THE USE OF WEB LOG ANALYSIS IN ACADEMIC JOURNALS – CASE STUDY

Azad I. Ali, Indiana University of Pennsylvania, azad.ali@iup.edu
Frederick Kohun, Robert Morris University, Kohun@rmu.edu

==========

ABSTRACT

Web log reports (the reports that are generated to list the statistical data on visitors of web sites) are often used by e-commerce sites to analyze their customer data. The same reports provide valuable information on the patterns of browsing by various visitors of the web site. This information can be analyzed and further studied to use for future decision making and for the benefit of the e-commerce organization in particular and for their web site in particular.

The benefits that can be gained from the web log reports are not limited to e-commerce web sites. Other organizations that conduct regular work on their web sites can benefit from this kind of reporting as well. Among the other organizations that may benefit from web log reports are research organizations that publish papers/journals and post them on their web sites.

This paper is to elaborate on the type of web log reports that are generated by one research institution that publishes a number of academic journals online. The journal editors for this institution receive web log reports where data for visitors of their web sites are presented in various formats. The process of producing these web log reports, their contents and the benefits that the reports may bring to the journals is discussed in this paper.

Keywords: Web Data Mining, Web Log Data Analysis, Web Analysis for Academic Journals, Data analysis for Journal Publications

INTRODUCTION

When hosting web sites, web servers can do more than simply displaying a web page or responding to certain users' requests. Web servers can generate various data about the visitors of the web pages, their geographical locations, the technologies used and other relevant data. These data are termed "web log data" and are saved in files that are then further tabulated and analyzed to generate patterns of visiting the site to aid the web site owners in future decision making regarding the web pages and the content displayed.

Analyzing web log data (or web log analysis) is more popular in e-commerce sites (for profit web sites) than others. In the e-commerce web sites, analyzing web log data may lead to studying patterns of visitors, browsing and purchasing habits. This then may lead to tailoring the design of the web site to fit certain group of visitors or certain other patterns that were discovered from analyzing the data [13].

Web log analysis is not limited to e-commerce sites, instead other organizations may benefit from analyzing data about the visitors of their web sites [12]. Some of the sites that reported benefiting from this type of analysis may include web sites for newsletters, service organizations, as well as other type of web sites.

This study is intended to report on one academic organization that uses web log reports to study the visitors of their web site. The Informing Science Institute (ISI) is an academic institution that publishes journals online. The ISI directors and journal editors receive periodic web log reports and display them to their editors. This paper is to describe the experience of this institution and their use of web log reports in assessing the visitors of their academic journals web sites.

The remainder of this paper is divided into four sections. First, a description of what is meant by web log analysis and the relevant terms is introduced. The second section describes the type of data that is generated from web log reports. Third the paper describes common uses of web log reports in e-commerce web sites as well as other types of web pages. Fourth, this paper elaborates on the work of ISI and their experience of using web log reports. A summary and suggestions for future research is included at the end.

About Web Log Analysis

Analyzing web data is relatively a new field and various terms and tools have been introduced to describe this process of collecting web data,

analyzing it and using it for further business decisions. The tools that are used to complete these reports, although are abundant; they vary and produce different reports. Thus an introduction to some of the terms covered in this process along with some of other related concepts is deemed to be helpful for this paper.

This section introduces some information about web log reporting and analysis. It begins by listing some of the terms that are used in this field along with their definitions. The section then elaborates on the steps that it takes to generate web log report and then to describe how these kinds of tools started to be used.

Terms Used and Definitions

Web log analysis is relatively a newer field. Therefore there is no single standard term that has been established to name it and there may not be one definition that can describe its various components and steps. Instead multiple terms have been introduced and different definitions have been used to describe web log reporting and analysis, their usage and their various components.

