A New Frontier of Informetric and Webometric Research: Mining Web Usage Data

Liwen Vaughan[1]

25 September 2008

Abstract

Many Webometric researches have been conducted since the early days of the Web. Most have focused on Web hyperlink as data source. Very few used Web usage data, a very rich data source that can be analyzed for various purposes. The objective of this paper is to encourage more research into this new frontier and to advance informetric and Webometric research with new Web data sources. I will first introduce the concept of Web data mining and discuss how Web data mining relates to traditional informetric research. I will then discuss types of Web usage data that are available and provide examples of studies that used these types of data. I will also discuss the limitations of Web usage data.

1 What is Web usage mining?

To answer this question, we first need to discuss what is Web data mining. Web data mining is also called Web mining. There are various definitions of the term. For example, Thuraisingham (2003) defines Web mining as "essentially mining the databases on the Web or mining the usage patterns and structure so that helpful information can be provided to the user" while Wikipedia (2008) definition of Web mining is "the application of data mining techniques to discover patterns from the Web". I define Web data mining simply as "analyzing large scale Web data to discover patterns, regularities and relationships". Web data mining is generally classified into the following three types: Web structure mining, Web content mining, and Web usage mining. Web structure mining, which examines Web hyperlink patterns, is perhaps the most familiar type to informetrics and Webometrics researchers. In fact, Webometrics started with analyzing Web hyperlink data (Ingwersen, 1998). Web content mining explores Website content, often the texts of Websites. For example, Liu, Ma, & Yu (2003) compared patterns of keywords on the Website of a company with that on the competitor's site to discover unexpected information for business competition. Vaughan and You (2008) combined Web content mining with Web usage mining to for business information. Web usage mining aims to determine navigation patterns of Web users. This type is the focus of this paper. Different forms of Web usage data and how they can be studied are discussed in detail below.

2 Why is Web usage mining a new frontier of informetric and Webometric research?

To address this question, we first need to examine the question of how Web data mining relates to informetrics, scientometrics, and Webometrics. As we know, informetrics, scientometrics, Webometrics are about analyzing large scale data, mostly quantitative data, to discover patterns and regularities. For example, all three basic informetric laws, i.e. Bradford law, Lotka's law, and Zipf's law, are about the patterns summarized from large scale data. The Bradford law, for instance, summarizes the pattern of distributions of papers on a particular over journals.

Informetric research has the tradition of studying usage data. For example, numerous studies ana-

[1]University of Western Ontario, Faculty of Information and Media Studies, Canada, lvaughan at uwo dot ca

H. Kretschmer & F. Havemann (Eds.): Proceedings of WIS 2008, Berlin
Fourth International Conference on Webometrics, Informetrics and Scientometrics & Ninth COLLNET Meeting
Humboldt-Universität zu Berlin, Institute for Library and Information Science (IBI)

lyzed library circulation data (Burrell & Fenton, 1994; Decroos et al, 1997; Tague & Ajiferuke, 1987) and library shelving data (Duy & Vaughan, 2006). From this perspective, studying Web usage data should be a natural extension of informetric research. However, very few studies have been carried out using Web usage data while most Webometrics studies so far focused on Web hyperlink data. There is no shortage of Web usage data. They can be found from various sources and in various formats. The two common types that will be discussed in this paper are Web server log data that records usage of a particular Website and Web traffic data (typically collected from Web toolbar users) that show Web visits to many sites. These data contain rich information on how Websites are being used. Finding patterns of regularities from Web usage data will contribute greatly to our knowledge of the Web. This rich data source has so far not been tapped into extensively by the informetrics community. This is why I believe that mining Web usage data is a new frontier of informetric and Webometric research.

3 Usage data of a particular Website

A Web server log is a record of the activities of the Web server. Computer programs to keep the log often come with server software, e g. Microsoft Internet Information Server has built in function to keep a server log and this function can be turned on or off as the site owner wishes. The following data are usually available from a server log: date and time of the request (when a Web user downloads a Webpage, it is a request to the Web server for that page); IP address of the requestor; HTTP request (the specific file that is requested); referrer URL (from whom did the user find the site, e.g. from a link pointing to the site or from the search result of a search engine); and user agent (requestor's browser and operating system).

