A Study of Web Logs for Personalizing the MultiLingual Information Access to The European Library

Maristella Agosti[1], Tullio Coppotelli[1], Giorgio Maria Di Nunzio[1], Nicola Ferro[1], and Eric van der Meulen[2]

[1] University of Padua, Italy
{agosti, coppotel, dinunzio, ferro}@unipd.it
[2] The European Library, The Netherlands
Eric.vanderMeulen@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 the analysis of the Web log data of The European Library service can contribute to the design and development of MultiLingual Information Access (MLIA) personalized services for such a system.
In particular, we discuss the present architecture of the system, how this architecture impacts on both Web logs and MLIA functionalities, and we present how the achieved results of log analysis can be exploited for obtaining clues about how to introduce MLIA functionalities in the system.

1 Motivation

The paper reports on the results of a ongoing collaboration between DELOS[1], the European Network of Excellence on Digital Libraries, and The European Library[2], which provides integrated access to the major collections of the European national libraries.

The study focuses on the analysis of Web log data to obtain information about users and their behaviour. In particular, we are interested in investigating how the analysis of Web log data can help us in implementing full MultiLingual Information Access functionalities in The European Library, where by full MLIA we mean the possibility for the users of the system to access and search the federated libraries in a personalized way that can allow them to access the collections of documents in their mother tongue and in other preferred languages.

The large use of Web logs on Web servers to record how a Web portal is used suggests that a deep analysis of these kinds of logs may bring to the surface some hidden and important information about users [4].

[1] http://www.delos.info/
[2] http://www.theeuropeanlibrary.org/

When the collection used for learning is composed by a set of collections each with documents in different languages and the users themselves have different geographical provenance, the need for MLIA functionalities gains a central role on the Digital Library (DL) usage. The collections managed by the European national libraries are, by their nature, suitable for learning and not only for browsing or referencing. The results obtained with this analysis can be generalized to similar services.

In particular, in this paper we answer the following questions: 1) are users interested in the use of multilingual services or cross-collections retrieval instruments? 2) what are the most valuable collections, and which ones require more effort to be correctly perceived and used, also with respect to the tools and resources needed to make them available in a multilingual setting? 3) is the behavior of a user influenced by hidden preferences about internationalization in software like Web browsers?

Moreover, we present two possible approaches to MLIA based on the specific architecture of The European Library and discuss their feasibility on the basis of Web logs.

The paper is organised as follows: Section 2 describes the architecture of the service and some issues related to it; Sections 3 and 4 describe solutions for dealing with the issues related to the system architecture from the viewpoint of Web log analysis and of MLIA functionalities; Section 5 describes how to exploit log analysis for gathering relevant information about MLIA; finally, Section 6 draws some conclusions.

2 Architectural Overview

The European Library service aims at providing a "low barrier of entry" for the national libraries that should be able to join the federation with only minimal changes to their systems [11]. This ease of integration is achieved by extensively using the Search/Retrieve via URL (SRU)[3] protocol in order to search and retrieve documents from national libraries. In this way, the user client can be a simple browser, which exploits SRU as a means for uniformly accessing national libraries.

With this objective in mind, The European Library service is constituted by three components: 1) a Web server: this provides users with access to the service; 2) a central index: this harvests catalogue records from national libraries, supports the Open Archives Initiative Protocol for Metadata Harvesting (OAI-PMH)[4], and provides integrated access to them via SRU; 3) a gateway between SRU and Z39.50: this makes national libraries, which would otherwise be accessible only through Z39.50[5], also accessible through SRU.

In addition, the interaction between the portal, the federated libraries, and the user mainly happens on the client side by extensive use of Javascript and

[3] http://www.loc.gov/standards/sru/

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

[5] http://www.loc.gov/z3950/agency/

Asynchronous JavaScript Technology and XML (AJAX)[6] technologies. Once the client, which is a standard Web browser, has accessed the service and downloaded all the necessary information from the Web server, all the subsequent requests are managed locally by the client. The client directly interacts with each federated library and the central index, according to the SRU protocol, it 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 organize user interaction.

Issues

The architecture and functioning of the service, as described above, pose some problems when gathering information in the Web logs and planning to introduce MLIA functionalities.

