Analysis of Web Server Log Files: Website of Information Management Department of Hacettepe University

Web Sunucusu Log Dosyaları Analizi: Hacettepe Üniversitesi Bilgi ve Belge Yönetimi Bölümü Web Sitesi

Mandana Mir Moftakhari*

Abstract

Over the last decade, the importance of analysing information management systems logs has grown, because it has proved that results of the analysing log data can help developing in information system design, interface and architecture of websites. Log file analysis is one of the best ways in order to understand information-searching process of online searchers, users' needs, interests, knowledge, and prejudices. The utilization of data collected in transaction logs of web search engines helps designers, researchers and web site managers to find complex interactions of users' goals and behaviours to increase efficiency and effectiveness of websites. Before starting any analysis it should be observed that the log file of the web site contain enough information, otherwise analyser wouldn't be able to create complete report. In this study we evaluate the website of Information Management Department of Hacettepe University by analysing the server log files. Results show that there is not adequate amount of information in log files which are provided by web site server. The reports which we have created have some information about users' behaviour and need but they are not sufficient for taking ideal decisions about contents & hyperlink structure of website. It also provides that creating an extended log file is essential for the website. Finally we believe that results can be helpful to improve, redesign and create better website.

Keywords: *Log file; users' needs; website of Information Management Department of Hacettepe University.*

Öz

Son on yılda, bilgi yönetim sistemleri analizinde önemli artış gözlenmektedir. çünkü log dosya analizinin bilgi sistemleri tasarımı arayüz ve web sistemlerinin oluşumuna neden olduğu ıspatlanmıştır. Online araştırmacıların bilgi arama süreçlerini ve kullanıcıların aradıklarını, gereksinimlerini, ilgi alanlarını, bilgilerini ve önyargılarını anlamak için log dosyası analizi en iyi yollardan biridir. Web arama motorlarının işlem loglarında depolanan verilerin kullanımı, kullanıcıların karmaşık etkileşimleri, amaçları ve davranışlarının araştırılmasında, web tasarımcılara, araştırmacılara ve web site yöneticilerine yardımcı olarak web sitelerinin verimi ve etkinliğini artırmaktadır. Herhangi bir analize başlamadan önce araştırılmak istenen web sitesi log dosyasının yeterli bilgi içerdiğine dikkat edilmesi gerekmektedir. Aksi takdirde analizörün tam ve kapsamlı bir rapor oluşturması mümkün görünmemektedir. Bu çalışmada sunucu log dosyaları analiz edilerek Hacettepe Üniversitesi Bilgi ve Belge Yönetimi Bölümü web sitesi değerlendirilmiştir. Sonuçlar, web sitesi sunucusu tarafından sağlanan log dosyalarındaki bilgilerin yeterli miktarda olmadığını göstermektedir. Oluşturduğumuz raporlar

* Librarian, Bilkent University Library. e-mail: mandana.mir@bilkent.edu.tr

kullanıcıların davranışları ve gereksinimleri ile ilgili bazı bilgiler içermekte olup, web sitelerinin içeriği ve bağlantıları ile ilgili ideal kararların alınabilmesi için yeterli değildir. Raporlar web sitesi için genişletilmiş log dosyasının oluşturulmasının gerekli olduğunu da göstermektedir. Gelinen noktada araştırma sonuçlarının web sitesinin yeniden tasarımına, geliştirilmesine ve daha iyi bir web sitesi oluşturulmasına yardımcı olacağına inanmaktayız.

***Anahtar Sözcükler**: Log dosyası; kullanıcıların gereksinimleri; Hacettepe Üniversitesi Bilgi ve Belge Yönetimi Bölümü web sitesi.*

Introduction

Today possessing a well organised website is one of the vital goals of any organization. A precise understanding of what users are like, why they use the website and how they might interact with it, is the most critical key for successful design of interactive systems.

According to Tidwell (2006) the users' goals, the specific tasks undertaken by users, the language or terminology used by users and the users' experience and skills are important information that a web designer should learn. Website user's activities and the interaction between the user and an information access system can be collected in server logs. This information can then be used to analyse. In order to discover the users' needs and to understand system requirements there are different methods and techniques like direct observation, interviews, surveys, personas and focus groups.

One of the most effective methods to evaluate the usability and effectiveness of a system is web log analysis. Web log analytics are used widely in different areas such as tourism, medical, health, economy, politics, marketing, management and education by researches (Croft, Cook, & Wilder, 1995; Jansen, Spink, & Saracevic, 2000; Jones, Cunningham, & McNab, 1998; Wang, Berry, & Yang, 2003). Logging the user interaction provides large amount of data for analysing the patterns of interface usage, frequency of requests, speed of user performance, or rate of errors (Shneiderman & Plaisant, 2005). Extracting user behaviour, extracting relevance of web communities, analysis of search keywords and visualization of access logs are some hot topics where log analysis can be used.

