Webometrics Rank Inspection: Proposed for Business Domain

Deepak Patidar
Department of CSE
PG Scholar
Shri Vaishnav Institute of Technology and Science, Indore, M.P (India)

Rajeev G. Vishwakarma
Department of CSE
Head of Department
Shri Vaishnav Institute of Technology and Science, Indore, M.P (India)

ABSTRACT
The management of any organization web sites is flattering more captious earlier than owed to the fast growth of the population reliant on the WWW as the most important information source. An organization can spread its achievements through its web sphere as well as accordingly achieve visibility and regulate from the web population. WRES assembles the parameters from the WWW repeatedly in stretchy periods and grant helpful knowledge immediately for the authenticated person of web sites. If the web performance of an organization is under the expected point according to their superiority, university authorities should go back over their web policies, elevating substantive increases of the degree and superiority of their electronic publications. The authenticated person on the website may espouse helpful approaches to endorse their rankings according to the hints given by the system.

Keywords
Web mining, ranking, webometrics, WWW, search engines, web site management

1. INTRODUCTION
The term webometric is a coinage from two modern English language words, "web" and "metric." The word web is a short form of World Wide Web. The scientific dictionary defined web as: "a hypermedia system that allows users to view and retrieve information from containing links and documents. On the other hand, metrics have to do with measurement or counting. Webster's Comprehensive Dictionary of English Language defined metrics as "the mathematical theory of measurement."

According to Bjorneborn and Ingwersen (2004), the explanation of webometrics expression is the survey of the quantitative aspect of the formation and use of knowledge possessions, structures and technologies on the Web illustration on bibliometric and informatics approach.

The expression webometrics was coined by Thomas Almind and Peter Ingwersen in 1997 identified the web as an important source for measuring documents and information. Information scientists recognized that many powerful web measurements could be conducted using the new powerful advanced search features of one of the top search engines of the day: AltaVista by Ingwersen, in 1998; Rodríguez I Gairín, in 1997. This was illustrated in a study that included a count of the amount of web pages in each Scandinavian country that linked to the pages of each other Scandinavian country (Ingwersen, 1998).

The ability to research web links using AltaVista proved particularly influential in triggering the research of webometrics. It was by the cause of hyperlinks are structurally similar to academic citations in the sense that they point from a source document to a target document. Academic citations have been used for many years before the web by those wishing to track or assess the impact of research (Borgman & Furner, 2002; Garfield, 1970; Moed, 2005; Nicolaisen, 2007). In parallel with this link analysis strand, other information scientists investigated the reliability and coverage of search engines and changes in the content of the web or individual collections of web pages (Bar-Ilan, 2004). These three types of web-based measurement research came to be collectively known as webometrics.

The term cybermetrics raised in parallel with the development of webometrics. This term was used to describe essentially the same research as webometrics and was the name of an electronic journal launched in 1997 as well as a series of workshops attached primarily to the biannual conference starting in 1996 of the worldwide humanity for Scientometrics and Informetrics. The term cybermetrics was particularly preferred in Spain, where the word webometrics has an unimpressive popular connotation.

Long after its creation, webometrics was given its accepted definition as "the survey of web-based phenomena using quantitative techniques and drawing upon informetrics method". The importance of this definition was its inclusion of informetric methods as a specifying characteristic, placing webometrics as a purely field of information science. Informetrics is a term used within information science to refer to quantitative research centered on measuring information. The definition thus excludes non-information science research based upon web measuring, such as statistical physics searches for mathematical laws of linking (Barabási & Albert, 1999) and computer science attempts to measure the size of the web (Lawrence & Giles, 1999), although in both these cases similarities could be drawn with prior informetric research respectively Marchionini Rousseau.

Webometrics is defined in a method-centered manner as the survey of net depicted object with principal quantitative methods for common skill investigate goals and using techniques that are not explicitly to one particular area of survey. The aim of this definition is to set webometrics free from informetrics and aim it at a wide social science audience. At the end, the definition is broad enough to encompass research by people who would not

describe themselves as webometricians but as researchers in other fields who were using web measurements.