Das and; Turkoglu [5] for example used the term "web usage mining" to describe this process and defined it as:

> Web usage mining is to analyze web log files to discover user accessing patterns of web pages. In order to effectively manage and report on a website, it is necessary to get feedback about activity on the web servers. (p. 6635)

On another hand, the web site "Web data mining .inc" [14] used the term "Web data mining" and looked at it from the prospective of crawling through web resources and further provided the following definition for it:

> The term **Web Data Mining** is a technique used to crawl through various web resources to collect required information, which enables an individual or a company to promote business, understanding marketing dynamics, new promotions floating on the Internet, etc. There is a growing trend among companies, organizations and individuals alike to gather information through web data mining to utilize that information in their best interest (p. 1).

Mobasher [9] used two terms to describe the same terms: first, web mining and second web usage mining and further elaborated on both terms:

> Web mining refers to the automatic discovery of interesting and useful patterns from the data associated with the usage, content, and the linkage structure of Web resources. It has quickly become one of the most popular areas in computing and information systems because of its direct applications in e-commerce, e-CRM, Web analytics, information retrieval/filtering, Web personalization, and recommender systems (p. 1).

Buzikashvili [3] described that the goal of researching through web log data is to understand the behavior of the visitors of the site as they browse through pages or search through search engines.

Corsini and Marcelloni [4] noted that a web log file stores the sequence of accesses to web pages that are managed by a web server

Although the definitions are numerous and the terms used are different, however there are some different characteristics that can be derived from these definitions:

- Web log data are abundant and it can be generated easily from the server hosting the web site
- Some of the web data that are generated are useful to the organization while others may not be so useful.
- The main goal of web log analysis is to find pattern of the web site visitors and to estimate their behavior based on their browsing habits.
- The value of the web log data can be increased by studying the relevant data and then derive a pattern of usage of the visitors of the web site.

How it Started

The wide use of web log reports came from the increasing use of web log data, the development of tools that can record this data and the benefits that were realized for studying the data.

Zhong (2008) explained the use and growth of web data mining and web reports into the following points:

- Due to rapid of development and expansion of the Internet, web data has increase.
- A technology needed to extract information from web sources.
- We data mining appeared to fill this gap. It combines together the traditional data mining technology and Web together.
- Web log reports take the data gathered from web data mining and generate reports that explain patterns of visitations to different web pages.

A study conducted by Yen (2004) explained that web data mining appeared because visitors spend "astounding" amount of time in navigating through what termed as "useless" or "redundant" pages. This, the study suggests that this kind of astounding information that is generated created a need to sift through all this data and create a pattern to better serve customer need for information on the web. Web data mining and web log reporting appeared to serve this purpose.

Tao, Hong, and Su [16] explained that data mining in general focuses on the techniques of non-trivial extraction of implicit, previously unknown, and potentially useful information from very large amounts of data. Since the Internet provides this kind of implicit and unknown data, the field of data mining has been expanded to include data on the web as well. As organizations put into use this kind of techniques in web data mining for the Internet data use, more understanding of the data is sought through this kind of reporting.

How it Works

To help realize how web log data is recorded an understanding of web browsing process and how web pages are presented on the browser may be necessary. Thus, the following explains the steps of web browsing and the steps that it follows behind the scene to display particular web pages.

Pabarskaite and Raudys [11] divided the steps it takes from typing a URL and displaying the web age into the following four steps:

1. Typing URL by the user in the web browser.
2. The browser sends a request through the server for the specified page to the internet.
3. If a server for the requested page is found, it sends the response back to the receiving server and then to the browser.
4. The interaction between the two servers (the browsing server and sending server) continues to send and receive pages as the user browse through the different pages.

Some of the steps above may be clear to the user as he/she types URL, browse through the site and observe the response from the browser. However, what goes behind the scene is that the server for the requested page receives additional information about the user. Some of the information is provided by the visitor's server (such as visitor location, technologies used and others) while other information is supplied by the server hosting the page being displayed (such as pages visited, visit duration, frequency of visits and other related data. This kind of data is logged into files that are termed "web log files". As additional visits are conducted by the same (or different) users, more data are added and additional information is recorded. The accumulated data is then tabulated differently and various reports can be produced from the visitors' log. The reports can then analyzed further and used for different purposes for the web site in particular and for the organization that uses the web site in general.

