

Web Log Mining: A Study of User Sessions

Maristella Agosti and Giorgio Maria Di Nunzio
Department of Information Engineering – University of Padua
Via Gradegnigo 6/a, 35131 Padova, Italy
{agosti, dinunzio}@dei.unipd.it

Abstract

The analysis of Web log files may give information that are useful for improving the services offered by Web portals and information access and retrieval tools, giving information on problems occurred to the users.

This study reports on initial findings on a specific aspect that is highly relevant for personalization services: the study of Web user sessions.

Categories and Subject Descriptors

H.3.m [Information Storage and Retrieval]: Miscellaneous; H.2.8 [Database Management]: Database Applications — *Data Mining*.

General Terms

Experimentation, Measurement, Algorithms.

Keywords

Data mining, Web log mining, User session analysis.

1 Motivations

Web log file analysis began with the purpose to offer to Web site administrators a way to ensure adequate bandwidth and server capacity to their organization. This field of analysis made great advances with the passing of time, and now e-companies seek ways to use Web log files to obtain information about visitor profiles and buyers activities [4]. The analysis of Web log may offer advices about a better way to improve the offer, information about problems occurred to the users, and even about problems for the security of the site. Traces about hacker attacks or heavy use in particular intervals of time may be really useful to configure the server and adjust the Web site.

From the point of view of users, the Web is a growing collection of large amount of information, and usually a great portion of time is needed to look for and find the appropriate information. Personalization is a possibility for the success of the evolving of a Web infrastructure. A customer or a visitor who finds easily what he was searching for is a customer or a visitor that will return. For this reason, Web sites are created and adapted to made contents more easily accessible, using profiles found to make recommendations or to target users with ad hoc advertising. An ideal environment would dispose of exact history and information about a user, permitting to know his tastes and information needs.

A way to evaluate the effectiveness of a Web site and its information access tools is through the mining of web log files. In fact, the three main approaches to evaluate an information access service are:

- the studies based on test collection analysis, the so-called “*Cranfield approach*” [3],
- the user studies, and
- the analysis of log data.

The collaboration with The European Library¹, a service born to offer access to combined resources (such as books, magazines, and journals – both digital and non-digital) of 47 national libraries of Europe, gave us the possibility to start a project for the analysis of the data contained in the log files of The European Library Web servers. The European Library is a service set up by the Conference of the European National Librarians (CENL)².

One of the scope of the study is to evaluate the information access service to give recommendations for developing possible future personalization services. The log data used for the present

¹<http://www.theeuropeanlibrary.org/>

²<http://www.nlib.ee/cenl/>

analysis refer to the collections from 27 out of the 47 national libraries that were full partners at the moment of the analysis. The aim of this paper is to report on initial findings on a specific aspect that is highly relevant for personalization: the study of user sessions. To reach this aim Section 2 reports on the proposed approach to address the problem, Section 3 reports on the initial analysis of experimental data, and finally Section 4 gives some final remarks and indications for the continuation of the work.

2 Approach

The extraction of the data from Web logs gives access to information that have to be managed efficiently in order to be able to exploit them for analyses.

The solution we have developed is based on database management methods which permit the definition of an application which maintains and manages the necessary data. The specifications of the proposed solution were presented in [1]. The database and the application which have been developed enable separation of the different entities recorded and facilitate data-mining and on-demand querying of the log data.

We concentrate our attention here only on deriving the information on user sessions from the analysis of the *HyperText Transfer Protocol (HTTP)* requests made by clients, grouped in sessions, using a specific heuristic.

A request represents the data of the HTTP request that are recorded in the Web log files. A session is a particular set of requests made in a certain interval of time by the same client. Sessions are found, when information about sessions are not available, as in our case, through empirical rules, the heuristics.

Organizing the HTTP requests in a single session permits to have a better view of the actions performed by visitors. A procedure, named “session reconstruction”, may be used in order to map the list of activities performed by every single user to the visitors of the site. We used a heuristic that identifies a single user with the pair IP address and user-agent, and permits only a fixed gap of time between two consecutive requests. In particular, a new request is put in an existing session if two conditions are valid:

- the IP address and the user-agent are the same of the requests already inserted in the session [6],
- the request is done less than fifteen minutes after the last request inserted [2].

The reason for the choice of the couple of IP address and user-agent as identifiers is to distinguish different users coming from the same proxy. Different people using the same proxy result in requests done by the same IP, despite of the real identity of the clients. The introduction of the user-agent permits to differentiate more clearly the source of requests.