This ranking system evaluates how strongly an organization is present in the web by its sub-pages, own web domain, scholarly articles, rich files and so on. The fundamental hypothesis of this approach is that web occurrence is a reliable pointer of the global concert and the status of the universities and as such. It is a tortuous way to measure all the university missions (teaching, research, transfer)

Webometrics orders world universities according to four indicators concerning the academic data shared on university web sites, and the indices are:

Size S, amount of web pages indexed by Google, Yahoo, and Live Search, and Exalead, weight 20%

Visibility V, the amount of incoming links indexed by Google, Yahoo, and Live Search, and Exalead, weight 50%

Rich files RF, amount of rich files in pdf, doc, ppt, and pps types indexed by Google, weight 15%

Scholar Sc, amount of scholar documents indexed by Google, weight 15%

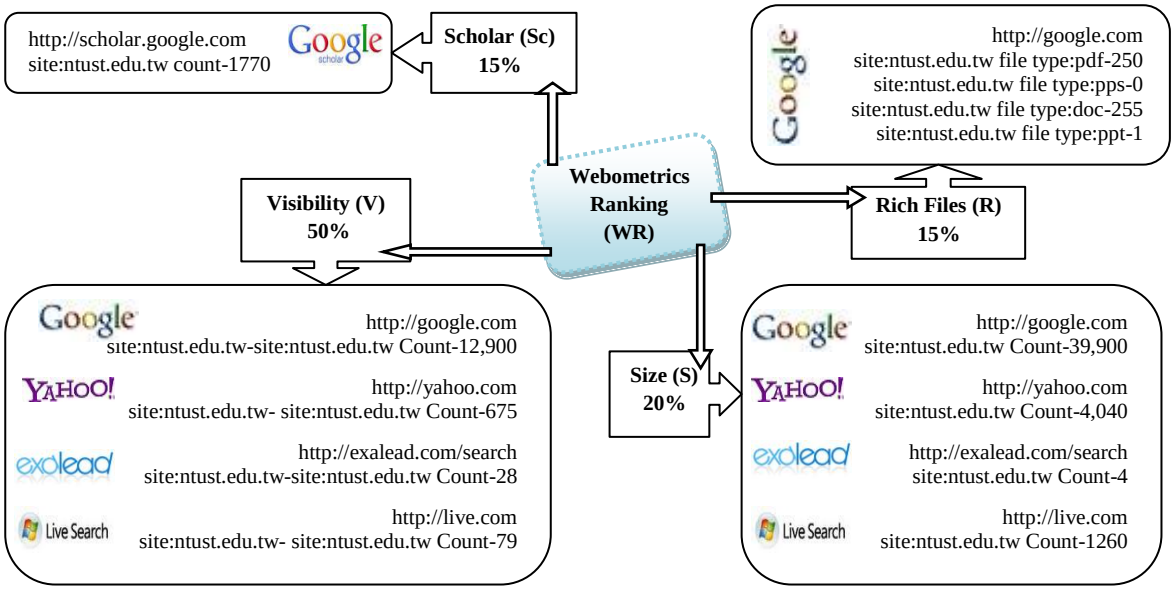

Figure 1 Queries of Webometrics Ranking for Different Search Engines

We can obtain the values of the indices by giving the queries shown in Fig. As an example to the corresponding search engines; Yahoo, Google, Live Search, and Exalead and then calculate the total score of a web site by substituting the values into the formula below:

WR= 0.2 × Score (S) + 0.5 × Score (V) + 0.15 × Score (RF) + 0.15 × Score (Sc)

2. BACKGROUND

The ranking of webometrics is well acknowledged as an important index of universities ready to promote them with the internet technology. ***[Shu et al]*** planned WRES as a before time caveat system for Webometrics Rankings [4].

WRES is constructed for the purpose of providing automatic early warnings for university website managers, and giving them more details and deeper insight of the WR ranking indicators. They employ web mining techniques to construct the early warning system WRES, which collects the WR indicator values and fills in the values in proper database tables for further display automatically.

With the fresh developed information visualization methods, ***[Rongying et al]*** analyzed the journalism themed at webometrics cited by Web of acquaintance [5].

Webometrics have become a new development direction and major research fields of Informetrics and has developed rapidly in various countries and regions in the world, ***[Rongying et al]*** [5] selected the co-citation data records by using "Webometrics" as the subject which were retrieved from ISI Web of Knowledge to make a visualization of statistical analysis, and obtained the following conclusions:

First, they draw a Knowledge map of cited-reference on Webometrics by using the information visualization tool.