Based on these data in a Web server log, much information can be extracted. From data about the date and time of the request, we can determine traffic patterns over time, e.g. peak time in a day or busy days over a month or a year. We can also find out what are the popular pages viewed and what are the common entrance page (the first page of the site that a user visited) and exit pages (a user visited this page and then left the site). We can also see what queries (either keywords or phrases) that users searched for on the site. This tells us what kind of information that users were looking for on this site. Data about the IP address of the requestor allow us to infer the geographic regions of people who visit the site. This is very useful information that tells us if the site is reaching a particular audience, e.g. international audience. Data about referrer URL inform us how users found the site, e.g. if they used a search engine to find the site and what queries they used at the search engine.

Because so much data are recorded for a single request, Web server log files are typically very large in size. The raw data are messy looking and they are not meant for human browsing. Computer programs are usually used to analyze Web server logs. Commercial software packages are available and a commonly used one is WebTrends. Customized computer programs or scripts can also be written to extract needed information.

4 Data sources about Web traffic to multiple sites

Server logs of individual Websites discussed above provide rich information about usage of a specific site. However, they do not tell us how a site compared with other similar or related sites. For example, the server log of a particular company provides information on visits to this site but it does not tell us how well this company site is doing compared with other similar sites or competing sites. The same applies to a university site or a government site. We often need to know traffic to other sites as well. This is the time that traffic data of multiple sites are useful. There are several such data sources. Alexa at http://www.alexa.com is perhaps the best known one. It is currently the only large scale free data source of this kind. Another large scale data source is Compete at http://www.compete.com but it is partially free. Several fee based data sources are available such as Hitwise at http://www.hitwise.com, comSocre at http://www.comscore.com, and

Nielsen//NetRatings at http://www.netratings.com. Alexa and Compete will be discussed in more detail in this paper because they are free or partially free.

Alexa data are collected from users of Alexa Toolbar (also called Amazon Toolbar). The first Alexa Toolbar was released in 1997 and currently there are millions of Toolbar uses around the world (Alexa, 2008). For a given website, Alexa will provide data on page views, reach, and traffic rank. Page views measure the number of pages viewed by site visitors while reach measures the number of users (Alexa, 2008). Alexa calculates traffic rank based on a combined measure of page views and reach. The traffic rank scores averaged over a three month period are used to rank all Websites. Alexa also provide other data about a site such as the breakdown of countries where users come from. One can compare traffic patterns of different Websites by entering the URLs in question. A graph is then produced to show the comparison. Figure 1 shows the comparison between www.uwo.ca (the University of Western Ontario, Canada) and www.hu-berlin.de (Humboldt-Universität, Germany).