In addition it should be mentioned that effective work on log analysis requires clean and well-defined log records. This means that analyst can create complete report if only when the transaction log data which have been collected and prepared is complex enough to cover every possible aspects of the system for the analysis (Agosti, Crivellari, & Di Nunzio, 2012).

In this paper web log analyser program is used to analyse the server logs of Information Management Department of Hacettepe University to get general statistics about hits, unique IPs, downloaded files, unique files, unique authors, visits per month, visits per hour per unique visitor, words in each search query, downloaded files created by a unique author and downloaded of a unique file.

Literature

A significant example of log analysis is a research study into online public access catalogues (OPACs) which was conducted by office of the Online Computer Library Center (OCLC) at beginning of the 1980s (Tolle 1983). From 1981 to 1983, OCLC implemented log analysis to determine to what extent current system features were used. Yu and Apps (2000) examined user behaviour in the Super Journal project by using transaction log data too. The researchers recorded 102,966 logged actions during 23 months from February 1997 up to December 1998. They related these actions to four subject clusters, 49 journals, 838 journal issues, 15,786 articles, and three Web search engines. Jansen and Pooch (2001) and Hsieh-Yee (2001) reviewed studies about web transaction log research that belongs to Web search engines and individual Web sites. They claim that most of the studies conducted between 1995 and 2000 examine and evaluate the effects of particular factors on search behaviour, involving information organization, kind of search task, Web experience, cognitive abilities, and emotional states. Wang, Berry and Yang (2003) and Spink (2004) describe approaches to transaction log analysis. Bar-Ilan (2004) evaluated the usage of web search engines in information science research by overviewing the search engines and user needs.

Suneetha and Krishnamoorthi (2009) have done analysis of web log data of NASA website to find information about a web site, important errors and potential visitors of the site. They claim that the obtained results of the study can help system managers and web designer to improve and increase effectiveness of NASA's system. Grace, Maheswari and Nagamalai (2011) give a detailed discussion about log files, their formats, their creation, access procedures, their uses, various algorithms used and the additional parameters that can be used in the log files which in turn gives way to an effective mining. They also provide the idea of creating an extended log file and learning the user behaviour. Goel and Jha (2013) suggested a log analyser tool which was called web log expert and used for determining the behaviour of users who access an astrology website. Deepti and Shweta (2014) have an overview of web usage mining. Also they offer some methods to detect users' behaviour from web log files.

Log File

As Grace, Maheswari and Nagamalai (2011) mentioned in their paper log files are files that save any transaction that has occurred between a client browser and a web server during a search. Log files keep all information about user name, IP address, time stamp, access request, number of bytes transferred, result status, URL that referred and user agent.

Rice and Borgman (1983) claim that transaction logs are data collection tool that automatically keeps the type, content, or time of users' transactions. Also Jansen (2006) describe log file as "an electronic record of interactions that have occurred during a searching episode between a Web search engine and users searching for information on that Web search engine" (Jansen, 2006, p.408).

According to Ratnesh, Kasana and Suresh (July 2009) the log files contain different

types of information and the basic information available in the log file are:

" • User name that identifies who had visited the site. Most of the time the identification of the user is the IP address that is assigned by the Internet Service provider (ISP).

• Visiting path that is taken by the user while visiting the web for searching.

• Path traversed that is taken by the user while visiting the web by using the various links.

• Time stamp is the time spent by the user in each web page.

• Page last visited which was visited by the user before leaving the web site.

• Success rate that can be determined by the number of downloads made and the number copying activity undergone by the user.

• User agent presents the browser from where the user sends the request to the web server.

• URL that reveals resource accessed by the user request type presents the method which is used for information transfer."

These are the contents which are kept in the log file and researchers analyse these information in order to learn users' needs and behaviour and to increase the effectiveness of web site.

Web Log Analyse

Web analytics are used widely by commercial, educational, political and health organizations to evaluate the effectiveness of their web sites so website design and management have become critical issue in web based applications. In the literature web analytics also named as web metrics, web log analysis, and web statistics is used to track, collect, measure, report and analysis data in order to optimize web sites (Kaushik, 2007).

Jansen (2006) claims that Web log analysis involves the following three major stages: collection, preparation and analysis. In collection step the interaction data is collected in a transaction log then the process of cleaning data is started and the last step data will be analysed.

