DESIDOC Journal of Library & Information Technology, Vol. 35, No. 6, November 2015, pp. 427-435

Search Engines and Alternative Data Sources in Webometric Research: An Exploratory Study

Samir Kumar Jalal,* B. Sutradhar,** Kalyan Sahu,*** Parthasarathi Mukhopadhyay,**** and Subal Chandra Biswas*****

Central Library, Indian Institute of Technology Kharagpur, Kharagpur-721 302
*E-mail: *jalalsk1971@gmail.com; **bsutra@library.iitkgp.ernet.in*

****Aricent Infotech Centre, 14/2 Milestone,*
Old Gurgaon, Delhi Road, Gurgaon, Haryana-122 016
E-mail: kalyanxfactor@gmail.com

*****Department of Library and Information Science,*
University of Kalyani, Kalyani, West Bengal-741 235
E-mail: psmukhopadhyay@gmail.com

******Department of Library and Information Science,*
University of Burdwan, Rajbati, Burdwan, West Bengal-713 104
E-mail: scbiswas_56@yahoo.co.in

ABSTRACT

Web contents are interlinked at each other through hyperlinks. Inter-linking nature of web explores significant sources of information. In the context of exploring hyper-linking behaviour of the web and retrieving relevant information, search engines and web crawlers play a predominant role as data sources but search engines had mostly withdrawn their supports after December 2011. An attempt has been taken to evaluate search engines (Google, AoL, Bing, Yahoo!) using some criteria and found that AoL has the highest coverage among these search engines. The paper also identifies various alternative data sources to carry out webometric research. The finding of the study shows that majestic.com is a predominant and comprehensive data source among alternative data sources in webometric research.

Keywords: Webometrics, alternate data sources, search engine, evaluation criteria, classification of backlinks, exploratory study

1. INTRODUCTION

The web is a massive collection of heterogeneous information interconnected at each other through hyperlinks. The information on the web are not organised properly because anybody can publish any type of information and nobody has a control over it. The heterogeneity nature of information leads to a chaotic situation on the web. In such an environment, search engines are normally used to search for information. The experience shows that search engines retrieve relevant information with more irrelevant information against a query. Therefore, it prompted the researchers to undertake such type of study to evaluate the performance of search engines and recommends some effective measures for improvement its efficiency. The efficiency and effectiveness of search engine depend on the strength of its search techniques and algorithms. Effectiveness of search engine implies the capability to find out the right information and efficiency indicates the quickness of search results. Different search techniques like Boolean search operators (AND, OR, NOT), query-based search, etc. are applied by different search engines to retrieve the results.

The role of search engine is not only to retrieve information against user's query but also to provide support in webometric research. The field Webometrics emerged from Bibliometrics, scientometrics, informetrics and cybermetrics. Almind & Ingwersen[1] first coined the term Webometrics in 1997. Webometrics measures the world wide web (www) to know various aspects of web indicators. According to Bjorneborn & Ingwersen[2], webometrics is 'the study of quantitative aspects of the construction and use of information resources structure and

Received 10 June 2015, revised 24 August 2015, online published November 2015

technologies on the web drawing on bibliometrics and informetrics approach'.

Webometric research requires data from web sources for analysis and decision making of any organisation. These web based data are normally collected either from search engines or from web crawlers or from both. The performance and the popularity of website depend on effective analysis of usage data through various metrics such as web impact factor, number of inlinks, total number of pages, number of hits, page rank. The success of Webometric research depends on the appropriate selection of search engines. There are some limitations for commercial search engines as pointed out by Rousseau[3] & Bar-Ilan[4] to accept them as data collection tool for Webometric research but most of the webometrician used AltaVista, Yahoo! and Google as data collection tool. It may be pointed out that, these commercial search engines, as a matter of company's policy, had withdrawn their support in Webometric research after the end of 2011.

Another important data source for webometric research is web crawler. Web crawlers are the computer programs that are capable of retrieving pages from the web and extracting the links from those pages. There are various types of web crawlers such as site management programs, research crawlers, site downloaders, etc. The most popular personal web crawlers developed by Mike Thelwall for link analysis research are LexiURL and SocSciBot. LexiURL is a computer program developed by Prof. Mike Thelwall for webometric research. At present, LexiURL has been discontinued and replaced by LexiURL searcher (http://lexiurl.blogspot.in/) and SocSciBot is also a web crawler for link analysis research developed by Prof. Milke Thelwall for strengthening webometric research. http://socscibot.wlv.ac.uk/

But, specialised web crawlers, which were developed for webometric research are either having limited access or having limited features. Therefore, alternative data source in webometric research have to be identified to enhance the progress of the research.

