A Web Application Path Analysis through Server Logs

Sevgi Koyuncu Tunç and Özgür Külcü
Department of Information Management, Hacettepe University, Beytepe, Ankara, Turkey

Keywords: Web Applications Usability, Server Log Analysis, Path Analysis, Human Computer Interaction.

Abstract: Web server logs contain certain information about accessing web pages. When this information is analysed using data mining methods, information about users' web page visiting habit, frequency and behaviour can be obtained. This allows us to improve service quality by measuring the usability and performance of web applications. The most important advantage of the remote usability evaluation method with server logs is that it is implemented without any communication with the user. While the user is using the system in a routine and natural environment, it is possible to access real user activity information with recorded logs. Users visit the web application one or more times. The web page calls during each visit bring up a session. User sessions must be distinguished to determine the route that users follow when using a web application. For this purpose, log records were processed through the following steps; Data Cleaning, Page Address Coding, Session Detection and Specification of all URL sequences in this research. There are 3 methods in the literature to distinguish the sessions from each other. The start of the sessions can be determined according to the total duration of the session, the time spent on the page, and the page source (Spiliopoulou et al., 2003). This research was based on total session duration, and the page calls made by the user in 10 minutes and 15 minutes were considered as a session. Hacettepe University Electronic Document Management System's (EDMS) 16 million access logs were pre - processed and analysed to obtain system usage data. The research revealed weaknesses of the EDMS web pages in terms of usability via path frequency analysis.

1 INTRODUCTION

Usability is the degree to which a software can be used by specified consumers to achieve quantified objectives with effectiveness, efficiency and satisfaction in a quantified context of use. Log analysis, a relatively new method for testing the usability of web applications is based on the analysis of users' server requests during web application usage. In this method that is called "remote usability evaluation", users and analyzers are separated by time and space. This makes it possible to analyze user's behavior in their natural, everyday environment, while removing the cost of bringing them to private laboratories.

A certain amount of server logs can be used to gather quantitative usage information. When logs are processed and interpreted, technical information such as usage intensity, server load, abnormal activity, failed client calls can be obtained at specific time intervals. However there are a number of studies that show that web server logs can be derived from not only statistical but also route-based inferences. By analyzing recurring patterns, it is possible to evaluate the usability and user experience, and for different types of software, it is possible to identify the interface problems by revealing the repetitive patterns that users make while using the system. (Kim et al., 2013) In this server path analysis study, it is aimed to determine the route that users use most when using Hacettepe University Electronic Document Management System to unreveal the problematic user interfaces.

2 LITERATURE REVIEW

Wang and Lee (2011) has proposed a compact graphical structure to record the way users navigate in the website. Munk and Kapusta (2010) has recommended improving the web portal based on hours of access to a web portal. Resul Das and Turkoglu (2009) aimed to investigate URL information related to accessing electronic resources using route analysis technique by pre-processing server log records.

Tunç, S. and Külcü, Ö.
A Web Application Path Analysis through Server Logs.
DOI: 10.5220/0007231804270430
In *Proceedings of the 10th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2018) - Volume 1: KDIR*, pages 427-430
ISBN: 978-989-758-330-8

Poggi and Moren (2008) has developed benefit-based web session management that can adapt itself using machine learning and markov-chain techniques. Liu and Keselj (2007) proposed a model that automatically classifies user navigation and predicts the future navigation of the user.

Even if the linking rules are not directly linked, the visited pages are used for exploring and revealing the common information of different groups with this information. In addition, it is possible to improve the web application by adding links between pages which are not related to each other, but which are shared by users (Chen et al., 1996, Punin et al., 2001, Batista et al., 2002, Zaiane et al., 1998).

Finding Web pages accessed one after the other makes it easier to develop users' activities, trends, and predictions for the next page to visit. Tree-like topology patterns and frequent route transitions have been investigated by Chen et al., 1996; Lin et al., 1999; Nanopoulos and Manolopoulos, 2000.

Borges and Levene (1999) use a probabilistic grammar-based approach, a Ngram model to capture user navigation behavioral patterns. The Ngram model assumes that the last N page to be browsed affects the probability of identifying the next page to be visited.

Jin et al., (2004) used Probabilistic Hidden Semantic Analysis (PLSA) to explore navigation traces. PLSA can be used to identify hidden semantic relationships between users and between Web pages and users.

Jespersen et al., (2003) Markov assumptions are used to investigate the structure of web application use.

3 RESEARCH

In this study, the 29 day 16 million access logs recorded in the Hacettepe University EDMS application server have been analyzed. The total log size is 12 GB.

For log records to be used for route analysis, they must go through the following preliminary steps. These are;

- Data Cleaning
- Page Address Coding
- Session Detection

3.1 Data Cleaning

When a web page call is made, the page's jpg, png, css files are also called. Since these additional files belong to the same page, they must be cleared before the route analysis (Liu and Keselj, 2007).

