Analysing HTTP Logs of a European DL Initiative to Maximize Usage and Usability

M. Agosti[1], G. Angelaki[2], T. Coppotelli[1], and G.M. Di Nunzio[1]

[1] University of Padua, Italy
{agosti,coppotel,dinunzio}@unipd.it
[2] The European Library, The Netherlands
Georgia.Angelaki@KB.nl

Abstract. In the context of an ongoing collaboration conducted between DELOS, the European Network of Excellence on Digital Libraries, and The European Library, we discuss how both the analysis of the Web log data of The European Library service and a user study can contribute to the personalization of services for such a system.

1 Introduction

The European Library is a non-commercial organisation[1]; it provides the services of a physical library and allows searching through the resources of many of the European national libraries, where the available resources can be both digital or bibliographical, e.g. books, posters, maps, sound recordings and videos. This paper presents results that have been achieved up to now on the analysis of who the users of The European Library are as well as how its functionalities are perceived by its users. The analysis involves two different and complementary strategies: 1) Web log analysis; and 2) user study.

HTTP logs of The European Library server are used to reconstruct the users sessions and to study the users behaviour. On the one hand, the effective number of users can be estimated and data can be obtained about users mean sessions length by using HTTP logs and by eliminating the access of crawlers. It is also possible to estimate if the users use advanced search functionalities or if they prefer to use the portal to search documents with the minimum effort, and obtain data about their geographical distribution. On the other hand, it is possible to understand how users exploit The European Library portal, what they expect from it, and what they would like to get by using additional information like user questionnaires. With the use of questionnaires, knowledge about their level of satisfaction can be obtained along with recommendations and hints about possible improvements.

2 The Initiative

The European Library initiative aims at providing a "*low barrier of entry*" for the national libraries that should be able to join the federation with only minimal

[1] http://www.theeuropeanlibrary.org/

D.H.-L. Goh et al. (Eds.): ICADL 2007, LNCS 4822, pp. 35–44, 2007.

changes to their systems [1]. With this objective in mind, The European Library service is constituted by three components:

- a Web server: which provides users with access to the service;
- a central index: which harvests catalogue records from national libraries, supports the *Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)*[2], and provides integrated access to them via *Search/Retrieve via URL (SRU)*;
- a gateway between SRU and Z39.50: this also makes national libraries accessible through SRU which would otherwise only be accessible through Z39.50[3].

In addition, the interaction between the portal, the federated libraries and the user mainly happens on the client side by means of an extensive use of Javascript and *Asynchronous JavaScript Technology and XML (AJAX)*[4] technologies. Once the client, which is a standard Web browser, accesses the service and downloads all the necessary information from the Web server, all the subsequent requests are managed locally by the client. The client interacts directly with each federated library and the central index, according to the SRU protocol, makes separate AJAX calls towards each federated library or the central index, and manages the responses to such calls in order to present the results to the user and to organize user interaction.

3 Web Log Analysis

The analysis was performed on seven months of The European Library Web log files, starting from October 1st 2006 to April 30th 2007. The structure of the Web logs conforms to the W3C Extended Log File Format [2].

This kind of log contains, among other things, the following useful information: 1) the *Internet Protocol (IP)* address and the user-agent which allow the identification of single users [3]; and 2) the referrer field, a *Uniform Resource Locator (URL)* address which communicates the last page viewed by the user, which can be used to know how visitors arrive at The European Library service.

The European Library Web logs also contain the cookie[5] saved on the client which reports extra information: 1) the language selected by the user during the navigation of the service; 2) the collections of documents selected during the query or query refinement; 3) the identifier of the session assigned by the server to a specific user.

A methodology for analysing these log files has been developed. It requires the use of a parser and a database for storing the data. Initial specifications of this database application were presented in [4]. The database enables separation

[2] `http://www.openarchives.org/OAI/openarchivesprotocol.html`

[3] `http://www.loc.gov/z3950/agency/`

[4] `http://www.w3.org/TR/XMLHttpRequest/`

[5] Cookies are plain text information stored locally by the client. The stored data are initially sent by a Web server to a Web client and then are sent back to the server on subsequent requests.

of the different entities recorded and facilitates data-mining and on-demand querying of the data.