2. LITERATURE REVIEW

Dwivedi, Joshi & Gupta[5] have evaluated three popular search engines, Google, Yahoo, AltaVista, based on three quality parameters which are in depth coverage, ideation and clarity and number of links of related article to query topic. Result shows that these three search engines do not differ significantly in their same query results.

Vaughan & Thelwall[6] have studied the coverage of commercial sites of four different countries using three major search engines (Google, AllTheWeb, Altavista) and they found significant differences in their coverage. They have also pointed out that web visibility of a site is very important to be indexed by search engines.

Bar-Ilan[7] studied on the performance over a time period of the search engines using a set of measures for testing and improving their functionality. The set of measures introduced in their work are: (a) technical precision;(b) technical relevance; (c) relative coverage; (d) new and totally new URLs; (e) forgotten, recovered, lost; (f) well-handled and mishandled; (g) self-overlap, number of rounds in which the URL is retireved; (h) persistent URLs. (a) and (d) played important role in their study results.

Chu & Rosenthal[8] evaluated and compared three search engines (Alta Vista, Excite, and Lycos) based on their search capabilities and retrieval performances. They found that Alta Vista was better performing than Excite and Lycos in both cases while Lycos had the largest coverage among them. For evaluating the web search engines, they also suggested some aspects namely composition of web indexes, search capability, retrieval performance, output option, user effort.

Spiteri & Richard[9] have used an evaluator in the studies which chooses the topic, formulates the search query and performs the relevance assesments on the documents retrieved from the search. Jalal[10] found that the relationship through link analysis that inlinks/outlinks to top ten Asian Universities are far less than top ten world universities. The percentage of inlinks and selflinks for top ten Indian and Asian universities were less than top ten world universities under study. Vaughan[11] reported in her study the problem of support in webometric research by commercial search engines and tried to find out the alternative data source such as Alexa and Yahoo! The result showed that there was a high correlation between Yahoo! and Alexa in terms of backlinks. Vaughan & Yang[12] studied three types of web data sources to resolve the lack of inlink data and recommended that Alexa inlink is better data source than Google URL and Yahoo!

3. RESEARCH QUESTIONS

The area of webometrics can broadly be categorised under: (a) Webpage content analysis, (b) web technology analysis, (c) web usage analysis, and (d) weblink structure analysis. Therefore, it is essential to know the data sources under each of these four categories of webometrics. The main purpose of search engines is to retrieve relevant information against a specific query from heterogeneous sources of information. Basically, the present study concentrates the link analysis research and its data sources. From the study above, a set of interlinked questions have been emerged:

(a) What are the criteria to be applied for evaluating search engines?

(b) What is the present status of webometric research in India?

(c) Do commercial search engines provide support in webometric research now-a-days?

(d) What are the alternative data sources in webometric research?

4. SEARCH ENGINES

Search engines are suits of computer programs that automatically find and download webpages and store them in a database. Search engines use program that links to the database to a user interfaces so that it can be interrogated through the internet. From the point of view of webometric research, search engines can broadly be categorised as: (a) Webometric supported search engines; (b) Webometric 'not supported' search engines.

4.1 Search Engines: Features

In the web environment, search engines played predominant role in information retrieval in general. Some search engines also support in webometric research through their special keywords. In information retrieval, generally, search engines are having simple search and advance search facilities with Boolean operators. One of the important features of search engines is that some search engines function as subject gateway. Subject gateway uses both automatic indexer and human indexer for indexing while search engine only uses automatic indexer. Other features of search engines are the use of metadata policy and handling web semantics. Strength of algorithm used in a search engine makes a difference in information retrieval especially in relevance and coverage. Handling web semantics carefully is the biggest advantage in improving the relevance of result. The performance of a search engine depends on the metadata policy they are using. Another feature of search engine is the capability of spam detection. Websites can increase its inlinks by increasing extra keywords and webpages (unwanted/spam).

4.2 Evaluation Criteria

At present, there are many search engines available. It is very difficult to choose the appropriate search engine at the time of their need in information. Wrong selection of search engine may yield more irrelevant and biased result[6]. Therefore, it is essential to evaluate search engines through some criteria to find out the most appropriate search engines.