Common Web Log Reports Content

The recorded web log contains large volume of data that examining them in total may not be totally useful. Thus examining the most pertinent and useful information in the weblog data may provide more specific information about the patterns for visitors of the web site. This section lists some items that are

commonly presented in web log reports. It explains each item, the meaning of the item presented in the report and also the benefits that it potentially brings to the web site from understanding and analyzing it.

Number of visits and number of unique visitors

Web log reports can contain two counts regarding visitors of each page: A count for unique visitors and another count for total visitors. The first count is sometimes called "hits" because it refers to the number of times a page was hit or browsed by any visitor of the web page.

Both counts provide useful information to the organization. It is always helpful to have more visitors of the web page and the organization can learn this from the number listed in visitor count. However, it is also helpful to have the same visitor repeat the visit to the site. This data can be obtained by looking at the unique visitors count. Thus studying the combination of both counts (visits and unique visitors) gives a better picture on the popularity of the web pages.

Visits duration and last visits

Both counts combined can give a better picture of the relevant use of the web page.

A short visit to web site may be useful but a general rule of thumb, if a visitor stays browsing pages on the site; the general perception is that the visitor is more interested in the site. However, this does not describe the idle time or the time that the visitor stayed on the site but did not browse, read or otherwise look at the page.

Last visit refers to the date that the page was last visited. A page that has not been visited for extended period of time may indicate a fault in the process of designing the page, navigating to the page or just increasing the ways that the page is searched.

Hosts list

From the outset, this may not seem very useful to many. However, a closer examination may reveal more helpful data. For example, through the host, it can be determined the geographical location of the host, this includes country, state and city in which the viewer of the page is located. This may help in redefining the page for certain population coming from a specific geographical area.

HTTP Errors

This item refers to the errors encountered when attempting access the page. A common practice on web site is to display an error page when someone clicks on a link to display a particular page. If the page is not found, the error page is displayed. If a page is frequently report HTTP error, this may indicate that there may be broken links to the page. Broken link means that when a user click on a link to display a page, the page is not displayed due to error in the link or due to the unavailability of the intended page.

File Types

Files put on the web can have different file types. Example, image files can have types of jpg, gif, bmp or others. Document files can have the type of DOC, PDF, TXT or others. This item of the report can provide a count of the type of the file that was accessed from the page. This count may give a clue for example if more document files are viewed and also specific document file (such as PDF) is viewed.

Browsers Used

While browsers data do not commonly provide useful information, but knowing what browsers are used to display the page may help understand the visits more fully. Sometimes, a visitor can access a site from different browsers, counting these visitors differently may help understand the number of visits more precisely

HTTP referrer

HTTP Referrer allows the server to know where people are visiting them from, and some additional identification. In other words, this item gives information on where the request for this page came from. HTTP referrer is used for statistical purposes in web log reports.

Other Items in Web Log Reports

In addition to the items listed above that are normally presented in web log reports, additional items are listed as well. Below is listing of some of the other items that may be included in a typical web log report:

- Number total page views
- Most viewed, entry and exit pages
- Search engines, key phrases and keywords used to find the analyzed web site
- Some of the log analyzers also report on who's on the site, conversion tracking, visit time and page navigation.
- Authenticated users, and last authenticated visits
- Days of week and rush hours
- Domains/countries of host's visitors
- Operating Systems used
- Robots

Common Uses of Web Log Reports and Analysis

A web site that specializes on the web data mining [14] noted three uses of web log reports: Business Intelligence, predictive analysis, and Internet marketing.

Mobashar [9] on the other hand, noted that web log reports can be used to e-commerce, e-CRM, Web analytics, information retrieval/filtering, Web personalization, and recommender systems.