3.1 General Information from the Analysis

The following data include all the requests and sessions that reached the portal, even those which can belong to automatic crawlers and spiders. These data can give a a first estimation of the trend of the traffic volume: a total of 22,458,350 HTTP requests were recorded in the log files, with a monthly average of 3,208,336 HTTP requests, a daily average of 105,936 requests, and an hourly average of 4,414 requests. Since The European Library service is a 7 days and 24 hour service with users from all over the world, it can be considered a busy service that answers to an average of 74 requests per seconds, 24 hours a day/7 days a week.

This traffic is generated both by human users and by automatic software agents. Before starting any personalization analysis, it is then mandatory to identify human requests from others. A simple analysis performed on the user-agents included in each request allows a first distinction between human requests and software-agent requests. The experimental data are reported in Table 1. The fields contained in the user-agent are decided by each browser and different versions of the same browser can have different fields. As a consequence, an automatic elaboration of this data is a hard task. However, the most used browsers use a standard-like string that allows the identification of such browsers. For example, the following user-agent, that is the most recurrent one in the analyzed logs, corresponds to Internet Explorer 6.0 (MSIE 6.0):

```
Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1)
```

There is a heterogenous quantity of data that browsers can store in this field. For example, it is possible to find complex user-agents such as:

```
Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR
+1.0.3705;+.NET+CLR+1.1.4322;+Media+Center+PC+4.0;+.NET+CLR+2.0.50727)
```

Nevertheless, the aim at this point is not to correctly identify the meaning of each field, but to use this information to distinguish a human from a crawler. To this aim, the first step is that of tagging as crawlers all the user-agents that contain key terms like crawler, robot, spider, etc. Other crawlers are identifiable because they disclose themselves in the user-agent, for example:

```
Mozilla/5.0+(compatible;+Yahoo!+Slurp;
+http://help.yahoo.com/help/us/ysearch/slurp)
```

is the well known Yahoo! crawler. There are some repositories that contain up-to-date lists of known crawlerse[6].

After crawler identification, all the user-agents generated by internal usage of the server can be tagged as software agents. In this category it is possible to find

[6] One of those repositories is available at the URL: http://www.botsvsbrowsers.com/

Table 1. Distribution of requests on the basis of user agents

origin	requests	percentage
human	17,006,566	75.72%
crawler	2,904,134	12.93%
software	2,460,867	10.96%
non standard	86,783	0.39%
total	22,458,350	100.00%

user-agents that correspond to requests made by: Verity *Information Retrieval System (IRS)* (the IRS that handles query execution in The European Library), Java or PERL applications (`Java/1.4.2_04`, `libwww-perl/5.801`), and other software tools like the MS FrontPage HTML editor. Some user-agents are visibly faked to resemble that of known browsers and then are tagged as non-standard; for example

`Mozilla/4.0+(compatible)` and `Mozilla/4.0+(compatible;+MSIE+5.00;)`

lack required information like the browser or the operating system. The remaining user-agents are manually checked and are tagged as human when they seems to have no anomalies.

The high percentage of human users obtained with this technique is partially biased by the behaviour of some crawlers that successfully fake a user-agent and are recognized as a human user. Despite this, the produced estimation of human requests can be considered good.

Other statistical information that is computable using these logs regards operative systems and browsers. The products of Microsoft are by far the most used: Windows alone is used by about 74% of the users; this tendency also affects the situation found in the browser analysis, with Internet Explorer as the most used browser, since it is used by 60% of the users. However, there has been an increase in the use of Mozilla Firefox, compared to a preliminary analysis performed in previous months of the logs (from November 2005 to January 2006) as reported in [4]; currently Mozilla Firefox is used by 13% of the users.

3.2 Session Reconstruction

With the term session a set of requests is intended that are performed by a single user during the browsing activity. Because a user is supposed to access the portal more than once during the analysed period, a time-out is applied to distinguish different sessions of the same user. When the user remains inactive for more then 15 minutes, his session is terminated and a new one is created when the next request is performed.

The reconstruction of sessions is an important process that allows the identification of single users (either humans or software-agents), their tracking during

the portal navigation and eventually their grouping on the basis of similar behaviors. Moreover, the session identification is an intermediate step that has to be performed in order to distinguish recurrent users from bouncers. We have proposed two different methodologies for reconstructing sessions: 1) a heuristic technique [3] that allows the identification of a single user using the IP address and the user-agent; and 2) an exact technique that takes advantage of the information contained in the cookies to reconstruct the sessions.

