# Evaluation of Digital Library Services Using Complementary Logs*

Maristella Agostii
University of Padua
Via Gradenigo 6/a, 35131
Padua, Italy
agosti@dei.unipd.it

Franco Crivellari
University of Padua
Via Gradenigo 6/a, 35131
Padua, Italy
crive@dei.unipd.it

Giorgio Maria Di Nunzio
University of Padua
Via Gradenigo 6/a, 35131
Padua, Italy
dinunzio@dei.unipd.it

## ABSTRACT

In recent years, the importance of log analysis has grown, log data constitute a relevant aspect in the evaluation process of the quality of a digital library system. In this paper, we address the problem of log analysis for complex systems such as digital library systems, and how the analysis of search query logs or Web logs is not sufficient to study users and interpret their preferences. In fact the combination of implicitly and explicitly collected data improves understanding of behavior with respect to the understanding that can be gained by analyzing the sets of data separately.

## Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]: User Issues; H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*Search process*; H.3.4 [**Systems and Software**]: User profiles and alert services

## General Terms

Algorithms, Design, Experimentation

## Keywords

Web Log, Search Log, User Study

## 1. INTRODUCTION

The interaction between the user and an information access system can be analyzed and studied to gather user preferences and to "learn" what the user likes the most, and to use this information to personalize the presentation of results. User preferences can be learned explicitly, for example asking the user to fill-in questionnaires, or implicitly, by studying the actions of the user which are recorded in the search log of a system. The second choice is certainly less intrusive but requires more effort to reconstruct each search session a user made in order to learn his preferences.

*Copyright is held by the author/owner(s).
SIGIR'09, July 19-23, 2009, Boston, USA.

Log is a concept commonly used in computer science; in fact, log data are collected by programs to make a permanent record of events during their usage. The log data can be used to study the usage of a specific application, and to better adapt it to the objectives the users were expecting to reach. In the context of the Web, the storage and the analysis of Web log files are mainly used to gain knowledge on the users and improve the services offered by a Web portal, without the need to bother the users with the explicit collection of information.

When research addresses the problem of studying log data in digital libraries, which are very complex systems, different characteristics regarding library automation systems and digital library systems need to be taken into account. In fact, for all the different categories of users of a digital library system, the quality of services and documents the digital library supplies are very important. Log data constitute a relevant aspect in the evaluation process of the quality of a digital library system and of the quality of interoperability of digital library services [2, 18]. With this concept in mind, it is also possible to think about new different logging formats which reflect how a generic DL system behaves [14].

This paper deals with the study of complementary types of logs in complex systems with the aim of finding new ways of using them to evaluate and personalize digital library services for the final users. The paper is organized as follows: Section 2 presents previous related work, Section 3 analyzes and presents different facets of the study and use of logs of complex systems, Section 4 presents the findings of the case study conducted in the context of the TELplus project[1] for the evaluation and personalization of the services of The European Library, and lastly Section 5 draws conclusions and indicates directions for the continuation of the work.

## 2. RELATED WORK

In the last decade, log analysis has become one of the main threads of research for understanding users of search engines as shown by the works presented at three major relevant conferences and that have been analyzed by us[2].

Those works study logs in different ways and for different

[1] http://www.theeuropeanlibrary.org/telplus/
[2] The three analyzed major conferences are:
SIGIR - http://www.sigir.org/
WWW - http://www.iw3c2.org/
JCDL - http://www.jcdl.org/

purposes, but they can be divided into two main classes: studies about search query logs, and studies about Web server logs. Since most of these research papers concern search engines, the focus of their research is more on improving queries and results and less on surfing the Web. The few exceptions to this classification will be analyzed later in the paper.

Query search logs can be used for: building knowledge, such as automatically building a search thesaurus [10], or acquiring ontological knowledge [24]; refining and expanding queries by means of analysis of search logs [4], or by means of correlations between query terms and document terms based on search query logs [11]; comparing of query extension techniques with pseudo-relevance feedback techniques [30]; organizing search results [29]; studying temporal changes and relationships, such as changes of queries on hourly basis in order to understand how user preferences change over time [5], analysis of multitasking user searches [6], issues related to ambiguity and freshness of queries [22], studies of causal relations between queries [27]; mining queries for extracting news-related queries [20], and association rules to discover related queries [25], or fast query recommendations [32].

Web logs can be used for: improving rank of results by replacing the adjacency matrix of the HITS algorithm with a link matrix which weights connections between nodes based on the usage data from Web server log traffic [21]; matching website organization with visitor expectations by means of Web log analysis [26]; finding user navigational patterns [9]; agents' detection [7].

There is also a recent emerging research activity about log analysis which tackles cross-lingual issues: [13] extends the notion of query suggestion to cross-lingual query suggestion studying search query logs; [16] leverages click-through data to extract query translation pairs. The interest in multilingual log analysis is also confirmed by initiatives promoted by the TrebleCLEF[3] coordination action which supports the development and consolidation of expertise in the multidisciplinary research area of multilingual information access (MLIA).