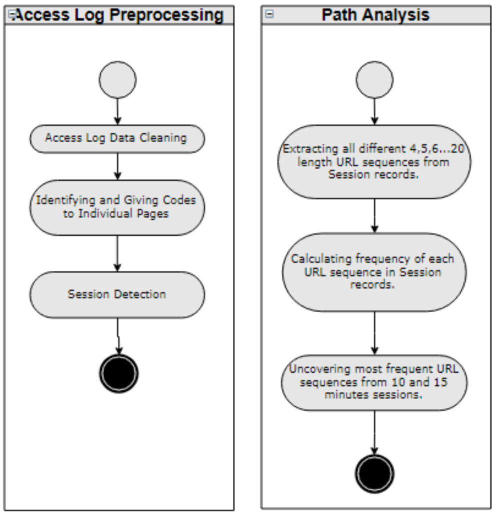
Access Log Preprocessing
Access Log Data Cleaning
Identifying and Giving Codes to Individual Pages
Session Detection
Path Analysis
Extracting all different 4,5,6...20 length URL sequences from Session records.
Calculating frequency of each URL sequence in Session records.
Uncovering most frequent URL sequences from 10 and 15 minutes sessions.

Figure 1: Research steps.

Lines that contain image, JavaScript, and style file invocations that do not point to any page in the access log records have been deleted using SQL text processing functions.

3.2 Page Address Coding

Individual page addresses are given letter codes to increase readability of URL sequences in a separate URL table.

3.3 Session Detection

Users visit the web application one or more times. The web page calls during each visit bring up a session. User sessions must be distinguished to determine the path that users follow when using a web application. There are 3 methods in the literature to distinguish the sessions from each other. The start of the sessions can be determined according to the total duration of the session, the time spent on the page, and the page source (Spiliopoulou et al., 2003). This research was based on total session duration, and the page calls made by the user in 10 minutes and 15 minutes were considered as a session. For example, if the time of the first page call for the S1 session is t0 and the time for the second page call is t1, then t1- t0 <= 10 must be present for the two-page calls in S1 (Spiliopoulou et al., 2003).

A desktop program in c# language developed for session detection. With this program, the ip-based session list and the URL sequence that accessed during each session are specified for the desired time interval (n-minute sessions). A data table named Session is designed to save sessions. The program follows these steps for each log record iteratively:

- If there is not a session record with the same IP with the i.th record in the Session Table, within a maximum of n minutes (n = defined session duration: 10 minutes / 15 minutes); a new session record is created. The page code is written in the "URL sequence" field.
- If there is a session record with the same IP with the i.th record in the Session Table, within a maximum of n minutes; the accessed page code is added to the URL sequence of the found session record.

After pre-process of access log data, path analysis has been implemented on session records as following.

3.4 Identification of URL Sequences

Once the sessions and URL sequences have been defined, a program has been developed to determine the paths that users use most frequently during their sessions while using the EDMS. This program performs the following operations:

3.4.1 Specify All Variety of URL Sequences

All different sequences of 4, 5, 6 .... 20 length URL sequences in session table entries were extracted.

3.4.2 Determine the Frequency of URLs Sequences

In all sessions, the total number of occurrences of URL sequence strings was counted and the URL sequence string and usage frequency data were recorded in the PATH_COUNT table.

3.4.3 Identifying the Most Common URL Sequences

On the Path_Count table, SQL queries that calculate the usage frequency statistics of URL sequences have been developed, and the most common routes have been determined by running these SQL queries.

4 RESULTS

In both the 10-minute and 15-minute session lists, the program's home page, the landing page "incoming documents page" and the page used to view and prepare correspondence appears to be the most visited consecutive pages.

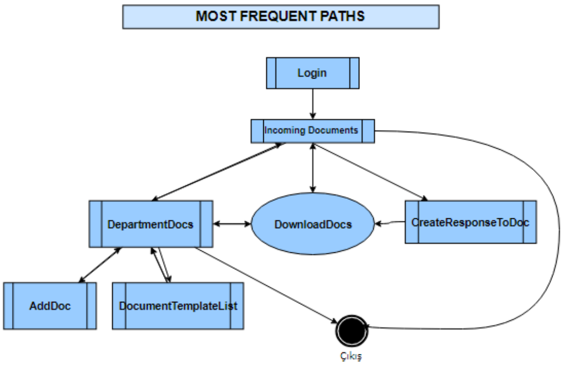

Figure 2: Most frequent paths of EDMS web application.

According to the path analysis the correspondence page is often followed by a "document download" call. The fact that the text of the correspondence cannot be previewed without downloading can be shown as the cause.

Another significant trend in most of the sessions is that some certain page calls are always followed by another specific page call. So, merging these two pages in one page may increase the efficiency of the system usage. For example, in folder merging process, user is being directed to a second page to choose folders for each folder and it means pages posted and loaded extra 2 times for a file merge process.

Documents Search screens appear to be not popular. It has been found that users query the correspondence list instead of the document search pages and download the document. The lack of frequent calls to Document Lookup and Detailed Document Lookup screens may be because these screens are not useful.