The first technique assigns each HTTP request to a session, including sessions of users that do not accept cookies. In this way, a great number of sessions (690,879) is reconstructed, but the drawback is that a large portion of them is made up of an extremely reduced number of requests, hence they are not exemplificative of human navigation behaviour. However, it is possible to analyze the different kinds of software agents that interact with the Web server. For example, there is a significant number of sessions that do not contain any HTTP request of a Web page, while the first request of a human user would clearly be a Web page.

The second approach to session reconstruction is therefore preferred when an analysis of human sessions is required, and it takes advantage of cookies to identify sessions. The European Library cookies contain a unique identifier, named TELSESSID, assigned runtime by the PHP[7] interpreter of the Web server each time a session is started by a user. This identifier is important for two reasons: 1) it more precisely distinguishes users that are hidden behind a proxy (therefore with the same IP); and 2) it enables human users to be separated from the other users of the portal because many of the automatic crawlers and softwares do not memorize cookies. The requests of the same session have the same TELSESSID value in the cookie field; in order to be coherent with the previous analysis, we decided to consider two sessions distinct when the delay between two requests of the same session is greater than 15 minutes. This approach cannot take into consideration sessions of users which do not enable cookies in their browser. Despite this drawback, the results are more accurate than those obtained with the first approach of sessions reconstruction. A deeper analysis and comparison of the results obtained with these two different approaches is under way and has already allowed the identification of odd behaviour in the sessions of users from some countries.

In the analysed period of time, we were able to reconstruct 209,900 different sessions on the basis of the cookies content. There is a significant number of sessions, almost 45% of the total number of sessions, which last more than 60 seconds regardless of the number of requests per session. Therefore, an analysis of the sessions which last more than 60 seconds and have more than 100 requests has been computed separately, since these sessions are valuable for the analysis of users for personalization purposes and to give an answer to the points of interest. Results shown in Figure 1 are important since they confirm that users do not only have a look at the home-page of The European Library portal, but they spend some time on the portal, interacting with it (more than 100

[7] `http://www.php.net/`

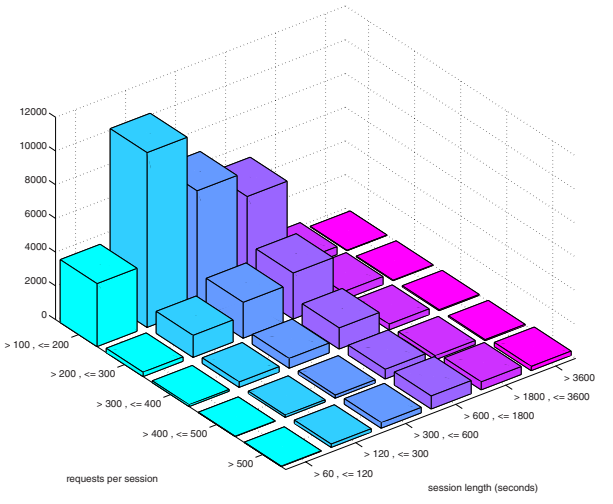


Fig. 1. Sessions (cookies) which last more than 60 seconds with a number of requests per session > 100

HTTP requests) and analysing the results (the majority of sessions last between 2 minutes to 10 minutes). Therefore, it seems reasonable to think that there is a sizeable number of users that do not only come across the portal but that actively use the functionalities that it offers, users that are really interested in the search of documents which are written in their own mother language and also in other languages.

3.3 Session Provenance

In this section, the analysis of sessions classified according to the different geographical areas are shown. The abbreviations that are used in the graphs are those adopted by the ISO 3166 standard[8], which is a three-part geographic coding standard for coding the names of countries and dependent areas, and the main subdivisions thereof.

The nations with the highest number of sessions reconstructed using the cookies are shown in Figure 2. Most of the accesses come from European nations that are active members of the The European Library service and, in particular, there is a noticeable increase in the number of accesses from the countries that recently joined the initiative. The United States of America are second in this list; however, we recorded a significant decrease in the number of sessions if compared with sessions reconstructed with HTTP requests. This fact indicates that the majority of crawlers have an IP address belonging to the geographic area of the United States.

