**Weightage**
Research Excellence	20%
Teaching Excellence	20%
International Faculty International Students	5% 5%
Academic Peer Review	40%
Employer Survey	10%

X. DISCUSSIONS OF THE RESULTS

Refer to Table-5 the value of rank correlation coefficient for the rank comparison between Webometrics ranking and ARWU ranking is .5277 which is quite significant indicating some agreement between the two rankings. Webometrics ranking is based purely on web based metrics whereas ARWU makes use of non-web based (conventional) parameters to rank the universities. The value of rank correlation coefficient for the comparison between Webometrics and Times is .033334 indicating lesser agreement between the two rankings as compared to between Webometrics and ARWU. Times also makes use of non-web based (conventional) parameters to rank the universities. The interesting result is for the ranking comparison between Webometrics and 4icu. The value of rank correlation coefficient in this case is .3333, indicating comparatively lesser agreement between the two ranking despite the fact that 4icu also makes use of web based metrics as Webometrics to rank the universities.

A. Immunity of the Web Based Rankings

Web and webometrics is an emerging field. Owners of the commercial web sites understand the significance and value of the ranking of web pages of their web sites. This is how the terms like Search Engine Marketing (SEM) [35] and Search engine Optimization (SEO) became the buzzword of the IT industry. People and owners of the web sites in the academic world (in the global context) are still not aware of the strategic significance of web based publications, not more than required to solve their day to day problems. As the administrators and web site owners of the universities become aware of the strategic significance of the web and web contents the emphasis will be on web publications. At the same time there is a need of caution particularly for the organisations which make use of webometrics to rank universities or hospitals or something else, to carefully devise the mechanism so that unethical attempts to influence the ranks are prevented. For example in case of webometrics refer to table-7, 50% weightage is assigned to visibility and 15% to scholar, similarly rest 35% is assigned to Size and Rich files, this 35% is absolutely within the control of the website owners, and thus can be easily manipulated to influence the ranking order. What if all the graduate or undergraduate students are required to submit their assignments through the university web site ? Whereas other 65% assigned to visibility and scholar is out of control of the web site owners therefore more difficult to influence. Out of 65%, 50% is for the visibility that is to be assessed through the search engine indexed pages and count of the backlinks and 15% for the Google scholar which is still in beta stage and produces much faulty results. For example a query run on google scholar for knowing the count of publications of "Vikram University Ujjain" since 2009 reports 3267 publications yet a slightest twist in the query like "Ujjain Vikram University" returns the correct result as 117. There are numerous examples of Google scholar [34] returning quite inflated results. Remaining 50% is liable to be manipulated by unethical search engine optimization tricks, though search engine company keeps monitoring such attempts, but successes cannot be completely ruled out, thereby influencing the ranking orders.

XI. CONCLUSIONS

A Comparison was made between the rankings of world universities carried out by various profit/non-profit/research organisations on the web. Ranking of the universities has been done using the conventional parameters like the research out put, quality of faculty, patents registered etc. or webometrics parameters. Webometrics is a newly emerging discipline which provides web based parameters like backlinks of a web site, indexed pages in a search engine etc. which may serve as indicators to quantify various quality attributes of the entities like universities. The main purpose was to make a comparison between rankings using the conventional parameters and web based parameters. We also made comparison between two web based rankings (Webometrics vs 4icu) results of which are found to be in little agreement with each other as compared to the results of rankings done using the conventional parameters. We also made caution against the manipulability aspects of web based ranking parameters. In case of Webometrics parameters used for ranking 35% weightage is assigned to the parameters which are in direct control of the web site owners and hence subject to unethical manipulation to influence the rankings. Rest 65% is assigned to visibility (50%) and scholar (15%), the information which is provided by the search engines. It is this 50% weightage assigned for visibility (backlinks) which is a billion dollar business in the search engine marketing, against which the academic institutions will be required to remain alert.