Search engine are normally evaluated through some criteria. These criteria may be categorised: qualitative and quantitative. The qualitative criteria for search engine evaluations are recall, precision, relevance, use of search techniques and language supports. The quantitative criteria are size, response time, and database size.

Spiteri & Richard[9] have used an evaluator in the studies who chooses the topic, formulates the search query and performs the relevance assesments on the documents retrieved from the search. They used many criteria such as search features (Boolean operators, display (output features), precisions of retrieved documents, help features, recall, database size, speed or response time, database content.

4.2.1 Recall

Recall of a search engine indicates the ratio of number of relevant document retrieved to the number of relevant documents in the collection. Recall relates to the ability to search to retrieve the relevant documents. In a word, recall is the ability to retrieve relevant document. There is a trade-off between recall and precision. In the web environment, it is very difficult to know the number of relevant document in the entire web. Besides, the concept of relevancy of document is associated with the perception of user and also the content in which the information is sought. In the context of web, it should be pointed out here that recall value cannot be computed due to the lack of value of number of document in the collection.

4.2.2 Precision

Precision of a search engine indicates the ratio of number of relevant document retrieved to total number of document retrieved. Precision relates to its ability not to retrieve non-relevant documents from the collection. In the context of web environment, the precision value can be calculated as it would be possible to get the value of numerator and denominator.

4.2.3 Relevance

The relevance of search result is a relative concept because the identified relevant result may vary depending on the intuitive and notional judgement applied while categorising the information as relevant or irrelevant. Another characteristic is that the concept of relevance is subjective by nature. Therefore, it is very difficult to know the exact number of relevant documents. In other words, the relevancy of result may be judged by the user, who needs the information against a particular query submitted to a search engines.

4.2.4 Coverage

The selection of search engines depends on its coverage or size. The coverage of a search engine can be measured in various ways like subject based, domain based, sub-domain-based, country-based, content-based, etc. The subject-based search may be executed through single or multiple keywords with or without search operators. Domain-based search can be executed using special keywords, i.e., site:domain name or domain:domain name.

In this study, a combined approach has been adopted to obtain unbiased result. The objective to adopt the combined approach is to have participation from individual domain names (7 premier IITs as academic domain), country-code top level domains, i.e., ccTLDs (SAARC countries) and gTLDs (generic top level domain) like .com, .org, .net. Webometric special keywords 'domain:' and 'site:' have been used for data collection. The data have been collected during 20-22 August 2014. Based on the collected raw data, ranking (both individual and combined) of search engines has been made on the basis of values of number of webpages.

The ranks of these search engines are derived on the basis of collected raw data on their coverage at a particular point of time. Table 1 shows that Google search engines got the first rank based on the data received from SAARC countries. It also reflects almost similar result for the case of IITs. The search engine's result may vary over time[13]. In the case of TLDs and domains, Google search engine provides better than others whereas AoL proves to be the best in general for generic TLDs with respect to coverage.

Table 1. Evaluation of search engine based on coverage

Search engines	TLD (SAARC)	IITs (7 no.)	.com	.org	.net	Combined rank
AoL	2	2	3	1	2	1
Yahoo!	3	4	1	2	1	2
Google	1	1	4	4	4	3
Bing	4	3	2	3	3	4

4.2.5 Response time

Response time is very important in webometric research, as fastness of a website is the key to keep end-users active on the concerned website. Among four popular search engines mentioned, only Google is providing the response time and the range of response time varies between 0.15 and 0.41 with reference queries executed.

4.2.6 Search Techniques

The efficiency of search engine depends on judicious application of algorithm supported at the backend for the development of the search techniques. All search engines under study do not use same algorithm to build the search technique. This mechanism leads to differentiate in the comprehensiveness of the coverage of search engine. In other words, if a user wants to get the result against a particular query, say, 'Network Security', it may be found different results from different search engines. Result revealed that Google (20,72,75,800) got comparatively higher value than AOL (20,51,29,300), Bing (20,41,32,900), and Yahoo! (15,28,39,600).

4.2.7 Language Support

Nandasara[14], *et al.*, studied webpages of Asian languages using 42 Asian countries domains to analyse the language support and found that there was a serious digital language divide exists in Asian countries. In such a situation, it is important to know the support extended by search engines while indexing the webpages. The spiders in all search engines may not have the same capabilities to index webpages written in regional languages.