[8] `http://www.iso.org/`

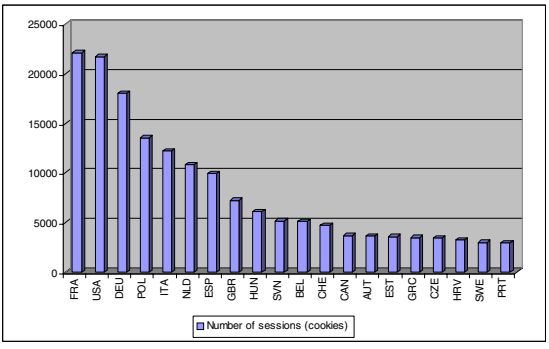


Fig. 2. The nations with the highest numbers of sessions (cookies) are shown

3.4 Advanced Search Usage

The users of The European Library have two different choices for performing advanced searches: use the advanced search functionality or personalize the set of searched collections. While we do not have any information on the first choice, the information on the collections selected by the user is saved in a specific field of the cookies and the analysis of this variable makes the analysis of the personalized searches possible.

In general, we can observe that users usually use the default collection selection instead of explicitly selecting which collections have to be searched; to give an idea, the ratio between explicit selection and the default selection is less than 1 to 10. However, the following analysis focuses only on those collections that are explicitly selected by the user and does not consider the collections assigned by default to the user. The study of these collections allows us to understand the behavior of users that are actually refining the query. The selection is equally distributed over these collections, with a mean of 1,708 selections, and a maximum of 11,000. Thus, it appears that only a reduced number of users actively selects a different set of collections; therefore, it is important to accurately select the initial set of collections in order to have a better exploitation of The European Library. Moreover, if this number of users corresponds to those users who perform an advanced search, then the behavior is comparable with that of other online services where only a limited number of users effectively uses advanced search tools.

4 User Study

Users surveys are a valuable method for understanding user behaviors in different situations. However, surveys usually require a significant amount of time and effort; for this reason, an accurate design of the process of studying users has to be carried out. Extensive methods can be used to broadly represent a population of users and investigate some characteristics of interest such as: the background,

the information need of the users, and the level of their satisfaction given a service. Surveys or questionnaires are types of extensive methods that can be used to interview users, and the goals of these methods can be those of learning how to better develop the service under investigation. Questionnaires can also provide simple feedback to build up an understanding of the different way users perceive the search tools provided by a service.

4.1 Study of the Willingness of Users for Interactive Search

Here, we want to discuss the kind of activities that users perform during the usage of The European Library service. We are particularly interested in studying the willingness of users for iterative search and the satisfaction of search tools offered by the service. Searching means both a simple action like specifying terms of interest, and a complex action like browsing results and iterating the search using more focused terms. In The European Library, we also have information about user preferences on collections during the search of documents: the country of the collection of the national library, the subject of the collection, and so on. The final aim of the analysis will be that of combining the observations carried out with the questionnaires and the results carried out with the log analysis.

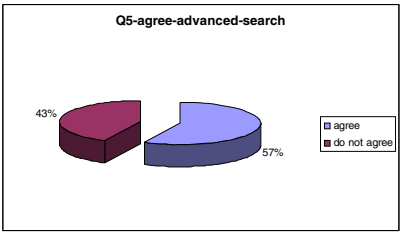
In the following, the results of a controlled study are presented. It was decided to conduct a controlled study, because previous studies on logs and observation in naturalistic settings, combined with interviews, seem more scientifically informative with respect to each of the two types of studies when conducted alone[5]. The final aim of the study is to gain insights on a specific group of data, and to use them in a more general way.

4.2 Controlled Study Set Up

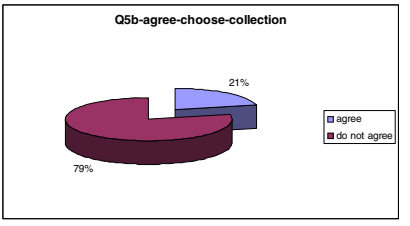
A controlled study was conducted for a group of users who were asked to freely crawl and navigate The European Library Web site and, after that, to fill in the questionnaire provided by The European Library to report and describe their impressions[9]. The goal of this controlled study was to combine the data in the Web log files of the sessions of the people who compiled the questionnaires with those that were 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.

The first important question of the questionnaire that gives a first impression about how users use The European Library portal is question Q3 which requires choosing between the following answers: 1) I prefer to use the Simple Search facility; 2) I prefer to use the Advanced Search facility. Results shows that 81% of users prefer the advanced search facilities. In Figure 3a and 3b, the percentage of users that agree with the following sentences are shown: (Q5a) An advanced search automatically searches all the collections held by The European Library; (Q5b) I usually use the "*choose your own collections*" option before undertaking an advanced search.