Empirical observations showed that there are high chances that different users are accessing the Web site when an amount of time greater than fifteen minutes passes between two requests from the same client.

With data so organized, it is possible to generate statistics about the visitors of the Web portal which can be used by a Web master in order to offer personalized information to the user.

3 Experimental Analysis

Experimental analysis was performed on the available Web log files, that correspond to eleven months of The European Library Web log files, starting from October 31st 2005 to September 25th 2006. We stopped the analyses before September 26th since, from that day, the records in the log file slightly changed in order to incorporate new data (such as cookies and track sessions). The structure of the log file record is conform to the W3C Extended Log File Format [5].

The analyses, that are presented in the following, cover software tools such as operating systems and browsers used by clients, sessions in terms of daily distribution, and time intervals per number of HTTP requests.

The numbers we are reporting include all the requests and sessions, even those ones which can belong to automatic crawlers and spiders.

A total of 25,881,469 of HTTP requests were recorded in the log files of the eleven months. Table 1 reports the distribution of HTTP methods which are present in the log files.

During this period, according to the heuristic we chose, 949,643 sessions were reconstructed. These numbers suggest that each client makes on average 27.25 accesses per session. The number of distinct pairs IP address and user-agent is equal to 285,158.

Figures 2a and 2b show the distribution of the operating systems and the browsers used by visitors respectively.

It is possible to see how the products of Microsoft are by far the most used by visitors of The European Library portal: Windows alone is used by about 75% of the users; this tendency also affects the situation found in Figure 2b, with In-

Table 1: Number of HTTP requests for each method.

HTTP method	total number
CONNECT	2
LINK	6
PROPFIND	760
PUT	3,640
OPTIONS	3,779
HEAD	33,770
POST	844,058
GET	24,995,454
Total	25,881,469

Internet Explorer as the most used browser. However, we noticed a significant increase in the use of Mozilla Firefox, compared to what we found in a preliminary analysis of a sample of the initial months of the logs (from November 2005 to January 2006) as reported in [1].

In Figure 3a, we present the number of sessions per hour of day. The distribution is the one expected for European countries, considering that the time recorded by the server refers to the Central European Time: a higher activity during the day compared to the one during the night.

In Figure 3b, the number of sessions per HTTP request intervals are shown. Sessions with less, or exactly, 25 requests are the 75% of the total number of sessions. This means that the great majority of sessions have a small number of requests. Sessions with more than 100 requests are only the 6% of the total, but those sessions play an important role in the study, since they may belong to deep browsing of the portal.

Because of this peculiar uneven distribution, we carried out a further analysis on the study of sessions according to both the number of requests and the length of the session. In Figure 4, a barplot which relates the number of requests per session to the length of sessions, computed in seconds, is shown. In order to make the barplots visible and clear enough, we split the figure in two parts: sessions with less than, or exactly, 25 requests (plot on the left side), and sessions with more than 25 requests (plot on the right side).

For the first set of data, we divided the intervals of the length of sessions into: less than, or exactly, 20 seconds, between more than 20 and 40, and more than 40 seconds. The analysis shows that sessions with less than 25 requests are very short, in terms of number of seconds, with very few exceptions. It is still under investigation the study of this minority, that represents sessions with few requests which are sent very slowly.

For sessions with more than 25 requests, plot on the right part of the figure, we chose a finer grain for the time interval of the length of sessions: 10 seconds for each interval, apart from the last row where the sessions that last more than 60 seconds are represented. Many sessions have less than 50 requests.

Nevertheless, there is a sizeable number of sessions, corresponding to the 16% of sessions, which last more than 60 seconds regardless of the number of requests per session. It is interesting to note that 12% of the sessions contain more than 50 requests. An analysis of the sessions with more than 100 requests has been computed separately, since we believe that these sessions are valuable for the analysis of users for personalization purposes. The results are shown in Figure 1. It is worth to note that the subset of sessions which last from 2 to 30 minutes is the biggest.

An experiment which is currently under investigation concerns a controlled study of a group of users who have been asked to freely crawl and navigate The European Library Web site and, after that, to fill in a questionnaire provided by The European Library to report and describe their impressions. The goal is to combine the data of the sessions of the people which have compiled the questionnaires, data which are present in the Web log files, with those that have been reported in the questionnaires with the aim of gaining insights from data on user sessions and judgements in the questionnaires to be used for personalization purposes.

4 Conclusions and Future Work

In this paper, we presented a preliminary analysis of eleven months of The European Library Web log data, according to a methodology for gathering and mining information presented in [1]. The aim of this work was to report on initial findings about the study of user sessions which have been reconstructed by means of heuristic methods, since no personal data was available to track each user.