5. WEBOMETRICS RESEARCH

Webometrics is promising research field in LIS, computer science and computing & information technology. Almind & Ingwersen[1] introduced the application of informetric methods to the www, so called webometrics. They proposed a number of specific informetric parameters such as hyperlinks per webpages, link density on webpages distributed over type of documents and domain names. Björneborn & Ingwersen[2] defined webometrics as: 'The study of the quantitative aspects of the construction and use of information resources, structures and technologies on the web, drawing on bibliometrics and informetrics approach.'

A detailed link topology, web node diagram and various terminologies were developed[15]. The scope of webometrics can broadly be categorised as: (a) webpage content analysis, (b) web technology analysis, (c) web usage analysis, and weblink structure analysis.

To determine the scope of webometrics, it is required to know the relationship between bibliometrics, informetrics, scientometrics, and webometrics. It may be noticed that webometrics is associated with bibliometrics and overlaps scientometrics to some extent[2]. Thelwall, Vaughan & Björneborn[16] contributed article on webometrics to demonstrate basic concept, origin, scope and coverage of webometrics and related reviews. The issues on data collection methods and measurement techniques of web related activities were addressed. Almind & Ingwersen[1] took an initiative to introduce and argue that it is possible to apply informetric methods to the web. webometrics covers all network-based communication using informetric or other quantitative measures. Thomas & Willett[17] described a webometric analysis of the linkages to websites associated with individual departments of library and information science (LIS) in UK universities. The findings of the study revealed that it was not possible to identify any significant correlation between the citation data and peer evaluations of research excellence embodied in the Research Assessment Exercise (RAE) ranking.

6. DATA SOURCES IN WEBOMETRIC RESEARCH

Mainly there are two categories of data sources in webometric research: (a) Commercial search engines, and (b) Personal web crawlers.

6.1 Search Engines as Data Sources

The most popular search engines in webometric research are AltaVista, Yahoo!, Google, Hotbot, Excite, MSN, etc. Bing emerged as a search engine in 2009 in place of LIVE search, previously known as Windows Live. Later Bing owned Yahoo! and AltaVista in 2011. Henceforth, Yahoo! has shut down its site explorer features and stopped to support webometric research. Similarly, AltaVista also withdrew its support of webometric research.

The webometric analysis is based on the data collected from the web using various search engines. In each search engine, there are some specific search keywords assigned by the search engines to retrieve the information from the Web.

Table 2 explains the webometric query syntaxes. Here 'abc' implies a particular domain to whom one wants to retrieve data. For example, if someone wants to know how many files (.doc) are there in IIT Kharagpur domain, he/ she should use the query like site:iitkgp.ac.in <space> filetype:doc

6.2 Web Crawlers as Data Sources

Another important data sources are personal web crawlers. The most popular personal web crawlers, which are used in link analysis research, are SocSciBot and LexiURL. These two web crawlers are developed by Prof. Mike Thelwal, University of Wolwerhampton, UK in order to find out alternative link analysis strategy. These web crawlers crawl webpages and download them in a local machine and then, tries to analyse them using integrated analytical software, i.e., Pajek, Cyclist, Ucinet, NetDraw, etc., to analyse the data and to build network graphs for representation of link data. It should be kept in mind that during the crawling processes, personal web crawlers take help search engines. Pajek is a computer program used for large network analysis. The present version of pajek is 4.05 available for download from http://mrvar.fdv.uni-lj.si/pajek/

Table 2. Webometric query syntax with results

S. No.	Search command	Results	Supported till Nov. 2011	Suppo-rted now
1.	domain:abc	Total no. of webpages	Google, AltaVista, Yahoo!	Google, AoL
2.	site:abc	Total no. of webpages	Google, Bing, Yahoo!	Google, Bing, Yahoo!
3.	linkdomain: abc –domain: abc	Total no. of inlinks	AltaVista, Yahoo!	No
4.	linkdomain: indomain: in	Total no. of self-links	AltaVista, Yahoo!	No
5.	linkdomain	Total no. of links	AltaVista, Yahoo!	No
6.	site:abc file:html	Report total no. of html files	Google, AltaVista, Yahoo!	Google, AltaVista, Yahoo!
7	site:abc filetype:doc	To know the file types under a particular domain	Google, AltaVista, Yahoo!	Google, AltaVista, Yahoo!