Methodology and Purpose of Research

AWStats is a freely available open source application for log file analysis offering a graphical display to the data in web server log files. This study analyses web usage statistics for the Information Management Department of Hacettepe University Web site for a two and half year period, from 15th May 2012 through 15th October 2014. The study starts with May 2012 because that is the first month the log files were saved. Web site was available at the current address: http://www.bby.hacettepe.edu.tr/yayinlar3.asp. The period of 30 months was selected to give enough data and time to identify trends in the site usage. Data from the AWStats reports were entered into an Excel spread sheet and graphed and sorted in order to show patterns of the use during the time.

Research Questions

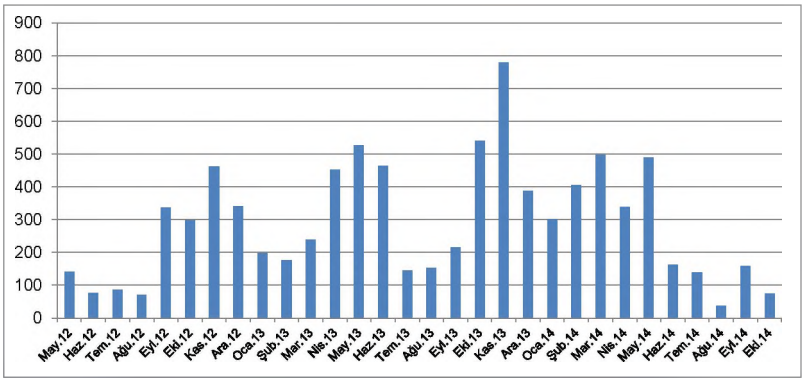
This case study will analyse the web site logs from the Information Management Department of Hacettepe University Web site in an attempt to answer the following questions:

- How many hits does the web site receive?
- How often the website was used (by month and hour)?
- How many words were used in each search query?
- How many of searches were from a unique visitor?
- How many times the files of a unique writer or author were downloaded?
- How many times a unique file was downloaded?

Findings

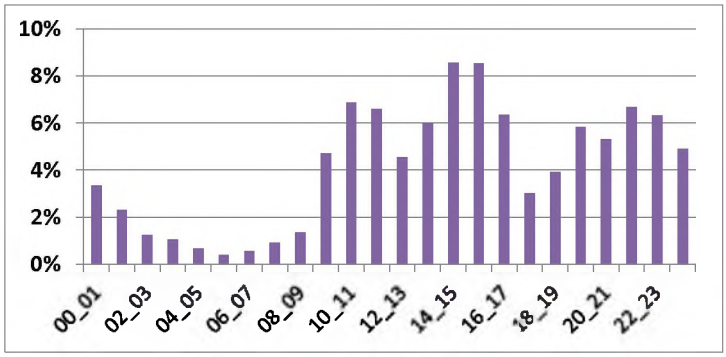
We observed a total of 8697 records. The Information Management Department of Hacettepe University Web site is not heavily used. Over the survey period, the site had an average of 290 visits per month. The number of repeat visits is 6308. Total number of hits is 8697. Total number of unique IPs is 2661. Total number of downloaded files is 1389. Total number of unique files is 338 and total number of unique authors is 103.

As we can see in figure1. total number of visits during these months: (May, June, July, August, September, October) for each year is follows: 1009 in 2012, 2046 in 2013 and 1062 in 2014.



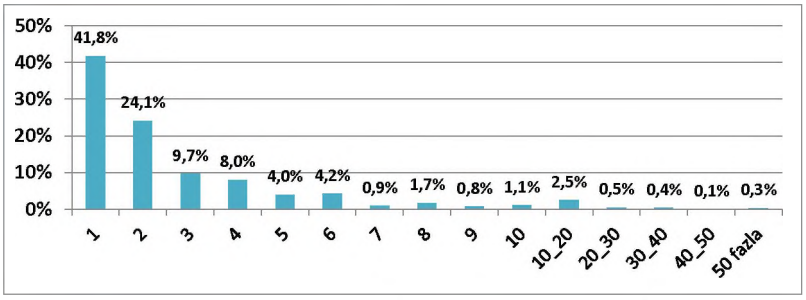
(Figure 1): Total number of visits per Month

As we can see the number of visitors had an increase from 2012 to 2013 but it had a decrease in 2014.