5 LIMITATIONS

Server log analysis gives some general information about the system usage but cannot show specific usability problems. Because of the caching mechanism it is not possible to handle user path for completing a task or we cannot know which page elements are problematic to the user by using only server logs.

Server log analysis may also show less users than the site had because multiple users could potentially use the same IP address. The data on entry and exit points will be unreliable. Users will look like they've left when they've simply gotten a new IP address. Their next page view will make look like a new user. The data on length of visit and time on each page will also be unreliable.

Another use of server logs and route analysis work is to capture tips on usability and improve the system by obtaining information about whether a process consisting of several steps is performed by following the route expected by the users, the most time-consuming process steps and the steps in which the process is left most. In the Hacettepe EDMS program, the process of "preparing correspondence" is a process consisting of inputting general information, writing text input, adding text input, and signing. But all the steps are done by one-page URL. For this reason, information such as the user's movement between steps, the amount of time spent in each step, and the step at which he left the process is not obtained from the access logs.

6 CONCLUSION

To be effective in a knowledge and information-based society, individuals need tools that allow them to collect, manipulate and distribute the information about their products. The problem for authors and maintainers of such distributed resources is that to measure the utility of information in such a process and the question is: How do you analyse that the information you have placed on the web is being accessed in a significant way?

Organizations seeking to use server logs to measure usability can reach more accurate data by adding the following information to the log records:

1. Giving a unique number to each user will ensure that sessions are distinguished precisely.
2. Specifying different URLs for each step of processes like creating a legal document or making a payment will enable analyzers to obtain the detailed path of the users.
3. Saving details about the clicks made by the user (scroll bar movements, time, etc.) will enable effective analysis of usage data.

REFERENCES

Batista, P., Ario, M., and Silva, J., 2002. “Mining web access logs of an on-line newspaper,”.

Borges, J. and M. Levene, 1999. Data Mining of user Navigation Patterns, Web usage Analysis and User Profiling, *LNCS*. Abbas, H.A., R.A Sarker and C.S. Newton, (Eds.) pp: 29-111.

Chen, M. S., Park, J. S., and Yu, P. S., 1996. “Data mining for path traversal patterns in a web environment,” in *Sixteenth International Conference on Distributed Computing Systems*, pp. 385-392.

Das R., Turkoglu İ., 2009. Creating meaningful data from web logs for improving the impressiveness of a website by using path analysis method[J]. *Expert Systems with Applications*, 36(3): p. 6635-6644.

Jespersen S., Pedersen T. B., and Thorhauge J., 2003. “Evaluating the markov assumption for web usage mining,” in *WIDM ’03: Proceedings of the 5th ACM international workshop on Web information and data management*. NY, USA: ACM Press, pp. 82-89.

Jin, X.; Zhou, Y.; and Mobasher, B. 2004. A unified approach to personalization based on probabilistic latent semantic models of web usage and content. In *Proceedings of the AAAI 2004 Workshop on Semantic Web Personalization (SWP’04)*.

Kim Y.B., Kang S.J., Kim C.H., 2013. System for Evaluating Usability and User Experience by Analysing Repeated Patterns. In: *Marcus A. (eds) Design, User Experience, and Usability. Design Philosophy, Methods, and Tools. DUXU 2013*.

Liu, H., & Keselj, V., 2007. Combined mining of web server logs and web contents for classifying user navigation patterns and predicting users’ future requests. *Data and Knowledge Engineering*, 61(2), 304–330.

Munk, M., Kapusta j., 2010. Data pre-processing evaluation for web log mining: reconstruction of activities of a web visitor. *Procedia Computer Science*, 1(1): p. 2273-2280.

Nanopoulos, A. and Manolopoulos, Y., “Finding generalized path patterns for web log data mining,” in *ADBIS-DASFAA ’00: Proceedings of the East-European Conference on Advances in Databases and Information Systems Held Jointly with International Conference on Database Systems for Advanced Applications*. London, UK: Springer-Verlag, 2000, pp. 215- 228.

Poggi, N., Moren, T., 2009. Self-adaptive utility-based web session management. *Computer Networks*, 53(10): p.1712-1721.

Punin, J., Krishnamoorthy, M., and Zaki, M., 2001. “Web usage mining: Languages and algorithms,” in *Studies in Classification, Data Analysis, and Knowledge Organization*. Springer-Verlag.

Spiliopoulou, M., Mobasher, B., Berendt, B., & Nakagawa, M., 2003. A framework for evaluation of session reconstruction heuristic in web usage analysis. *INFORMS Journal on Computing*, 15(2), 171–190.

Zaiane, O. R., Xin, M., and Han, J., 1998. “Discovering web access patterns and trends by applying olap and data mining technology on web logs,” in *ADL ’98: Proceedings of the Advances in Digital Libraries Conference*. Washington, DC, USA: IEEE Computer Society, pp. 1-19.

Wang Y., Lee A., 2011. Mining Web navigation patterns with a path traversal graph. *Expert Systems with Applications*, 38(6): p. 7112-7122.