With respect to Web logs, much of the interaction between the client and The European Library service cannot be recorded in the logs, since it happens outside the system and it is scattered around the clients and the various national library systems. Therefore, the logs of the Web server mainly contain information about the navigation and browsing within the portal while the information about the interaction with the national libraries is only partially contained in the logs.

With respect to MLIA, the service has no control on queries sent to the national libraries, since the browser directly manages the interaction with national library systems via SRU. As a consequence, introducing MLIA functions into the central system would have no effect on the national library systems. Therefore, in order to achieve full MLIA functionality, not only the system but also all the national library systems would have to be modified. This is an unviable option as it would require a major effort and disregards the "low barrier of entry" guideline adopted when designing The European Library service.

On the other hand, the central index harvests catalogue records from national libraries, which beside catalogue metadata may contain other information useful for applying MLIA techniques, such as an abstract. Since the central index is completely under the control of the system, we plan to extend its functionalities by adding a component able to translate the catalogue records in order to perform MLIA on them.

3 Web Log Analysis

The structure of The European Library Web logs conforms to the W3C Extended Log File Format [10]. They contain, among other things, the following useful information: 1) the Internet Protocol (IP) address and the user-agent which allow the identification of a single user [3]; 2) the referrer field, a Uniform Resource Locator (URL) address which communicates the last page viewed by the user. This is very useful to know the way visitors get to The European Library service.

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

Tel Web logs also contain the cookie saved on the client which reports extra information too: 1) the language selected by the user during the navigation of the service; 2) the document collections selected during the refining of the query; 3) the session identifier assigned by the server to a specific user.

Due to the reduced interaction between client and server and the high quantity of data, we built our own methodology for analysing these log files. Our methodology copes with the specific structure of the portal, enables the separation of the different entities recorded and facilitates data-mining and on-demand querying of the data. Upon the reception of a new log file, the file is elaborated by our parser which separate the different information and store them in a database [4]. The database enables the ad-hoc analysis of the data, allowing us to focus on the interesting aspects of the service usage.

The investigation of the use of the portal can be used to tailor specific MLIA needs according to the type of human users. It is important, before to proceed with MLIA solutions, to understand if the user spends few time in the portal, making only a query, and wants the results immediately or if it spends more time on the portal and then needs some advanced translation services. Moreover, the understanding of user needs, their preferences about languages, and specific collections of documents of interest allow the personalization of user experience.

4 MultiLingual Information Access

A feasibility study [1,8,9] has been conducted with the aim of proposing and assessing alternatives able to cope with the architectural characteristics of the system and, at the same time, provide The European Library users with rich MLIA functionalities.

A two-step solution is suggested consisting of two complementary approaches: *isolated query translation* and *pseudo-translation*. The first provides a basic cross-language search functionality for the entire system; the second operates on the central index.

Isolated Query Translation This can be considered as a sort of pre-processing step where the translation problem is treated as completely separate from the retrieval. Before actually submitting the query to the national libraries or the central index, the user asks for a translation of the query and the browser sends the query, via SRU, to a dedicated component of the system which translates it and also applies query expansion techniques to reduce the problem of missing translations [6]. At this point, the user can interactively select the translation which best matches his needs or can change some query term to refine the translation. In the latter case, the translation process may be iterated. Once the desired translation of the query is obtained, the retrieval process is executed using both the translated query and the original one.

Pseudo-translation In this approach, the complete bibliographic records contained in the central index are translated into different target languages.

Indeed, although query translation is more often adopted, [7] reports as document translation can be very competitive in some general cross-language retrieval settings. Document translation is usually very resource-demanding in the case of large documents but in the case of the central index records their brevity makes document translation a viable solution. In addition, since the translation is made for retrieval purposes and not for presentation, there is no need to produce a syntactically correct or semantically unambiguous document as the translation will remain hidden to the end user. Therefore, the whole translation process can be implemented in a lighter way and, as described in [8], the obtained translations are still effective for retrieval purposes.

Although the results of the feasibility study were encouraging, they are a long way from solving the problem of implementing true MLIA functionality in The European Library. Indeed, both solutions were proposed only in a language-to-language context (i.e. with queries in one language against target collections in a second language) while The European Library currently manages 20 different languages and in the near future another 8 languages will be added.

This setting would need to be able to manage around 400 language pairs. The linguistic resources required for offering so many languages would be incredibly huge and for many pairs of languages linguistic resources are probably non-existent.