Cyclist is a text search engine, not a link analysis program. It works with Socscibot for text analysis purpose; UCINET 6 is a software package for the analysis of social network data. It works with NetDraw visualization tool. https://sites.google.com/site/ucinetsoftware/home; and NetDraw is visualisation tool developed by S.P. Borgatti used for large data. Current version is 2.155 as on August 2015.

7. CHALLENGES IN WEBOMETRIC RESEARCH

Webometrics works on scholarly document and non-scholarly documents. Scholarly documents include scholarly publications including e-journals, e-books, patents, technical reports, etc. Non-scholarly documents include webpages commercial, academic, social networking sites, etc., published by individuals, blogs and portals where no peer-review system is being followed. The biggest challenges in webometric research are in the area of finding out the data sources and the techniques for data collection. Among the four areas of webometric research, link analysis is getting affected more and more after commercial search engines had withdrawn supports in link analysis research. Therefore, there is a need to find out alternative data sources.

8. ALTERNATIVE DATA SOURCES

In the first decade of 21st century, most of the search engines supported webometric research through their special keywords such as site: domain, linkdomain, linkfromdomain, etc. In 2012 onwards, there was a tremendous change reflected in the data source of webometric research as a matter of policy undertaken by the owner of search engine. As a result, most of the search engines have withdrawn their support in webometric research. Researchers in the field of webometric were tried hard to search for 'Alternative Data Source' to carry out webometric research. Followings are some of the sources where webometric data can be collected.

8.1 Alexa Internet (www.alexa.com)

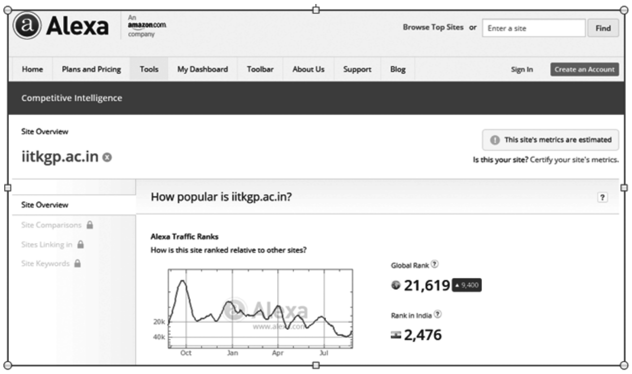
Alexa Internet was founded in 1996. As an SEO tool, Alexa collects data on browsing behaviour of users while visiting in websites and through its analytical tools, data are being analysed to provide global rank, country rank, web traffic data, total sites linking to a particular domain etc. 'Total Sites Linking In' implies that a particular domain i.e., iitkgp.ac.in is getting inlinks (2231) from different sites. In other words, IIT Mumbai has received inlinks from 6043 unique sites. Table 3 provides data for premier IITs. Table 3 shows that IIT Mumbai has received highest links as compared to other IITs listed above. The table also identifies an inverse relationship between country rank and 'Total Sites Linking In'. Global rank represents the popularity of website and it is being calculated using a combination of average daily visitors to this site and page views on this site over the past three months. A site's ranking is based on combined measures of unique visitors and page views. Unique visitors are determined by the number of unique Alexa users who visit a site on a given day. Page views are the total number of Alexa user URL requests for a site. It is also possible to made comparison between two or more sites.

8.2 Alexa's Toolbar Service

Alexa's Toolbar Service is a small software program to be installed in a computer. On installation with agreed terms and conditions, Alexa's toolbar service collects and stores information about the webpages, websites, and the other websites.

Table 3. Alternative data source through alexa.com

Name	Domain	Global rank	Rank in India	Alexa traffic (%)	Total sites linking In
IIT, Mumbai	iitb.ac.in	12155	897	10.7	6043
IIT, Madras	iitm.ac.in	23592	2209	14.9	3965
IIT, Kanpur	iitk.ac.in	15628	1166	14.8	3404
IIT, Delhi	iitd.ac.in	18847	1599	17.7	2752
IIT, Kharagpur	iitkgp.ac.in	19403	2063	13.3	2231
IIT, Roorkee	iitr.ac.in	64510	4233	12.3	1206
IIT, Guwahati	iitg.ac.in	128870	9733	23.9	622



Site Overview

iitkgp.ac.in

This site's metrics are estimated

Is this your site? Certify your site's metrics.