(Figure 2): Total number of visits per Hour %

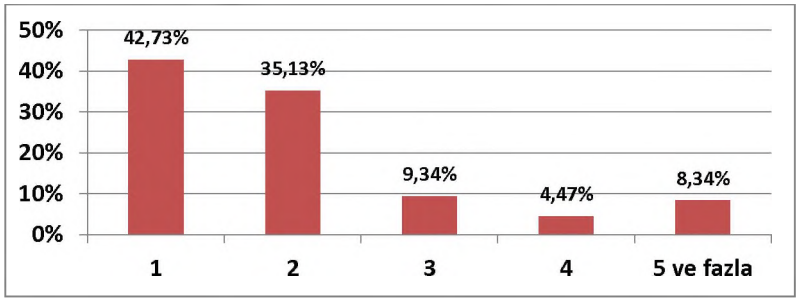
The hours of heaviest use were during 10 am-17 pm and the peak of visits was between 2-4 pm showing that users use the site during traditional work hour (see Figure 2).



(Figure 3): Total number of searches by a unique visitor %

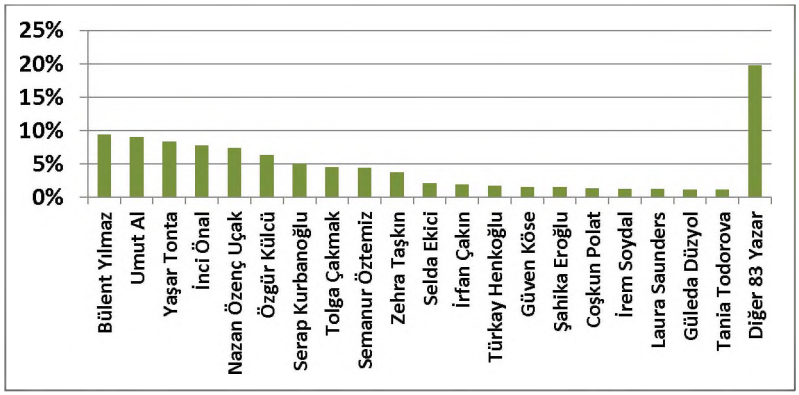
The average number of visits per visitor was 3.37. Also 42% users have visited the page only one time during the period of about 2.5 years. This shows 42% users are not returning to the web site again (see Figure 2).

To tell why this is so, further research is needed to answer the question. But this could represent a problem, indicating that perhaps the web site is not able to provide the information needs of its users, and it is causing them to not return.



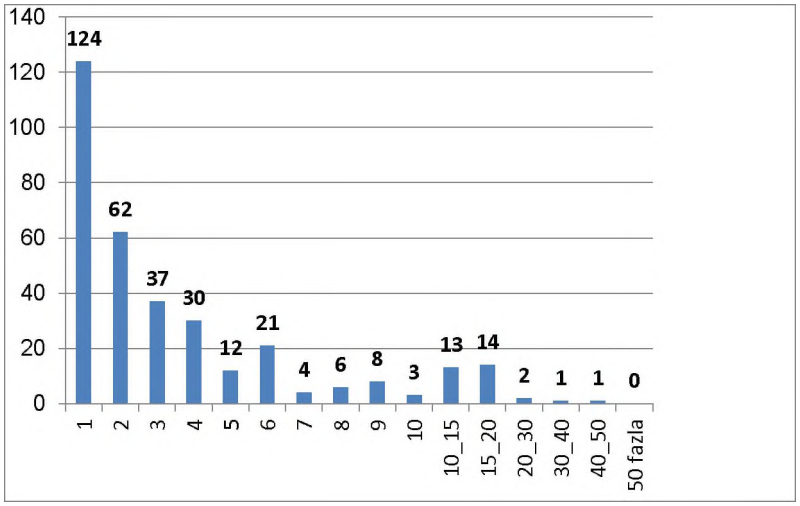
(Figure 4): Total number of words in each search query %

78% of searches were done by using 1-2 key words. It shows these users could find their files by using few keywords.



(Figure 5): The number of downloaded files created by a unique author%

80% of downloaded files belonged to 20 authors and remaining 20% belonged to 83 authors.



(Figure 6): The number of downloading of a unique file

There are 124 files which were downloaded only once.