Kaushik [8] explained that web data mining can used for predictive analysis, traditional decision support systems, large data warehouses, business intelligence systems and tools.

Huang [7] noted that searching web log data may help business owners better understand their customers, improve business operations and make better decisions.

Pabarskaite and Raudys [11] listed the following uses of web data mining

- Restructuring websites
- Improving navigation
- Specialized/Intelligent adverts
- Turning non-customers into customers increasing the profit
- Monitoring efficiency of the web site

Boving and Simonsen [2] used the term Web-based information (WIS) and noted that they are often are often used in distributed organizations to support communication, collaboration, and coordination.

The remainder of this section elaborates on some of the common uses of web log reports in general.

Restructuring Web Sites

Web log reports can be helpful in restructuring and redesigning web sites. Restructuring the design of web pages is more common in e-commerce web sites. In these instances, e-commerce web sites may have data extracted from cookies or from previously entered data and saved into databases. These data in turn can be used to customize the page view according to previously collected data about customers [10].

In regards to customization of web sites, Zhong [17] noted that one of the most beneficial uses of web data mining is for the organizations to be able to make information processing on the web personalized and automatic.

A prime example of this customization and personalization is what is used in Amazon.com web site. A customer logs in, then the viewed page may show previously purchased items and suggests similar item that may attract the user to purchase.

In web log reports, such specific data is not available. Instead the simplest forms, web log reports can indicate the pages that are not visited and study the reasons that it may be turning away visitors to these particular pages. Further study of web log data may reveal some fault in design of the web site, thus owners and web developers can use some of this information in redesigning or restructuring the web site.

Improving Navigation

Navigation among web pages in a particular web site is important. It is essential to make the navigation links between the different pages more usable, visible

and error free. A slow visited page indicated in a web log report may have some causes that make it difficult to visit or see by visitors for the web site [6]. Two items in the web log report may indicate faulty navigation in the web site. HTTP errors and key phrases used to find the pages. If a particular page encounters more HTTP errors, this may indicate a broken link(s) to the page. At the same time, finding the most common key phrase used to reach the page may help predicting what key phrases to use for the search engine as well as within the structure of each of the pages.

Web log analysis in Academic Journals

This section explains about the experience of one academic institution and their use of web log analysis for their web site. The Informing Science Institute (ISI) is a research organization that publishes periodic journals, books, and conference proceedings and presents them on their web site (www.informingscience.org) for public viewing. The following description was extracted from this institute's web site about the mission of the organization:

> The Informing Science Institute (ISI) is an organization of colleagues helping colleagues. We draw together people who teach, research, and use information technologies to inform clients (regardless of academic discipline) to share their knowledge with others. The Informing Science Institute Learning Object Repository (ISI LOR) is the newest Open Source LOR created by a team of give individual. We are currently Beta testing the ISI OSLOR on this site.

The ISI publishes seven periodic journals in online format. Each journal has its own editor and editorial board. The seven journals, although may work semi-independently by having their own editorial board, but all the journals work under the ISI and this organization overall has a managing editor that oversees the operation of all the journals. Below is a list of journals that this institute publishes each year:

- Informing Science: The International Journal of an Emerging Transdiscipline (inform)
- Journal of Information Technology Education (JITE)
- Interdisciplinary Journal of E-Learning and Learning Objects (IJELLO)
- Interdisciplinary Journal of Information, Knowledge, and Management (IJIKM)
- International Journal of Doctoral Studies (IJDS)
- Journal of Information, Information Technology, and Organizations (JIITO)
- Issues in Informing Science and Information Technology Journal (IISIT)

Figure 1 below shows the different journals that are published by the ISI and the relationship between the journals and the main institution (the ISI).