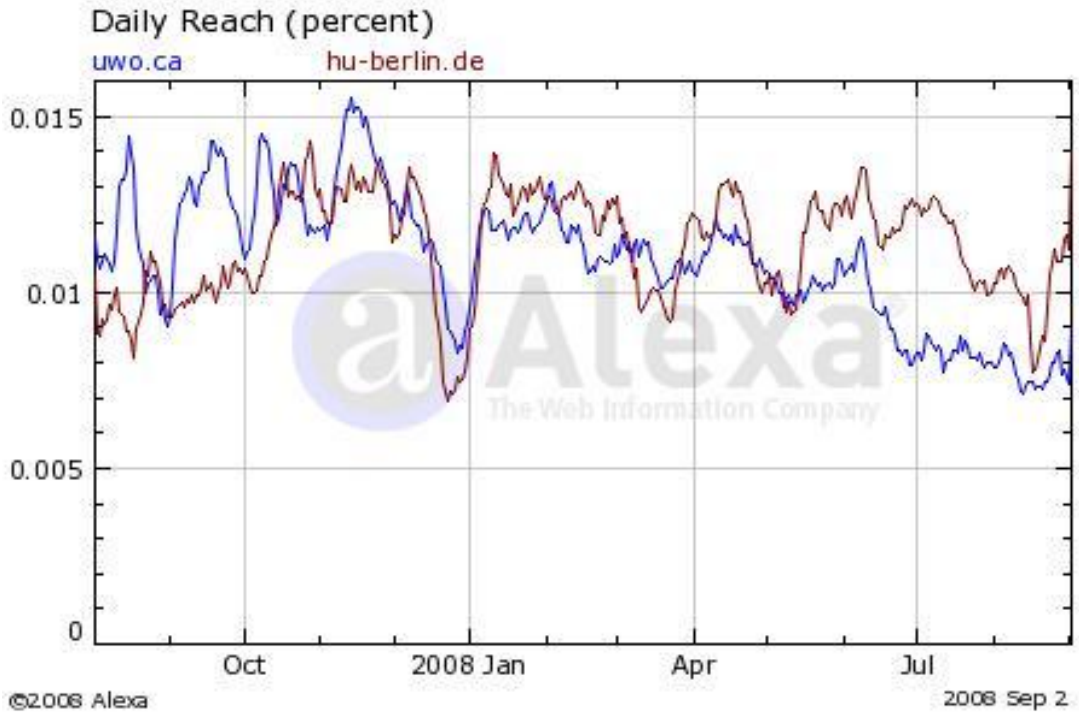

Figure 1 Traffic Comparison of two universities at Alexa

Compete data are collected from over two million U.S. Internet users (Compete, 2008). These users are recruited through multiple sources, including ISPs, the Compete Toolbar and additional opt-in panels. Compete provides various measures of Web traffic such as "Unique Visitors-monthly" and "visits-monthly". These two types of data are available free of charge for the period of the last 12 months. If you want data for a two year period, then a fee is needed. It should be noted that these two metrics are different from Alexa's page view and reach data. Compete does have a "page views-monthly" data, which would be more comparable to Alexa's page view data, but it is not free of charge.

Care must be taken when using these data. One needs to know the specific definition of a metric. For example, Compete's "Unique Visitors-monthly" metric only counts a person once no matter how many times the person

visited a site in a given month. On the other hand, the "visits-monthly" metric measures the number of visits made to a site. A person can only be counted as one person in a month, but can make multiple site visits. Another important issue to be aware of is from whom the data are collected. Alexa collect data from a global user base while Compete data are collected from American users only. This is important to know when compare Websites across countries. Figure 2 shows Compete's comparison of "unique visitors-monthly" between the University of Western Ontario and Humboldt-Universität of Germany, the same two universities compared in Figure 1. The relative positions of the two universities are very different in the two Figures although the comparisons cover the same time period and both measure the number of users. Figure 1 (Alexa data) shows that the two universities are similar but Figure 2 (Compete data) shows that the University of Western Ontario attracted many more users than Humboldt-Universität did. The difference here could be explained by the fact that Compete data came from American users only while Alexa data came from international users. The University of Western Ontario is located in Canada which is geographically much closer to the U.S.A. than the Humboldt-Universität in Germany does.