How popular is iitkgp.ac.in?

Alexa Traffic Ranks

How is this site ranked relative to other sites?

Global Rank 21,619

Rank in India 2,476

Figure 1. Alexa traffic ranks from www.alexa.com dated 28-08-2014.

8.3 Who.is

In 2005, Who.is came up as a web portal to look up domain information of any organisation. Who.is has offered a unique tool to look up IP addresses, location, DNS name server, related domain availability, etc., for the organisation or university, etc., for which domain has been subscribed for. It also provides data on domain registration, date of expiry, date of updation and domain popularity. Domain popularity is calculated by using a combination of average daily visitors and pageviews on a particular domain over the past month. The site with the highest combination of visitors and pageviews is ranked number one. Besides, the portal also provides contact information, content data, and traffic data.

8.4 Webconfs.com (www.webconfs.com)

Webconfs.com is another source of webometric data using the web tools and the SEO (Search Engine Optimisation) tools. Various methods to rank the websites 'Age of Website' is a major indicator. This value can be calculated also using Domain Age Tool[18]. The result is reported in terms of months by differentiating the date of creation of the website and the time of study (in this case 13 Aug 2014). It provides powerful tools to find out the age of website, last update, domain name, domain ID, expiry date, address of the organisation etc. http://whois.domaintools.com/. This tool display the approximate age of website (www.webconfs.com/domain-age.php).

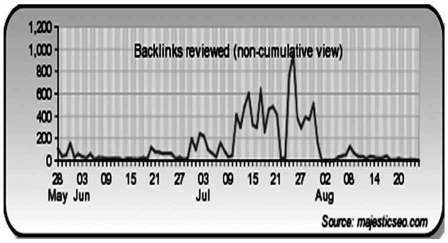
8.5 Majestic SEO Tool (http://www.majesticseo.com)

Backlinks, popularly known as inlinks or incoming links or inbound links are links to a website. Backlinks are one of the indicators of the popularity of a website. Majestic.com is one the best tool for backlink checker[19]. Search engines also use backlinks as one of the indicators to measure the popularity and rank of the website. Majestic.com is an SEO tool which provides much information on the website.

Backlinks are popularly known as external links or incoming links or inbound links. Backlinks is a link received by any node of University A (i.e., English and Foreign Languages University) from any web node. Node may be any webpage or directory or website, etc. Figure 2 shows backlinks for English and Foreign Languages (EFL) University retrieved as on 30-08-2014. It has been noticed here that a sudden increase of backlinks during 9 July to 31st July 2014, whereas number of backlinks received by EFL University, Hyderabad has found to be significantly less during early June and throughout August 2014.

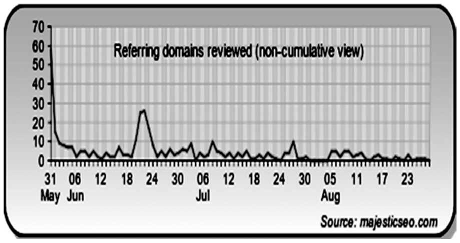
Domain is the organisation/institute/university's unique descriptor (e.g., efluniversity is the domain of English and Foreign Language University). It lies within its URL, i.e., http://www.efluniversity.ac.in/. Now, referring domain is a domain from which a backlink is pointing to a page or link. Therefore, Fig. 3 provides an idea about the nature of referring domains for the case of English and Foreign Languages University on 30-08-2014.

Figure 4 shows the external backlink profile and referring domain profile. It demonstrates the positive relationship between citation flow and trust flow for both the cases. Citation flow is a metric which predicts the influence of a link in a site. It does not judge the quality of a link. Trust flow indicates the trustworthy of a link. The value of trust flow depends on quality of backlinks. If trust flow increases, the citation flow will increase but not vice versa.



Source: majesticseo.com

Figure 2. External Backlinks reviewed over the last 90 days.



Source: majesticseo.com

Figure 3. Referring Domains reviewed over the last 90 days.

8.6 Linkvendor.com

Linkvendor is a professional SEO tool powered by searchmetrics[20]. It provides basically all data on site visibility and social visibility. Site visibility is analysed through PageRank[21], Metatag analyser. keyword tools, server and domains (domain age, domain popularity, reverse IP), link tools (link popularity, backlink counts, link value, outbound links). Social visibility provides data on social links from Facebook, Twitter, Linkedin, Google+, etc.