The heuristics used to identify users and sessions suggested that authentication would be required since it would allow Web servers to identify users, track their requests, and more importantly create profiles to tailor specific needs. Moreover, authentication would also help to solve the problem concerning crawlers accesses, granting access to some sections of the Web site only to registered users, blocking crawlers using faked user agents.

As a follow up of the cooperation with The European Library, the Office of The European Library has implemented the changes, that the re-

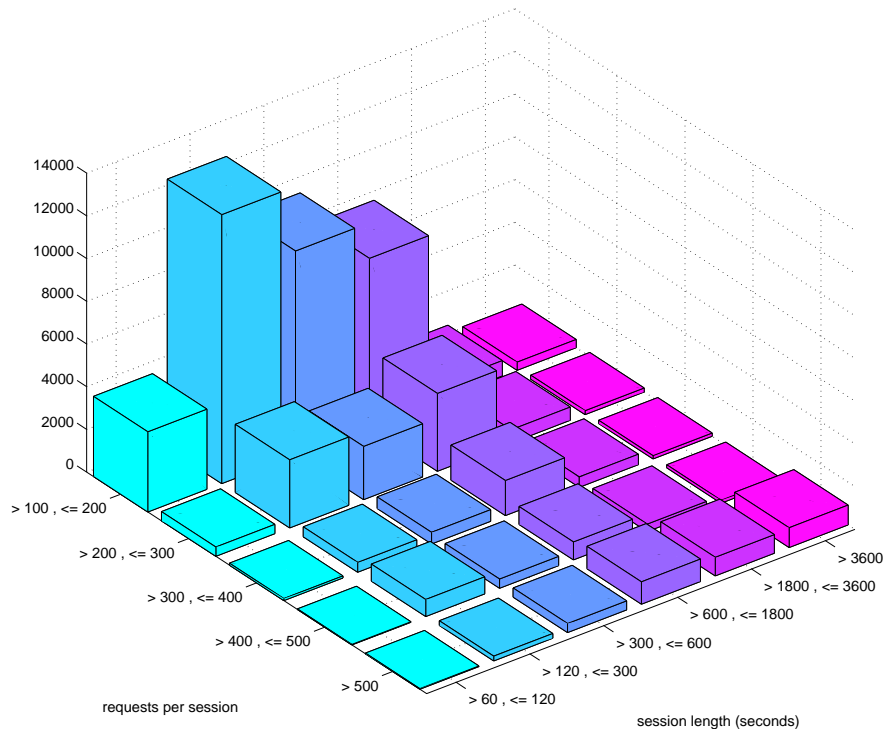


Figure 1: Sessions per time-intervals in seconds with respect to the number of requests per session. The reported sessions are those with at least 100 HTTP requests.

sults of the analysis here reported were suggesting, in its HTTP server logging system (September 2006) and that The European Library Office has also established a user authentication procedure (since August 2006).

Acknowledgements

The work reported in this paper is conducted in the context of a joint effort of the DELOS Network of Excellence on Digital Libraries and the The European Library project.

The work has been partially supported by the DELOS Network of Excellence on Digital Libraries, as part of the Information Society Technologies (IST) Program of the European Commission (Contract G038-507618).

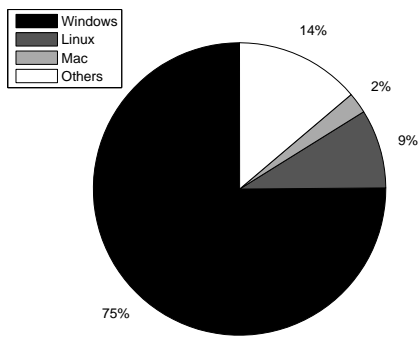
The authors wish to thank all the staff of The European Library for the continuing support and cooperation. Sincere thanks are due to Tullio Copotelli for the useful discussions.

References

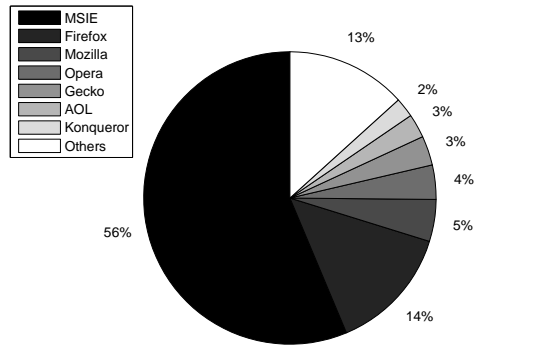
[1] M. Agosti, G.M. Di Nunzio and A. Niero “From Web Log Analysis to Web User Pro-

filin” In DELOS Conference 2007. Working Notes. Pisa, Italy, 2007, pp 121–132.