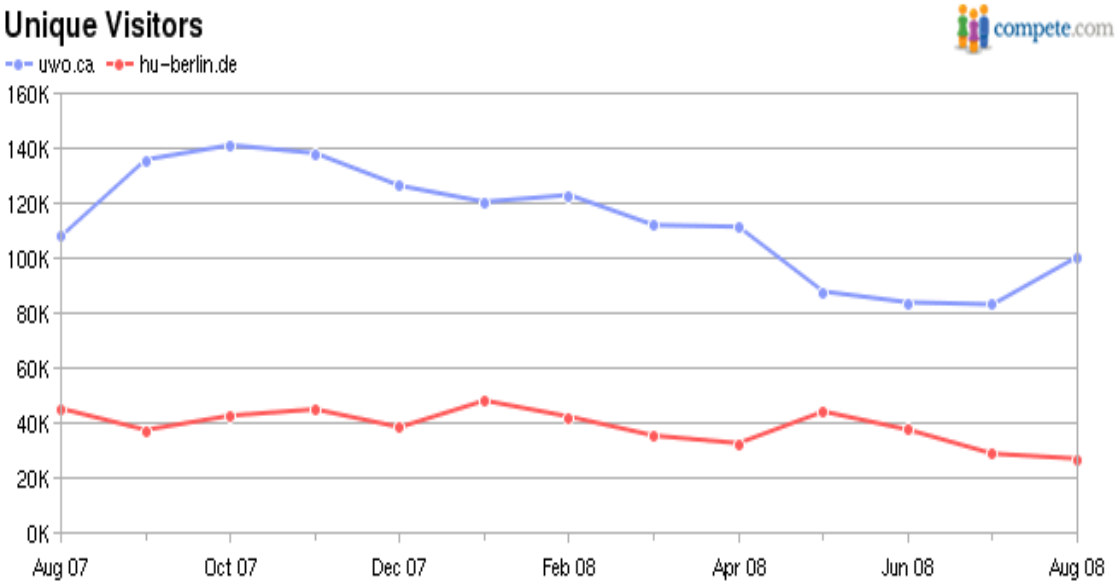

Figure 2 Traffic Comparison of two universities at Compete

The above example illustrates two important issues when using this type of traffic data. We need to be aware of the source of data, i.e. from whom the data are collected. We also need to be clear on specific definitions of the metrics, i.e. how raw data are turned into specific metrics. In short, we need to be aware of the validity and reliability issues of these data. However, this is not to say that these data should be ignored because they may have quality issues. We can examine the data by comparing them with other data or other measurements to find out if the data are chaotic and meaningless or if they contain useful information. Two example of this kind of study are presented in the next section.

5 Examples of studies that used Web usage data

Numerous studies have analyzed Web server logs to find various kinds of information. Silverstein et al (1999) was among the first papers that studied server logs of Web search engines. Later, many studies were carried out using the Excite server logs that were made available to a group of researchers (Spink, Jansen & Ozmultu, 2000; Spink, Wolfram,

Jansen & Saracevic, 2001). They found that users typically used very short queries, rarely employ advanced search techniques and usually viewed only the first 10 to 20 retrieved pages. Ross & Wolfram (2000) further refined the method of query log analysis by examining term pairs rather than individual terms to gain more information on user search interests. In contrasting to these studies that analyzed search engine query logs, other studies analyzed server logs of individual Websites. Nicholas, Huntington, Lievesley & Wasti (2000) examined visits to news Websites while Huntington, Nicholas & Jamali (2008) analyzed site navigation of virtual scholars.

Accessing scholarly journals through electronic media rather than print media is a growing trend. More and more academic libraries provide users with electronic access to journals. Server logs of this kind of access provide opportunities to study journal usage patterns. Duy & Vaughan (2003) compared server log data from vendors of electronic journals with data from a university library proxy server to measure the reliability of the vendor data. Duy & Vaughan (2006) compared the new electronic usage data with more established print usage data and found that the two correlate significantly, which suggests that electronic usage data can be reliable indicators of journal usage. Recently, a very large scale study of usage data, the MESUR project, attempts to define, validate and cross-validate a range of usage-based metrics of scholarly impact (Bollen, Van de Sompel & Rodriguez, 2008). On the other hand, Mayr (2006) proposed a simple quantitative method which establishes indicators by measuring the access and download pattern of open access documents and other Web entities of a single Web server.

From these examples, we can see that Web server logs have been studied for years. Most studies aim to find user interests (what kind of information they are looking for) and user search behavior (how they surf the Web). However, very few studied from an informetrics point of view. Much can be done in this direction. Even fewer are studies that analyze Web traffic to multiple sites. I will briefly describe two studies of this kind that I have carried out to illustrate the potential of Web traffic data. Both studies used Alexa traffic data. The main purpose of the studies was to find out if the Web traffic data provide useful information that can be mined (uncovered). The approach that I took to examine this question was to find out if the traffic data correlate with traditional, more established data that measure the performance of organizations. The logic behind this analysis is that if they correlate, then the traffic data contain useful information in that they measure the performance of an organization. As such, we can tap into this new data source for further information. I analyzed two different types of organizations, academic organization and business (commercial) organizations.