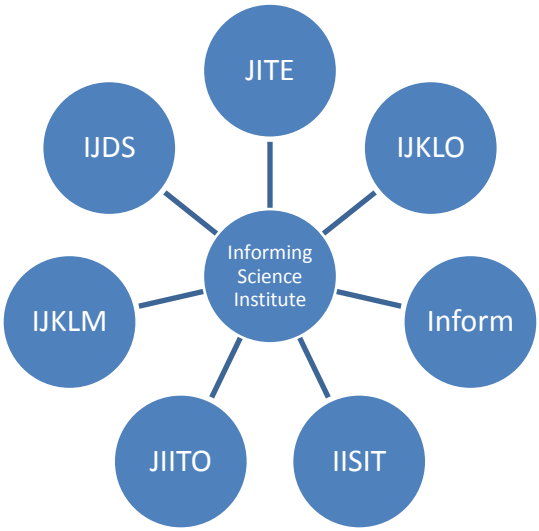

Figure 1 - ISI Journals

Each journal has its own web site but they are all linked from the main ISI web site. The journals publish articles that are saved in Acrobat file format so it will be compatible with online publication formatting. Thus, visitors of the site can view the articles, download them or print them – they are freely available online.

Visitors of the ISI web site can come from anywhere around the world. Due to the multi-national of the members of the ISI, visitors from different countries

around the world often visit and read articles published by the ISI.

Visitors for the ISI web site can visit the site once or can visit it multiple site. Each visit is called a hit. Thus, a visitor can have multiple hits or can have one visit. A better description will be to consider them unique visitors or returning visitors.

Since the papers are published to be viewed and read by visitors to the web site, a measure of the popularity of any published paper is how many hits and visitors it receives. It will help the editors also to learn which countries of origin for visitors of the different articles. Moreover, it will be helpful to the editors to have additional statistical information about the visitors of their journals and the technologies used.

In order to facilitate all this and provide the journal editors data about visitors of their journals, each editor receives a monthly report showing the web log activity for the journals on his/her web site. The report is different from one journal to another based on the request of the editors. But in general, the report is divided into some general categories as well as some customized sections.

The remainder of this section describes some of the categories in the web log reports that are supplied to each editor for the journals that are published by the ISI.

General Statistics

This contains summery information about total hits, visitors, page views and bandwidth. The information is useful to the editors because it provide a general (overall) view of the total visitors to their web pages. The editors, thus can gauge the popularity of the site and make decisions accordingly.

Activity Statistics:

The activity statistics breaks down the total visits by weeks (or by days if needed). It shows charts of the visits for the month. , Some of the information that can be extracted from this category includes hits, page views, visitor count, bandwidth and others. This category can show a chart for the data and can be divided by week, by day or by other specified period, thus it gives a glance of the overall activity for the pages or articles viewed on the site.

Access Statistics

This category provides information on the top other files that were accessed by the visitors. It gives information on the file name, the number of hits/visitors, incomplete requests and bandwidth.

Visitors

This section of the report is probably is the most useful for the journal editors. It shows how many times each PDF file is viewed or downloaded. This indicates the popularity of the paper.

Referrers

This shows the pages that referred to the visited page and it shows what phrase is used to find the page. Each published paper contains a list of keywords that are listed in case of someone searches for a particular topic. The keywords are also used for categorization of the topic that the paper falls under. So knowing the keywords that are most commonly used in the search may help the editors to refine their keywords so to increase the likelihood that their papers receive more visitors.

Browsers

This section shows the browsers used to view the page and the count for each browser. Although this report may not be as useful, but a closer look at the report may reveal the most popular browser used. Customization may be completed to the web site so to gear to move the design of the page more toward one browser or another.

Customized Reports

Different journal editors at the ISI may request customized reports according to what data they consider to be most helpful. Among the most helpful items that are requested is a total count of downloads (or hits) for each article in each country. A reporting of downloads of articles in different countries is requested by the some editors. Sometimes, a combination of both counts can be combined in one report.