REFERENCES

1. Tomas C. Almind and Peter Ingwersen (1997). "Informetric analyses on the World Wide Web: Methodological approaches to 'webometrics'". *Journal of Documentation* 53 (4): 404–426.
2. Mike Thelwall (2009). Introduction to Webometrics: Quantitative Web Research for the Social Sciences. Morgan & Claypool. ISBN 978-1-59829-993-9.
3. M Thelwall, L Vaughan, L Björneborn , Annual Review of Information Science and Technology.
4. B. Thackray and H. B. Brock, Eugene Garfield: history, scientific information and chemical endeavor, In B. Cronin and H. B. Atkins (eds.) The Web of Knowledge: A festschrift in honour of Eugene Garfield (Information Today, inc. ASIS Monograph Series: Medford, NJ, 2000) 11-23.
5. E. Garfield, Citation Indexing: Its theory and applications in science, technology and the humanities (Wiley Interscience, New York, 1979)
6. M. Thelwall, Bibliometrics to Webometrics, Journal of Information Science, 34 (4) 2007 pp.1-18.
7. C. L. Borgman and J. Furner, Scholarly communications and bibiliomet6rics, Annual Review of Information Science and Technology 36 (2002) 3-72.
8. G. K. Zipf, Human behavior and the principle of least effort: An introduction to human ecology. (Addison Wesley, Cambridge, MA, 1949)
9. H. F. Moed, Citation analysis in research evaluation, Information Science and Knowledge Management. (Springer, New York, 2005)
10. L. Leydesdorff, Why words and co-words cannot map the development of the sciences. Journal of the American Society for Information Science 48(5) (1997) 418-427.
11. L. Bjorneborn and P. Ingwersen, Toward a basic framework for webometrics, Journal of the American Society for Information Science and Technology 55(14) (2004) 1216-1227.
12. T. C. Almind and P. Ingwersen, Informetric analysis on the World Wide Web: Methodological approaches to 'Webometrics', Journal of Documentation 53(4) (1997) 404-426.
13. P. Mayr and F. Tosques, Google Web APIs: An instrument for webometric analysis? (2005)
14. I. F. Aguillo et al., Scientific research activity and communication measured with cybermetrics indicators, Journal of the American Society for Information Science and Technology 57(10) (2006) 1296-1302.
15. P. Ingwersen, The calculation of Web Impact Factors, Journal of Documentation 54(2) (1998) 236-243.
16. B. Cronin, Bibiliometrics and beyond: some thoughts on web–based citation analysis, Journal of Information Science 27(1) (2001) 1-7.
17. M. Thelwall, Extracting macroscopic information from web links, Journal of American Society for Information Science and Technology 52(13) (2001) 1157-1168.
18. O. Thomas and P. Willet, webometric analysis of departments of librarianship and information science. Journal of Information Science 26(6) (2000) 421-428.
19. F. Barjak and M. Thelwall, A Statistical analysis of the web presences of European life sciences research teams, Journal of the American Society for Information Science and Technology (2008)
20. F. Barjak, X. Li, and M. Thelwall, which factors explain the web impact of scientists' personal home pages? Journal of the American Society for Information Science and Technology 58 (2) 2007 200-211.
21. B. Cronin, H.W. Snyder, H. Rosenbaum, A. Martinson and E. Callahan, Invoked on the Web, *Journal of the American Society for Information Science* 49(14) (1998) 1319–1328.
22. K. Kousha and M. Thelwall, Google Scholar citations and Google web/URL citations: A multidiscipline exploratory analysis, Journal of the American Society for Information Science and Technology 58 (7) 2007 1055-1065.
23. L. Vaughan and D. Shaw, Bibliographic and web citations: what is the difference? Journal of the American Society for Information Science and Technology 54 (14) (2003) 1313-1322.
24. L. Vaughan and D. Shaw, Web Citation data for impact assessment: A comparison of four science disciplines, Journal of the American Society for Information Science and Technology 56 (10) (2005) 1075-1087.
25. K. Kousha and M. Thelwall, Motivations for URL citations to open access library and information science articles, Scientometrics 68(3) (2006) 231-288.
26. S. Lawrence and C. L. Giles, Accessibility of information on the web. Nature 400 (6740) (1999) 107-109.
27. J. Bar-Illan and B. C. Peritz, Evolution, continuity, and disappearance of documents on a specific topic on the web: A longitudinal study of 'informetrics', Journal of American Society for information Science and Technology 55 (11) (2004) 980-990.
28. Blaise Cronin, Journal of Information Science 2001; 27; 1, Bibliometrics and beyond: some thoughts on web-based citation analysis.
29. http://www.4icu.org/top200/ last accessed on 20th Aug. 2010
30. http://www.webometrics.info/top8000.asp last accessed on 20th Aug. 2010
31. http://www.arwu.org/ARWU2009.jsp last accessed on 20th Aug. 2010
32. http://www.timeshighereducation.co.uk/ last accessed on 20th Aug. 2010
33. http://www.webometrics.info/methodology.html last accessed on 20th Aug. 2010
34. Péter Jacsó, *University of Hawaii, Hawaii, USA* Google Scholar revisited, Journal: Online Information Review Volume: 32 Number: 1 Year: 2008 pp: 102-114.
35. Mike Moran, Bill Hunt. Search Engine Marketing, Inc.: Driving Search Traffic to Your Company's Web Site, Prentice Hall PTR Upper Saddle River, NJ, USA.
36. Thorste Joachims, Optimizing search engines using click through data, Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, 2002.
37. Ravi Sen, Optimal Search Engine Marketing Strategy, International Journal of Electronic Commerce, Volume 10, Number 1: Fall 2005, pp 9 – 25.
38. Zolt´an Gy¨ongyi, Hector Garcia-Molina, Link Spam Alliances, Proceedings of the 31st international conference on Very large data bases, 2005 - portal.acm.org .
39. Z Gyöngyi, H Garcia-Molina - Web spam taxonomy, Information Retrieval on the Web, 2005 – Citeseer.
40. Carlos Castillo, Debora Donato et. al. A reference collection for web spam ACM SIGIR Forum Volume 40, Issue 2 (December 2006), Pages: 11– 24, Year of Publication: 2006, ISSN:0163-5840.