Therefore, in order to reduce the number of language pairs needed, an or pivot language could be used. Instead of directly translating from one language to another, we have a preliminary translation from the source language to the pivot language and a further translation from the pivot language to the destination language. A number of studies have attempted to evaluate the performance loss that can be expected with a pivot language and strategies to reduce this have been proposed [5].

Thus, a first question we aim to investigate with the analysis of the Web logs is when we need to provide direct translation, i.e. manage the language pair, and when we can rely on the pivot language. Multiple factors may influence this decision: the geographic distribution of the users, the fact that some national libraries are queried more often from users that come from another nation, the fact that some collections exist only in one language and it is thus valuable offering access to them from other languages, or the need for promoting the use of less known collections which might be less accessed only because of linguistic barriers. By analysing the Web logs, we aim at finding information about these influencing factors in order to be able to answer the original question.

Moreover, the "Isolated Query Translation" approach involves some user interaction, since the user may iteratively refine the translations before sending the query. The way in which the interaction is constructed should depend on the behaviour and the profile of the user. For example, there may be users who prefer a very quick interaction and rely on the first translation proposed by the system or there may be users who spend more time with the system and would prefer to have a richer interaction and more advanced tools for refining the translation.

Therefore, analysing the Web logs for studying user behaviour and gathering information about the session length for different user typologies can provide us with interesting insights about how to design the "Isolated Query Translation" feature.

5 Exploiting Log Analysis for MLIA

The following analysis performed on the Web log data corresponds to seven months of The European Library Web log files, starting from October 1st 2006 to April 30th 2007. A total of 22,458,350 HTTP requests were recorded in this period of time, with a month's average of 3,208,336 HTTP requests, a daily average of 105,936 requests, and an hourly average of 4,414 requests.

5.1 Human Users' Session Analyses with Cookies

The analysis of users' sessions takes advantage of the information contained in the cookies to reconstruct the sessions. In fact, cookies contain a unique identifier, named *TELSESSID*, assigned at runtime by the PHP[7] interpreter of the Web server for each session started by a user. The use of these identifiers is important for two reasons: we are able to better distinguish users that are hidden behind a proxy, and also to separate human users from the other users of the portal, which in most cases are crawlers and banners. The drawback of this choice is that session reconstruction may be biased by those users which do not enable cookies in their browser and it does not allow to distinguish first time users from users that return many times (with different TELSESSID values).

In the analysed period of time, we were able to construct 209,900 different sessions on the basis of the cookies content.

There is a sizeable 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 first point of analysis that has been raised in Section 3. 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 but they spend some time on the portal, interacting with it (more than 100 HTTP requests) and analysing the results (the majority of sessions last between 2 minutes and 10 minutes). This answers to our first question and confirms the feasibility of Isolated Query Translation. The extra effort required to the user seems to be compensated by his need of information.

5.2 Geographical and Collection Analyses

This part introduces the results on the analysis of sessions. Sessions are classified depending on the different geographical areas from where they have been

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

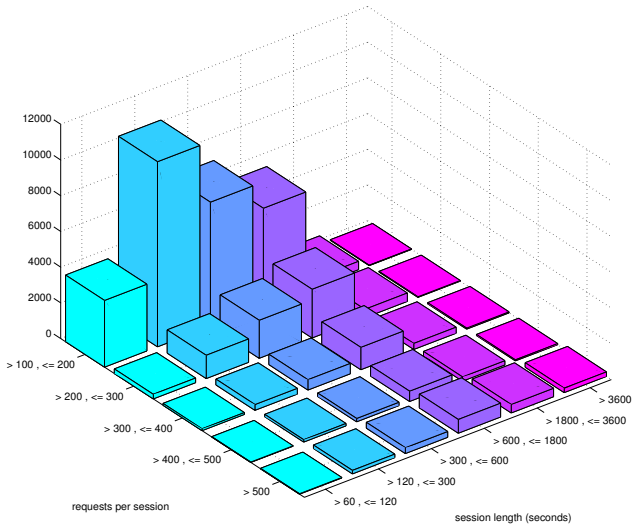

Fig. 1: Sessions (cookies) which last more than 60 seconds with a number of requests per session > 100.

conducted.The analysis focuses on the sessions reconstructed using the cookies with the main aim of identifying the use of the collections which is carried out by the human users to answer questions 2 and 3 in Section 1. The following figures and tables use the three-letter country codes of the ISO 3166-1:2006 standard[8].