Summary and Suggested Future Research

This paper wrote about using web log reports and analysis for web sites that hosts academic journals. It started by writing on web log reports in general, the different terms used in this regard, their definitions. It then discussed the contents that are generated from web log reports and their common uses. Later, the paper focused on the experience of one academic institution – the Informing Science Institute and showed how this organization uses web log report to analyze visitors viewing or downloading articles from their web site.

While this paper introduced the information above in general terms, much focus was not placed on what can be done to get more use of web log reports and to analyze them further. Instead, the paper listed how these reports are currently used but stopped short of suggesting more use of these reports. The plan is to conduct a second study where the focus of the future paper will be on how to present more useful information from these weblog reports so it can help enhancing the presentation of journals on the ISI web site.

REFERENCES

1. Berry, M. & Linoff, G. (2000) *Mastering Data Mining*. New York: Wiley Pulishing.
2. Bøving, K., Billeskov; Simonsen, Jesper (2004). HTTP Log Analysis. *Scandinavian Journal of Information* Systems, 16, 145-174.
3. Buzikashvili, Nikolai (2007). Sliding window technique for the web log analysis. *Proceedings of the 16th international conference on World Wide Web*, 1213-1214. Retrieved May 13, 2010, from ACM Digital Library http://www.acm.org/dl..
4. Corsini, Polo; Marcelloni, Francesco (2006). A Fuzzy System For Profiling Web Portal Users From Web Access Log. *Journal of Intelligent & Fuzzy Systems*, 17 503–516.
5. Das, Resul; Turkoglu, Ibrahim (2009). Creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method. Expert Systems with Applications; Apr2009 Part 2, Vol. 36 Issue 3, p6635-6644, 10p.
6. Fang, X.; Holsapple, C. W. (2007). An empirical study of web site navigation structures' impacts on web site usability. *Decision Support Systems*, 43(2), 476-491.
7. Huang, X (2007). Comparison of Interestingness Measures for Web Usage Mining: An Empirical Study. *International Journal of Information Technology & Decision Making*. 6 (1), 15–41.
8. Kaushik, A. (2007). Data Mining and Predictive Analytics On Web Data Works? Nyet!. Retrieved May 10, 2010 from http://www.kaushik.net/avinash/2007/09
9. Mobasher, B. (2010). Web Data Mining. Retrieved May 10, 2010 from http://facweb.cs.depaul.edu/mobasher/classes/ect584/syllabus.html.
10. Mulvenna, M. D. ; Anand, S., S. & Buchner, A., G. Personalization on the net using web mining, *Communications of the ACM*, 43(8) (August 2000) 122–125.
11. Pabarskaite, Zidrina; Raudys, Aistis (2007). A Process of Knowledge Discovery from Web Log Data: Systematization and Critical Review. *Journal of Intelligent Information Systems*, 28 (1), 79-104.
12. Rubin, Jeffrey H. (2004). Log Analysis Pays Off.(weblogs). *Network Computing*, 15 (18), 76-78.
13. Velayathan, Ganesan; Yamada, Seiji (2006). Behavior based web page evaluation. *Proceedings of the 16th international conference on World Wide Web*, 1317 – 1318. Retrieved May 13, 2010, from ACM Digital Library http://www.acm.org/dl.
14. Web Data Mining .net. (2010). *Web Data Mining.* Retrieved May 10, 2010 from http://www.web-datamining.net/.
15. Yen, Benjamin (2004). Structure-Based Analysis of Web Sites. *Proceedings of the 2004 IEEE International Conference on e-Technology, e-Commerce and e-Service* (EEE'04).
16. Tao, Y.; Hong, T.; & Su, Y. (2008). Web usage mining with intentional browsing data. *Expert Systems with Applications*, 34(3), 1893-1904. Retrieved May 14, 2010 from Science Direct Digital Library
17. Zhong, S. (2008). Information Intelligent System based on Web Data Mining. *Proceedings of the 2008 International Symposium on Electronic Commerce and Security*, 514-517. Retrieved May 14, 2010 from IEEE Computer Society Digital Library.