Second, it is identified that "Webometrics", "information", "world-wide-web", "impact factors", "internet" are the hot areas and the frontier fields are: "websites", "social-network", "link-analysis", "research-groups", "power-law", "matching-urls", "hit-count", "business performance", "world-connectors", "web-publications", "url-citations", "small-world-connectors", "search-engines", "scientific-web", "school-web", "link-patterns" on Webometrics.

It has been observed that the research outcome need to gather speed the progress of internationalization, improve the international fame, increase the competitive ability in the world, and maintain international scientific and technical cooperation on sustainability to advance the ability of the international competition.

[Shu et al] proposed a novel cloud-based information providing system for Webometrics Rankings which employs web mining techniques to gather the WR indices. The proposed system is deployed on Microsoft Azure cloud platforms in order to give up-to-data and deeper insights of the WR ranking indicators for its users.

In 2012, there are at least 120,000 universities in the globe under estimate of WR; hence the cycle for a WR to process the complete data is six months. The proposed system on the cloud affords high speed data update and multiple result presentations to its users [6]. Moreover, it provides customized analyzed results and the foreknowledge of their superiority and weakness for the university website managers and authorities about website management.

As the role of webometrics playing in organization management and information management of network appears more and more important. This article [3] discussed the feasibility of extracting the new data sources of webometrics relying on Dublin Core (DC) element set based on the current data sources in order to offer a better data support for webometrics.

If each information publisher expresses and organizes information resources according to certain criteria, such as Dublin Core or other indicators of webometrics, therefore, network information resources exist in a standardized form. It will be easier to find more laws and more convenient for information users to use the web. So the standard interfaces to the information sources of webometrics could be found from the utilization and standardization of the existing rules and norms by the webometrics based on metadata, and then serve for webometrics and provide a simpler channel of information sources for the organization and management of network information. On this basis, it can also enrich webometrics methods.

In unusual world grading systems, educational institutes compete to achieve higher grading. The aim of search engine optimization is to find out whether ease of access can be employed as a device or not. Results indicate that there is a positive relation between webometrics and WCAG conformance. As for search engines ranking and WCAG conformance, results showed a weak positive trend almost in all cases which indicates that the accessibility could be recommended as a tool for search engine optimization.

The primary objective of this study was to examine the effect of conformance to WCAG on Webometrics ranking. In addition, this research also looked at the relation between WCAG conformance and SEO. This research [7] has proven that accessibility does influence Webometrics ranking and WCAG 2.0 should be used by higher learning institutes.

Ranking is the key problem for text applications and information retrieval. Recently, the ranking methods based on learning to rank (machine learning optimization), become the focus of researchers and practitioners. The main goal of these methods is to apply the various existing and effective algorithms for machine learning to rank. In this paper [8], we investigate the important papers in this sense; the pros and cons of the recent-proposed framework and algorithms for ranking are analyzed and the relationships among them are discussed.

From the practice now, the probabilistic approach appears to be attractive because of its robustness. Both Rank Net and List Rank can perform excellently in the noise and real data sets. Another is the list-wise method. The list-wise method which appears in 2007 is not perfect, and may have much vital force in learning to rank. Even the list-wise algorithms may have another set of the input and output and the evaluation if necessary.

In paper [9], they exploited several possible ways to develop a prototype (Link Discoverer) that collects data from both search engines and real-time links. The prototype consists of two parts: a crawling part of collecting real-time link data from a given domain or site and a search engine part for harvesting link data from search engines by using specific search commands.

Further studies are however required to deal with issues such as website edge demarcation, processing of script links and link filtering, inequity of multi-domain institutions.

The proportion of the green region in each academia that is captivated through orbiter can be employed as world academy rating indices. In this exertion [10], green images captured from Google Earth and have been processed during a segmentation phase by means of the hsv threshold method. At the effecting timely investigation, web processing consumes 19.5% greater than the dealing out that is done exclusive of web crossing point. Finally, the result shows that the hsv thresholding scheme is 21% enhanced than global thresholding scheme and 59% improved than local thresholding scheme. Image execution time using a website is approximately 122.23 seconds and not using the website is 98.39 seconds. The difference in time between the two of them is 23.83 seconds.