- [2] B. Berendt, B. Mobasher, M. Nakagawa, M. Spiliopoulou “The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis”, WEBKDD 2002, LNAI 2703, pp 159-179, 2003.
- [3] C. W. Cleverdon “The Cranfield Tests on Index Languages Devices”. In Readings in Information Retrieval, Morgan Kaufmann Publisher, Inc., San Francisco, California, pp.47–60, 1997.
- [4] F.M. Facca, P.L. Lanzi “Mining interesting knowledge from Weblogs: a survey”, Data and Knowledge Engineering Vol. 53, No. 3, June 2005, pp 225-241.
- [5] P.M. Hallam-Baker, B. Behlendorf “Extended Log File Format, W3C Working Draft WD-logfile-960323” <http://www.w3.org/TR/WD-logfile.html>.
- [6] D. Nicholas, P. Huntington, A. Watkinson “Scholarly journal usage: the results of deep log analysis”, Journal of Documentation Vol. 61 No. 2, 2005.

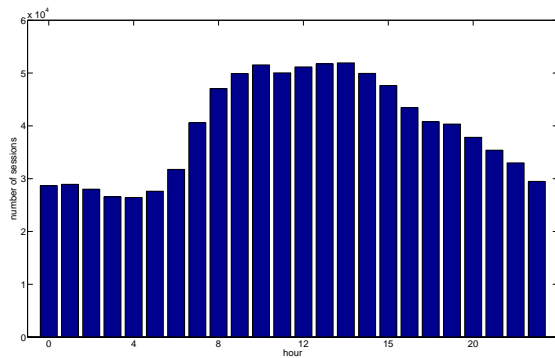


(a) Operating systems used by clients.

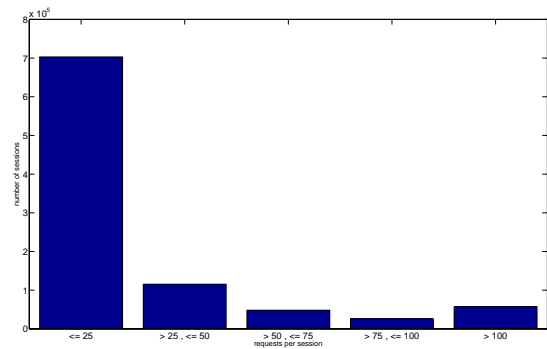


(b) Browsers used by clients.

Figure 2: Web log general statistics concerning HTTP requests.



(a) Sessions per hour of day.



(b) Sessions per number of requests.

Figure 3: Web log analysis of sessions per hour and number of requests.

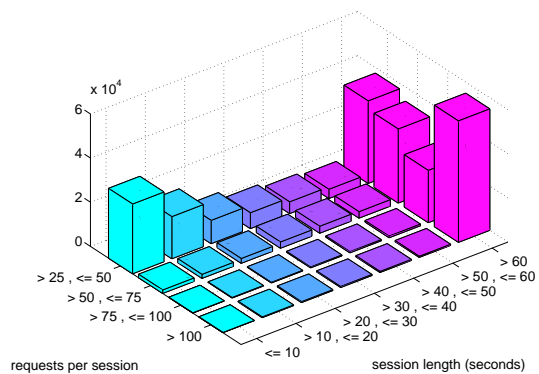
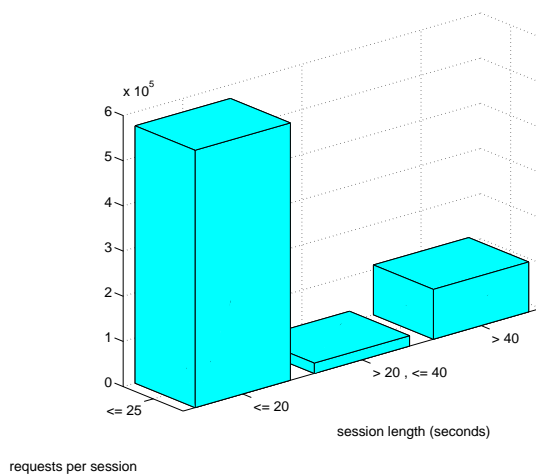


Figure 4: Sessions per time-intervals in seconds with respect to the number of requests per session.