For the study of academic organizations (Vaughan, 2008), I used the list of top 100 universities from Times Higher Education– QS World University Rankings 2007. There are several different university rankings available but the Times–QS ranking is generally considered as one of the most reputable rankings. Another reason to choose the Times–QS ranking is that this ranking is based on traditional measures such as peer reviews and citations to faculty publications. Comparing this ranking with the Aelxa data allows us to explore the relationship between these traditional indicators and the new Web indicators. For universities on this list, I collected data on Web traffic to these university Websites from Alexa. As mentioned above, Alexa provides various kinds of data for a given Website, such as page views and reach. I used the "three month rank" data which is a combined measure of page views and reach over a three month period. The three-month period rather than the one-day period was used in order to cancel out the effect of daily fluctuation. The universities were ranked based on Alexa data and this ranking was compared with Times–QS ranking. A significant correlation between the two was found. In general, universities that are ranked higher by Time-QS attracted more traffic to their Websites. This suggests that Web traffic could be an indicator of the quality of the universities. Data on Web hyperlinks to these universities were also collected. A significant

correlation between to the traffic data and the hyperlink data was found. Since many earlier studies have found that the number of hyperlinks to a university Website can be an indicator of the calibre of the university (Smith & Thelwall, 2002; Thelwall, 2001; Vaughan & Thelwall, 2005), the correlation found here further confirms that the traffic data contain useful information of the quality of the universities.

The study of business organizations analyzed the relationship between Alexa Web traffic data and business performance measures. The top 100 information technology (IT) companies listed in the July 2007 issue of BusinessWeek magazine were chosen as the candidates of the study. The Magazine provides business performance data of these companies. This data source is considered authoritative and reliable. For companies in the study, Alexa's three-month traffic rank data were collected, as was done in the study of academic sites described above. Significant correlations between traffic data and the business performance measures of revenues and net profits were found. Generally speaking, the more traffic to a company's Website, the higher the company's revenue and net profit. The study was presented at the 2008 Annual Conference of Canadian Association for Information Science but no paper has been published yet.

6 Concluding Remarks

The studies reviewed above show that Web usage data do contain useful information but they are under studied in informetrics and Webometrics research relative to other forms of data. Web traffic data are particularly underutilized and yet we can discover information that may otherwise be unavailable to us. For example, business performance measures are often not readily available, especially for private companies. Now we may use traffic to company Websites as an indicator of the company's business performance because the study described above suggested a correlation between the two measures. Of course, more research is needed to further establish the relationship between the two.

It is acknowledged that usage data have their limitations and the reliability of the data is a concern. However, we should not give up on this valuable data source because they have problems. We need to study the data to know what kind of problems they have and how to overcome the difficulties. We can also triangulate data from different sources to examine the reliability. We can use multiple data sources to counter possible bias of a single data source. In short, these large scale data contain rich information. What we need to do is to find ways to extract valuable information from large amount of "noisy" raw data. This is data mining is all about, just as we extract gold from raw ore.

To summarize, Informetric research has the tradition of studying usage data such as library circulation data and library shelving data. Web usage data provides both new opportunities and challenges for us to study. It has great potential as the Web continues to play an increasing role in our society. It is a fertile and yet under explored ground. More studies in this area are needed to gain more knowledge and insights into how these new data sources can complement traditional sources and how we can take advantage of these new data sources. Thus, mining Web usage data is a new frontier of informetirc and Webometric search.

References

Aelxa (2008). *About the Alexa Traffic Rankings*. Retrieved Sept. 5, 2008 from http://www.alexa.com/site/help/traffic_learn_more#page_views.

Bollen, J, Van de Sompel, H. & Rodriguez, M. (2008). Towards usage-based impact metrics: First results from the MESUR project. JCDL '08, June 16–20, 2008, Pittsburgh, Pennsylvania, USA.