There are 31 files which were downloaded more than 10 times

Conclusion

During the last 10 years log analysis has raised important points of discussions and has become an integral part of many organizations' operations. There are many researches which are down by using web log analysis tools in order to provide lots of technical information regarding server load, unusual activity, or unsuccessful requests. Researchers believe that acquired results from log analysing can help website maintainers, website analysts, website designers and developers to increase the quality of their system by determining occurred errors, corrupted or broken links. By analysing the *website of Information Management* Department *of Hacettepe University we revealed these results:*

- There is no regular increase in the number of users during the years.
- The hours of heaviest use were during 10 am-17 pm and the peak of visits was between 2-4 pm showing that users use the site during traditional work hours.
- 42% users have visited the page only one time during a period of 2.5 years.
- The average number of visits per visitor was 3.37. Also 42% users have visited the page only one time during a period of about 2.5 years. This shows that 42% users are not returning to the web site again. It can be concluded that the web site could not provide the information wants and needs of its users so it is causing them to not return.
- 78% of searches were done by using 1-2 key words. It shows on these web site these users could find their files by using few keywords.
- 80% of downloaded files belonged to 20 authors and remaining 20% belonged to 83 authors.
- There are 124 files which were downloaded only once.
- There are 31 files which were downloaded more than 10 times.

We believe that the results of this work help increase the of website's effectiveness.

Recommendations

- Log file should save more details about users' searches and downloads. Such as:
  - Kind of resources they prefer more to download (book, article...).
  - Years they prefer more.
  - Kind of key word they use more for searching (author or the name of resources or the year).
- The web page should be redesigned.
- Instructors should encourage students to use the page.
- English page should be designed.

References

Agosti, M., Crivellari, F., & Di Nunzio, G. M. (2012). Web log analysis: a review of a decade of studies about information acquisition, inspection and interpretation of user interaction. *Data Mining and Knowledge Discovery*, 24(3), 663-696.

Bar-Ilan, J. (2004). The use of web search engines in information science research. In B. Cronin (Ed.), *Annual review of information science and technology* (Vol. 33, pp. 231-288). Medford, NY, USA: Information Today.

Croft, W., Cook, R., & Wilder, D. (1995, June). Providing government information on the Internet: Experiences with THOMAS. Paper presented at the Digital Libraries Conference, TX: Austin.

Deepti, s. & Shweta, M., (2014). Detecting users behavior from web access logs with automated log analyser tool, *International Journal of Computer Science and Information Technologies*, 5 (4) , 5106-5109

Goel, N., jha, c.k.(2013). Analyzing users behavior from web access logs using automated log analyzer tool. *International Journal of Computer Applications,* 62(2), 29-33.

Grace, L. K., Maheswari, V., & Nagamalai, D. (2011). Analysis of web logs and web user in web mining. *International Journal of Network Security & Its Applications (IJNSA)*, 3(1), 99-110.

Hsieh-Yee, I. (2001). Research on web search behavior. *Library & Information Science Research,* 23(1), 168-185.

Jansen, B. J., (2006) .Search log analysis: What it is, what's been done, how to do it. *Library & Information Science Research,* 28, 407–432.

Jansen, B. J., & Pooch, U. (2001). Web user studies: A review and framework for future work. *Journal of the American Society of Information Science and Technology*, 52(3), 235-246.

Jones, S., Cunningham, S., & McNab, R. (1998, June).Usage analysis of a digital library. Paper presented at the Third ACM Conference on Digital Libraries, Pittsburgh, PA.

Kaushik, Avinash. (2007). *Web analytics: An hour a day*. Indianapolis: Wiley Publishing.

Ratnesh K. J., Kasana R. S. and Suresh, J. (July 2009). Efficient web log mining using doubly linked tree, *International Journal of Computer Science and Information Security, IJCSIS*, 3(1).

Rice, R. E., & Borgman, C. L. (1983). The use of computer-monitored data in information science.*Journal of the American Society for Information Science,* 34(4), 247–256.

Shneiderman, B., & Plaisant, C. (2005). *Designing the User Interface*. 4 th edition. ed: Pearson Addison Wesley, USA.

Spink, A. (2004). Multitasking information behaviour and information task switching: An exploratory study. *Journal of Documentation*, 60(3), 336-345.

Suneetha, K. R., & Krishnamoorthi, R. (2009). Identifying user behavior by analyzing web server access log file. *International Journal of Computer Science and Network Security*, 9(4), 327-332.

Tidwell, J. (2006). *Designing interfaces*. Sebastopol, CA: O'Reilly.

Tolle, J., (1983.) Transactional log analysis: Online catalogs. In: Kuehn JJ (ed) *Proceedings of the 6th annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR'83*. ACM, New York: Association for Computing Machinery, pp 147–160.

Wang, P., Berry, M., & Yang, Y. (2003). Mining longitudinal web queries: Trends and patterns. *Journal of the American Society for Information Science and Technology*, 54(8), 743-758.

Yu, L., & Apps, A. (2000). Studying e-journal userbehavior using log files: The experience of superjournal. *Library & Information Science Research*, 22(3), 311–338.