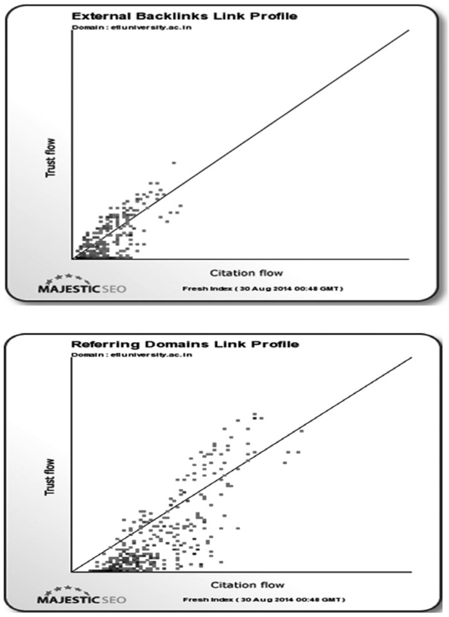


Figure 4. External backlink and referring domain profile of efluniversity.ac.in.

9. CLASSIFICATION OF BACKLINKS

Backlinks are important source of information. The page rank of a domain increases depending on the quality backlinks. Classification of hyperlinks may be on broad categories or sub-categories. It would be an interesting to know under which sub-category, more and more links are received by an institute/university.

Table 4 shows the classification of back links URL. It also tries to focus under broad categories and sub categories. The Table also shows a sample of five links and its classification.

10. DISCUSSIONS

With respect to the findings of search engines, Dwivedi, Joshi & Gupta[5] found that search engines (Google, Yahoo! and AltaVista) do not differ significantly whereas Vaughan & Thelwall[6] agreed that search engines (Google, AllTheWeb and AltaVista) found

Table 4. Classification of backlinks

S. No.	Backlink URL	Category	Broad category
1.	http://www.osmania.ac.in/AboutUs-OtherLinks.htm	Osmania University	Higher Education
2.	http://career.webindia123.com/career/website.htm	Education & Career	Education & Career
3.	http://www.kangwon.ac.kr/english/menu1/sub_01_05.php	Kangwon National University	Higher Education
4.	http://www.caluniv.ac.in/useful-links/useful_links.html	Calcutta University	Higher Education
5.	http://blogs.iucr.net/crystalmath/	Blog	Blog

significant differences in their coverage. On the other hand, Chu & Rossenthal[8] found that Lycos had the largest coverage than Altavista and Excite. In this context, the present study is conducted using search engines (AoL, Google, Yahoo! and Bing) and found that in case of TLDs and domains, Google search engines perform very well whereas AoL proves to be the best among these search engines in general.

As far as alternative data sources in webometric research are concerned, Vaughan & Yang showed that Alexa inlinks is better data sources than Google URL and Yahoo! The present study revealed that majestic SEO tool is another comprehensive data source in webometric research.

11. CONCLUSIONS

It is a fact that webometric study is gradually gaining lot of importance due to increasing dependence on the web. The study tries to list out some criteria for evaluating search engines to know at least the coverage. The support of commercial search engines in webometric research has become limited with respect to web link analysis especially due to withdrawal of keyword like ‘linkdomain’ in search query syntax. Therefore, alternative data sources are highly required to make the web-related data available to strengthen the webometric research. Some alternative data sources have been mentioned in this paper. Yet, further research is required to find out validity and acceptability using statistical techniques of data retrieved from various alternative data sources.