[9] The European Library. Questionnaire on The European Library's Portal, 2006.



(a) User agreement on advanced search tools. Question Q5a.



(b) User agreement on using their own collections. Question Q5b.

Fig. 3. Users agreement on search tools and collections

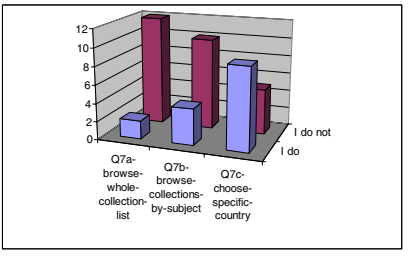
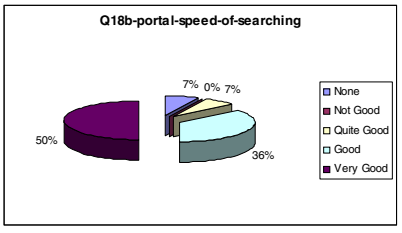


Fig. 4. How users use the "*choose your own collections*" option. Question Q7.

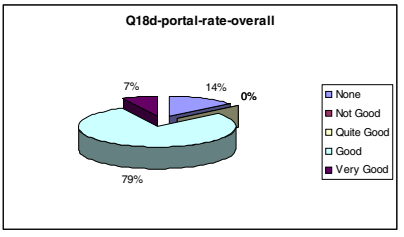
The numbers indicate that there is almost a complete disagreement among users about how many collections are used during an advanced search. Moreover, most of the people using advanced search tools do not make use of the option that allows users to select their preferred collections. However, there is still a significant number of users (14%) who do not know after this experiment how to use the "*choose your own collections*" option (question Q6 of the questionnaire). In Figure 4 the number of users who use and how they use the "*choose your own collections*" option is shown. The possible answers were (question Q7 of the questionnaire): (Q7a) I browse the whole collections list to find relevant collections; (Q7b) I use the "*browse the collections by subject*" option; (Q7c) I choose a specific country and look at collections; (Q7d) I use the "*search collections by description*" option.

For those users who do use this option, few of them like to browse the whole list of collections (Q7a) or browse the subject of the collections (Q7b). Instead, the majority of users prefer to browse the collection choosing the countries of interest (Q7c). None of them, not shown in the figure, have ever used the search collection by description option.

In Figure 5a and 5b, two different types of evaluation of the portal are shown from the point of view of the speed of search and the overall rating. Half of the users are not satisfied with the portal speed of searching; however, and most importantly, more than 80% of the users at least rate the portal as good.



(a) Speed of portal rating (Q18b).



(b) Overall portal rating (Q18d).

Fig. 5. Web portal rating. Question Q18.

5 Conclusions

Preliminary analysis using both HTTP requests and user surveys has shown that users of The European Library come from different geographical areas, especially from countries that recently joined the initiative, some of them are willing to spend some time to perform advanced searches and to select collections of documents different from the default ones. Further analysis is underway to better combine the results of the HTTP logs analysis with that of the user study to give a better understanding on the usage of The European Library service and give directions towards an innovative personalization of the service.

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).

References

1. van Veen, T., Oldroyd, B.: Search and Retrieval in The European Library. A New Approach. D-Lib Magazine 10 (2004)
2. Hallam-Baker, P., Behlendorf, B.: Extended Log File Format, W3C Working Draft WD-logfile-960323 (1996), `http://www.w3.org/TR/ WD-logfile.html`
3. Agosti, M., Di Nunzio, G.: Web Log Mining: A study of user sessions. In: PersDL 2007. Proc. 10th DELOS Thematic Workshop on Personalized Access, Profile Management, and Context Awareness in Digital Libraries, Corfu, Greece, pp. 70–74 (2007)
4. Agosti, M., Di Nunzio, G., Niero, A.: From Web Log Analysis to Web User Profiling. In: Thanos, C., Borri, F., Candela, L. (eds.) DELOS 2007. LNCS, vol. 4877, pp. 121–132. Springer, Heidelberg (2007)
5. Ingwersen, P., Järvelin, K.: The Turn. Springer, The Netherlands (2005)