3. PROBLEM IDENTIFICATION AND PROPOSED SOLUTION

Here we study a number of papers related to webometrics ranking as given in references, and we found some new problems which are described here with their proposed solution:

(i) Any research organization known by their researches and their provided solutions in the given domain, they are published in their achievements and research using the websites. The web site administrator is responsible to keep updated the articles and research papers up to date, but sometimes it is unable to update the proper manner and organizations are lake in finding good grades. Webometrics provide the ranking for these kinds of websites by which web site administrator keeps track their records and their researches.

(ii) Due to the rare and limited area of use of this kind of web domain too few persons get knowledge about the webometrics. For the website management and webometrics design too few studies and literature are available, and this matrix is created only for the universities and their ranking according to their content. But the problem is that not only universities are working in research and other business directions. Required a generalized framework where web site managers of the different domain can register their web sites and view the ranking and position where they are in their competitive areas.

The proposed solution for the above problem is Design a new web metrics for all other business domains. Generalize the concept of webometrics. Represent their standing points for the website manager. We proposed an analysis web tool to improve the business quality and services in research direction.

4. APPLICATION

The application domain of webometrics and its ranking is described here:

4.1 Analysis of Link: Networks and Impact Measurements

Analysis of link drove the early webometrics research, principally through a combination of the development of improved methods and applications to a range of different contexts. Two types of studies emerged, link network analysis and link impact analysis.

Essentially the study of Link impact analysis compares the numbers of hyperlinks pointing to each website within a pre-defined set, such as all universities in a country or all departments within a discipline in a country.

Link network research created a network of the links among specified collections of websites in order to identify patterns of connectivity.

Two important components of link analysis are the software and methods for extraction of link data. Researchers were for many years able to gather hyperlink information from commercial search engines like AltaVista, Bing and Yahoo! via their advanced link search commands, only these tools were all eventually withdrawn.

4.2 To the Altmetrics from Web Citation Analysis

The succeeding type of webometrics to grow to be trendy was a web citation analysis: calculate online credentials to available academic papers. The rationale behind the before time research was to appraise whether the web could substitute traditional illustration databases to appraise the impact of articles in open entrée online journals and consequently also for all journals. This untimely study found that while counts of web credentials correlated with credential counts from standard databases, numerous of the web documents derived from non intellectual sources. As an outcome, the web appeared to be a mediocre source of citation bang indication for journals or personage journal articles.

This string of webometrics research gave path to further specific investigations into picky types of web credentials to scholarly publications such as documents from PowerPoint presentations, online syllabi and Google Books that within restricted areas, web-based citation counts can reveal different types of crash from the scholarly crash reflected by conventional quotation enumeration. For instance, online curriculum records could reflect the instructive impact or appraise of articles.

4.3 Analysis of Web Data for the Social Sciences as a Service

Webometrics research has expanded from general or academic web analyses to investigations of social websites, frequently by automatically downloading data from those websites either through a web crawler or through data requests sent through permissive routes. For example, exploiting the information-centered research approach, RSS feeds and blogs have been analyzed to detect public fears about science while social network sites have been investigated to detect language use and friendship patterns. YouTube has been analyzed for the factors associated with discussions attached to online videos and Twitter for the sentiment of public reactions to major media events. In all cases, the methods of the research have been webometrics – large scale data gathering and analysis for social science purposes – but the findings of the research have been targeted at disciplines outside information science, such as media studies, politics and science communication. Many of the programs used are now publicly available in the free software Webometrics Analyst.

4.4 Search Engines

A momentous sum of webometrics research has evaluated profitable search engines. The two main investigation topics have been about the extent of the coverage of the web and the accuracy of the reported outcome. Research into developing search engine algorithms (information retrieval), and into how search engines are used (information seeking) are not part of webometrics.

At 1999, a survey of the main search engines estimated that none covered more than 17.5% of the 'indexable' web and that the overlap between search engines was surprisingly low. Here the 'indexable' web is roughly the set of pages that a perfect search engine could be expected to find if it found all web site home pages and followed the links to find the remainder of pages on the sites. The absence of comparable figures after 1999 is due to three factors:

First, an obscure Hypertext Transfer Protocol technology, the virtual server, has rendered the sampling method of Lawrence and Giles ineffective.