The nations with the highest number of sessions reconstructed using the cookies are shown in Figure 2. As expected the nations that interact more with The European Library portal are the European countries, with the exception of the United States. We have also noticed an increase over time of the number of accesses by the new members of the European Union.

When a user enters the service for the first time, a default list of collections is assigned to him. These collections are decided in advance and do not depend on the user (e.g. nationality or language). Last April, this list counted 23 collections, each one from a different country (i.e. SBN OPAC for Italy, Online Catalogue of the National Library of Lithuania for Lithuania). When an user executes a query, the search is performed only over those collections that are supposed to contain the most valuable information. For these reasons, in the first analysis they appear to be the most used.

Figure 3 shows the six collections mostly selected excluding the default list with respect to the country of the user. Nations has been ordered according to the usage of collections a0412. If different users from different nations have had similar behaviour then the graph would have shown a similar ordering for each

[8] `http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=39719`

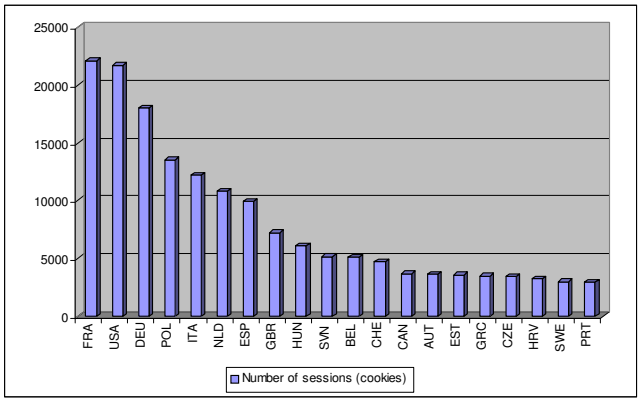

Fig. 2: The nations with the highest numbers of sessions (cookies) are shown.

nation. The presence of many switches in positions for each collection confirms that users coming from different countries require specific collections. In general the language of these collections may differ by the language spoken by the user.

5.3 Discussion

As stated before, in The European Library there are two kinds of information about the user: the nationality and the language. Some initial results about the language selection shows a different behaviour from the one derived from the nationality and it requires a more deep analysis.

What appears to be a limitation of using a single predefined list of collections for all the users is that no information about the user is considered, while it would be interesting to assign, as an example, more collections in the language selected by the user, or on the basis of his nationality. The language is given by the user when he selects the language to use in the portal, while the nationality can be derived from the user IP address.

The information on the collections selected by the user is saved on a specific field of the cookies and the analysis of this variable makes the analysis of the most selected collections possible. In general, we can observe that users usually use the default collection selection instead of explicitly selecting which collections have to be searched.It seems then that there is no distinction on the user behaviour on the basis of their geographical provenance. Our guess is that this is a consequence of portal constraints that do not allow a user to personalize his own query (i.e. it is not possible to execute a query only on collections on some specific language). To overcome this limitation, the analysis focused only on those collections that are explicitly selected by the user and did not considered 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

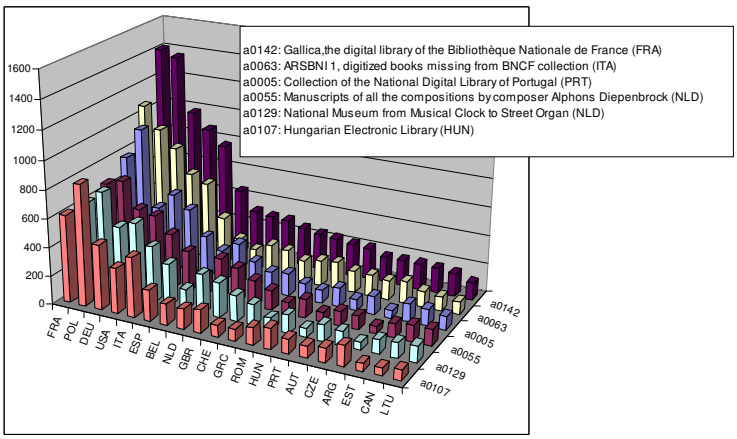

Fig. 3: Most frequently selected collection versus user nationality.

mean of 1,708 selections, and a maximum of 11,000. It appears then 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.