REFERENCES

1. Almind, T.C. & Ingwersen, P. Informetric analyses on the world wide web: Methodological approaches to Webometrics. *Journal of Documentation,* 1997, **53**(4), 404-26.
2. Björneborn, L. & Ingwersen, P. Towards a basic framework for webometrics. *J. of the Amer. Soc. for Inf. and Tech.*, 2004, **55**(14), 1216-27.
3. Rousseau, R. Evolution in time of the number of hits in keyword searches on the internet during one year, with special attention to the use of the word euro. *In* Proceedings of the 8th International Conference on Scientometrics & Informetrics. IISI, Sydney, 2001, pp. 619-27.
4. Bar-Ilan, J. Data collection methods on the web for informetric purposes—A review and analysis. *Scientometrics*, 2001, **50**(1), 7-32.
5. Dwivedi, N.; Joshi, L. & Gupta, N. Statistical analysis of search engines (Google, Yahoo and Altavista) for their search result. *Inter. J. of Comp. The. and Engi.,* 2013, **5**(2), 298-301.
6. Vaughan, L. & Thelwall, M. Search engine coverage bias: Evidence and possible causes. *Inf. Proc. and Manag.*, 2004, **40**(4), 693-707.
7. Bar-Ilan, J. Methods for measuring search engine performance over time. *J. of the Amer. Soc. for Inf. Sci. & Tech.,* 2002, **53**(4), 308-19.
8. Chu, H. & Rosenthal, M. Search engines for the world wide web: A comparative study and evaluation methodology. ASIS, 1996.
9. Spiteri, F. & Richard, N. Evaluation of internet search engines: Methodological issues and assumptions. *In* Proceedings of the 27th Annual Conference of the Canadian Association for Information Science, 1999.
10. Jalal, S.K. A comparative weblink analysis among top Indian, Asian and World universities. *DESIDOC J. of Lib. & Inf. Tech.*, 2013, **33**(2), 131-40.
11. Vaughan, L. An alternative data source for web hyperlink analysis “Sites Linking In” at Alexa internet. *COLLNET J. of Scient. and Infor. Manag.*, 2012, **6**, 31-42.
12. Vaughan, L. & Yang, R. Web data as academic and business quality estimates: A comparison of three data sources. *J. of the Amer. Soc. for Inf. Sci. & Tech.,* 2012, **63**(10), 1960-72.
13. Bar-Ilan, J. Search engine results over time: A case study on search engine. *Cybermetrics,* 1999, **2/3**(1), Paper-1. http://cybermetrics.cindoc.csic.es/articles/v2i1p1.html (accessed on 10 October 2015).
14. Nandasara, S.T. *et al.* An analysis of Asian language webpages. *The Inter. J. of Adv. in ICT for Emer. Reg.,* 2008, **1**(1), 12-23.
15. Björneborn, L. & Ingwersen, P. Perspectives of webometrics. *Scientometrics*, 2001, **50**(1), 65-82.
16. Thelwall, M.; Vaughan, L. & Björneborn, L. Webometrics. *Ann. Rev. of Inf. Sci. & Tech.*, 2005, **39**, 81-135.
17. Thomas, O. & Willett, P. Webometric analysis of departments of librarianship and information science. *J. of Inf. Sci.*, 2000, **26**(6), 421-28.
18. DomainAgeTool. http://www.webconfs.com/domain-age.php (accessed on 10 October 2015).
19. Majestic. https://majestic.com/ (accessed on 11 October 2015).
20. Searchmetrics Rapid. http://rapid.searchmetrics.com/en/tools/ (accessed on 11 October 2015).
21. PageRank. Check PageRank of Web site pages instantly. http://www.prchecker.info/check_page_rank.php (accessed on 10 October 2015)

About the Authors

Dr Samir Kumar Jalal is presently working as Deputy Librarian at Central Library, Indian Institute of Technology Kharagpur, West Bengal. He has 41 publications in many journals and conferences. He has 14+ years of working experiences in reputed institutes and universities. His key research interests are: Webometrics, information retrieval, and digital library services.

Dr B. Sutradhar is presently working as Librarian in Central Library, Indian Institute of Technology Kharagpur, West Bengal. He has more than 22 years of working experiences in reputed institutes. Currently, he is actively involved in National Digital Library (NDL) project.

Mr Kalyan Sahu obtained his MTech on Information Security from the Department of Computer Science and Engineering, Birla Institute of Technology Mesra, Ranchi. He received GATE fellowship while pursuing MTech. At present, he is working as Software Engineer in Aricent Infotech Centre since October 2014.

Dr Parthasarathi Mukhopadhyay is working as Associate Professor and Head in the Department of Library and Information Science, University of Kalyani, West Bengal. He has 16 years of teaching experience in the field of Library and Information Science. His areas of interests are: Webometrics, information retrieval, community information and library automation, etc.

Dr Subal Chandra Biswas is working as Professor in the Department of Library and Information Science, University of Burdwan, West Bengal. He has 30 years of teaching experience in the field of Library and Information Science. He has guided many research scholars. His areas of interest are: Indexing languages and systems, community information, information seeking and retrieval in digital environment.