Second, the sense of the rise of dynamic pages is no longer reasonable to talk in terms of the "total number of web pages".

Finally, given that search engine coverage of the web is only partial, the exact percentage is not particularly pertinent, unless it has substantially changed. One outcome of this research, however, was clear evidence that meta-search engines could give more results through combining multiple engines.

4.5 Measurement of Web 2.0

Web 2.0 is a term coined by the publisher Tim O'Reilly mainly to refer to web sites that are driven by consumer confine, such as blogs, Wikipedia and social network sites. In addition to the style of research on data mining, there have been many studies of Web 2.0 sites in order to describe their contents and explain user behavior in them.

A webometrics study of MySpace has indirectly investigated activity levels but focused on member profiles. Amongst other findings, this showed that about a third of registered members accessed the site weekly and the average reported age was 21. Although other research found that MySpace close friends tended to reflect offline friendships, both male and female users preferred to have a majority of female friends. Another study looked at the geography of friendship, finding that the majority of friends tended to live within a hundred miles, although a minority lived in the same town or city. Finally, market research companies have published many statistics about Web 2.0, despite the uncertain provenance of this data; the results sometimes seem reasonable and also, because of the cost of obtaining the data. An example is the announcement by HitWise that MySpace had supplanted Google as the most visited website by US users by December 2007.

5. CONCLUSION

Hence WR was designed to emphasize the sharing of web resources provided by the world organization. By the WR platform, any organization in developing countries can show their importance in the same manner as the ones in developed counties. In future this generalized system helps all the business and commercial organization that works in the

competitive environment. This may be useful for small industries, local business, universities, research labs, computer and IT, and various other domains of knowledge, science, engineering, technology and many more domains.

6. ACKNOWLEDGEMENT

With due respect we would like to inform, that the article which we had proposed is prepared for academic scenario, that is not feasible for industrial research purpose the references which we had taken from listed here, one or not only references there are many other references which we had taken from real world and from daily life aspects. So it is vary through to remember all those references.

7. REFERENCES

[1] Aguillo, Fernandez, and Ortega "Webometrics Ranking of World Universities: Introduction, Methodology, and Future Developments Proceedings of Higher Education in Europe", pp. 233-244, May 2008.

[2] Aguillo, I. and Ortega, "Mapping World-class universities on the Web", pp. 272-279, June 2006.

[3] Zhao Qingfeng, Shi Longqing, "Data Sources of Webometrics", pp. 1312-1315, IEEE 2011

[4] Shu-Ming Hsieh, Ssu-An Lo, Chiun-Chieh Hsu, Da-Ren Chen, "Applying Web Mining Techniques for Constructing Webometrics Ranking Early Warning System".

[5] Rongying Zhao, Jing Wang, "Research of International Webometrics in Frontier Domains in Visualized Information", 978-1-4244-5326, IEEE Sep, 2010.

[6] Shu-Ming Hsieh, Ssu-An Lo, Chiun-Chieh Hsu, Wei-Chao Lin, "Novel Cloud Service for Improving World Universities Ranking", 978-0-7695-4654, IEEE April 2012.

[7] Mohamed Ahmed Elgharabawy, Media Anugerah Ayu, "Web Content Accessibility and its Relation to Webometrics Ranking and Search Engines Optimization".

[8] Chuan He, Cong Wang, Yi-Xin Zhong, Rui-Fan Li, "A Survey on Learning to Rank", 1-4244-2096-4, IEEE 12-15 July 2008.

[9] Bo Yang and Jian Qin, "Data Collection System for Link Analysis", 978-1-4244-2917-2, IEEE 2008

[10] Riri Fitri Sari, Muhammad Sulaiman, Muhamad Nur Fajar, "Green Image Segmentation Analysis of Google Earth Image to Rank World Universities using PHP and MATLAB", 978-1-4577-0255-6, IEEE 2011.

[11] Samuel C. Avemaria Utulu, "Webometric Ranking and Nigerian Private Universities: A Case Study of Bells University of Technology".

[12] S. Thanuskodi, "A Webometric Analysis of selected Institutes of National Importance Websites in India", International Journal of Library Science 2012, 1 (1): 13-18.