It appears that there is a demand by the users for MLIA solutions but the current implementation of the portal does not allow users to express this need.

The limitations of the portal do not allow to answer general questions like in what languages users search (these data are not present in the HTTP logs) or if English users search only on English collections (it is not possible to restrict a query only on collections in a specific language). What can be answered is: do users navigate the portal displaying data in their mother tongue or do they prefer to use the default language (English)? The HTTP logs showed that the 80% of users do not choose any language, thus they interact with the portal using the default language. This result is quite strange because we have experienced that when an user has the possibility to select his own language it actually does it. We cannot say if this odd behaviour is a consequence of the particular nature of The European Library users or it is a consequence of the consolidate practice to navigate web sites in English language. Moreover, in the 80% of sessions where a language is selected, the users select their own mother tongue.

6 Conclusions

This paper has reported the results of the analysis conducted on the Web logs of The European Library service for MLIA purposes. The main goal of the investigation has been to correlate the nationality, the language, and the collections selected during the search for the design of a multilingual service.

The results gave important hints for the possible design of personalization services to support users with multilingual digital material harvested by The Eu-

ropean Library. To conclude, users behave differently according to their country and mother language, therefore, it is useful for them to have access to advanced services to exploit The European Library resources dynamically. In particular, the isolated query translation appears to be an interesting solution while we were not able to find in which case it would be better to use the pseudo-translation.

References

1. M. Agosti, M. Braschler, and N. Ferro. A Study on how to Enhance TEL with Multilingual Information Access. In *DELOS Research Activities 2006*, pages 115–116. ISTI-CNR at Gruppo ALI, Pisa, Italy, August 2006.
2. M. Agosti, M. Braschler, N. Ferro, C. Peters, and S. Siebinga. Roadmap for MultiLingual Information Access in The European Library. In *Proc. 11th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2007)*. LNCS, Springer, Heidelberg, Germany, 2007 (to be published).
3. M. Agosti and G.M. Di Nunzio. Web Log Mining: A study of user sessions. In *Proc. 10th DELOS Thematic Workshop on Personalized Access, Profile Management, and Context Awareness in Digital Libraries (PersDL 2007)*, pages 70–74, Corfu, Greece, Jun 2007. `http://www.dblab.ntua.gr/persdl2007/papers/72.pdf` [last visited 2007, July 30]
4. M. Agosti, G.M. Di Nunzio, and A. Niero. From Web Log Analysis to Web User Profiling. In *DELOS Conference 2007. Working Notes*, pages 121–132, Pisa, Italy, Feb 2007.
5. L. A. Ballesteros. Cross-Language Retrieval via Transitive Translation. In *Advances in Information Retrieval: Recent Research from the Center for Intelligent Information Retrieval*, pages 203–234. Kluwer Academic Publishers, Norwell (MA), USA, 2000.
6. L. A. Ballesteros and W. B. Croft. Phrasal Translation and Query Expansion Techniques for Cross-language Information Retrieval. In *Proc. 20th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1997)*, pages 84–91. ACM Press, New York, USA, 1997.
7. M. Braschler. Combination Approaches for Multilingual Text Retrieval. *Information Retrieval*, 7(1/2):183–204, 2004.
8. M. Braschler and N. Ferro. Adding MultiLingual Information Access to The European Library. In *DELOS Conference 2007 Working Notes*, pages 39–49. ISTI-CNR, Gruppo ALI, Pisa, Italy, February 2007.
9. M. Braschler, N. Ferro, and J. Verleyen. Implementing MLIA in an existing DL system. In *Proc. International Workshop on New Directions in Multilingual Information Access (MLIA 2006)*, pages 73–76. `http://ucdata.berkeley.edu/sigir2006-mlia.htm` [last visited 2007, March 23], 2006.
10. P. M. Hallam-Baker and B. Behlendorf. Extended Log File Format – W3C Working Draft WD-logfile-960323. `http://www.w3.org/TR/WD-logfile.html` [last visited 2007, June 15], March 1996.
11. T. van Veen and B. Oldroyd. Search and Retrieval in The European Library. A New Approach. *D-Lib Magazine*, 10(2), February 2004. `DOI:10.1045/february2004-vanveen`