Burrell, Q. L. & Fenton, M. R. (1994). A model for library book circulations incorporating loan periods. *Journal of the American Society for*

Information Science, 45(2), 101-116.

Compete (2008). *Where does Compete Ranked List data come from?* Retrieved Sept. 5, 2008 from http://www.compete.com/help/q32.

Decroos, F., Dierckens, K., Pollet, V., Rousseau, R., Tassignon, H. & Verweyen, K. (1997). Spectral methods for detecting periodicity in library circulation data: a case study. *Information Processing & Management*, 33(3), 393-40.

Duy, J. & Vaughan, L. (2006). Can electronic journal usage data replace citation data as a measure of journal use? An empirical examination. *The Journal of Academic Librarianship*, 32(5), 512-517.

Duy, J. & Vaughan, L. (2003). Usage data for electronic resources: A comparison between locally-collected and vendor-provided statistics. *The Journal of Academic Librarianship*, 29(1), 16-22.

Huntington, P., Nicholas, D. & Jamali, H. R. (2008). Site navigation and its impact on content viewed by the virtual scholar: a deep log analysis. *Journal of Information Science*, 34 (1), 598-610.

Ingwersen, P. (1998). The calculation of Web Impact Factors. *Journal of Documentation*, 54(2), 236-243.

Liu, B, Ma, Y., & Yu, P. S. (2003). Discovering business intelligence information by comparing company Web sites. In Zhong, N., Liu, J., & Yao, Y. (Eds) *Web Intelligence*. Berlin: Springer, 105-127.

Mayr, P (2006). Constructing experimental indicators for Open Access documents. *Research Evaluation*, 15(2), 127-132.

Nicholas, D., Huntington, P., Lievesley, N. & Wasti, A. (2000). Evaluating consumer website logs: a case study of The Times/The Sunday Times website. *Journal of Information Science*, 26(6), 399-411.

Ross, N.C.M. & Wolfram, D. (2000). End user searching on the Internet: An analysis of term pair topics submitted to the Excite search engine. *Journal of the American Society for Information Science*, 51(10), 949-958.

Silverstein, C., Henzinger, M., Marais, H. & Moricz, M. (1999). Analysis of a very large Web search engine query Log. *SIGIR Forum*, 33(1), 6-12.

Smith, A. & Thelwall, M. (2002). Web Impact Factors for Australasian universities. *Scientometrics*, 54(3), 363-380.

Spink, A., Jansen, B.J. & Ozmultu, H.C. (2000). Use of query reformulation and relevance feedback by Excite users. *Internet Research: Electronic Networking Applications and Policy*, 10(4), 317-328.

Spink, A., Wolfram, D., Jansen, B. J. & Saracevic, T. (2001). Searching the Web: The public and their queries. *Journal of the American Society for Information Science and Technology*, 52(3), 226-234.

Tague, J. & Ajiferuke, I. (1987). The Markov and the mixed-Poisson models of library circulation compared. *Journal of Documentation*, 43(3), 212-235.

Thelwall, M. (2001). Extracting macroscopic information from Web links. *Journal of the American Society for Information Science and Technology*, 52 (13), 1157-1168.

Thuraisingham, B. (2003). *Web Data Mining*

and Applications in Business Intelligence and Counter-Terrorism. Boca Raton, Florida: CRC Press, p. xvi.

Vaughan, L. (2008). Exploring New Web Data Sources for the Evaluation of Academic Performance. In *Book of Abstracts, 10th International Conference on Science & Technology Indicators*, pp. 217-218, Vienna, Austria, Sept. 17-20, 2008.

Vaughan, L. & Thelwall, M. (2005). A modeling approach to uncover hyperlink patterns: The case of Canadian universities. *Information Processing & Management*, 41(2), 347-359.

Vaughan, L. & You, J. (2008). Content assisted Web co-link analysis for competitive intelligence. To appear in *Scientometrics*, 77(3).

Wikipedia (2008). *Web mining*. Retrieved Sept. 9, 2008 form http://en.wikipedia.org/wiki/Web_mining.