## 3. LOGS OF COMPLEX SYSTEMS

Present digital library systems are complex software systems, often based on a service-oriented architecture, able to manage complex and diversified collections of digital objects. One significant aspect that still relates present systems to the old ones is that the representation of the content of the digital objects that constitute the collection of interest is still done by professionals. This means that the management of metadata can still be based on the use of *authority control* rules in describing author, place names and other relevant catalogue data. A digital library system can exploit *authority data* that keep lists of preferred or accepted forms of names and all other relevant headings. This is a dramatic difference between digital library systems and search engines, and it is usually overcome with the analysis of log data. In fact a *search engine* often becomes a specific component of a digital library system, when the digital library system faces the management and search of digital objects by content in the same manner as information retrieval systems and search engines [1]. In all other types of searches, either the digital library system makes use of authority data to respond to final users in a more consistent and coherent way through a search system that is a sort of a new generation of online public access catalogue (OPAC) system, or the system supports the full content search with a service that gives the final users the facilities of a search engine.

Search query logs or Web logs alone give only a partial view of the stream of information that users produce. [28] show how to combine two different streams of data, search query logs and click-streams, in order to analyze re-finding behavior of a group of users under observation for a period of one year.

Moreover, log analysis can be supported and validated by user studies which are a valuable method for understanding user behavior in different situations. User studies require a significant amount of time and effort, so an accurate design of the process has to be carried out. In general, user studies and logs are used in a separate way, since they are adopted with different aims in mind. Ingwersen and Järvelin report in [17] that it seems more scientifically informative to combine logs together with observation in naturalistic settings. Pharo and Järvelin in [23] suggest systematic use of the triangulation of different data collection techniques as a general approach in order to get better knowledge of the Web information search process. An example of this type of combined studies is [15], where that authors claim that fully understanding user satisfaction and user intent requires a depth of data unavailable in search query logs but possible to acquire from other sources of data, such as one-on-one studies or instrumented panels.

The combination of implicitly and explicitly collected data improves understanding of behavior with respect to the understanding that can be gained by analyzing the sets of data separately. In particular for digital libraries, where the evaluation of the different services is difficult if logs are used alone, the combined sets of data provide the opportunity of reaching insights towards user personalization of digital library services.

From this starting point we have developed a method for collecting data derived from the user interaction log, "implicit" data, and data collected from user questionnaires, "explicit" data, for analyzing the interaction between users and digital libraries. This means that the conceived method is based on the combination and analysis of the following data sources: HTTP log which contains the HTTP requests sent by the Web client to the Web server during a user browsing session; search log which contains the actions performed by the user during a search; questionnaire data which are collected at the end of a user browsing and searching session.

The possibility of studying and correlating different sources of data was envisaged during the study of the Web portal of The European Library[4], which provides a vast virtual collection of material from all disciplines and offers interested visitors simple access to European cultural heritage.

[3] `http://www.trebleclef.eu/`

[4] `http://www.theeuropeanlibrary.org/`

# 4. RESULTS OF THE CASE STUDY

The European Library is a free service that offers access to the resources of 48 national libraries of Europe in 20 languages with about 150 million entries across Europe. The European Library provides a vast virtual collection of material from all disciplines and offers interested visitors simple access to European cultural heritage.

To validate the proposed method, a study was conducted in a controlled setting at the end of 2007 – beginning of 2008, in the computer laboratories of different faculties of the University of Padua, Italy, where students were requested to conduct a free navigation and search for information on The European Library portal and to fill in a questionnaire specifically designed to harvest the data that can be used to extract information on users satisfaction on the use of different parts of the portal. A total of 155 students participated in the study, mostly Italians, equally distributed between males and females, and with an age range typical of students of Bachelor and Master Degree (in most cases between 19 and 25 years old).

The analysis of the results was done in the following order: the analysis of each stream of data - i.e. HTTP log, search query log, questionnaires - was first conducted, while the analysis of possible interrelation among these sources was conducted later. The description of the analysis of each single stream is reported in [3], here we concentrate on the aspects which emerge from the correlation of the different sources of information.

Table 1 summarizes one of the important features when doing log analysis: session length. In particular, the table shows how different these lengths are according to the source that is analyzed. The "Search log" column shows the statistics of the times, in minutes, of sessions found in the search logs, and between brackets the times of sessions of users who registered to the portal. This shows that logging on is a clear intention of users who are willing to spend time in the portal and search more, compared to random users. The "HTTP log" column shows the times of sessions found in the HTTP logs computed in October 2007, and between brackets the times of the sessions of users who participated in the user study at the University of Padua. In this case, there is a strong bias of the students of the user study due to the time slot which was about 30/45 minutes. The times of random users are comparable to those found in the search logs. The last column shows the times of sessions for filling-in the questionnaires, which are obviously very similar to the times of HTTP sessions of the user study. There is one important aspect which emerges from the data: sessions are very short, browsing and searching activity lasts less than 2 minutes in 50% of the cases. This particular situation can be explained only by studying the answers of the users to the questionnaire where there are clear indications about some difficulties they found in understanding how to read the list of the results, and how to use some functions of the interface. These are also the reasons why they would have left the portal sooner if they had not been asked to stay and fill in the questionnaire.

An important interrelation was found among questionnaires and log data which may explain the short length of a user session. One of the outcomes of the questionnaire was the disorientation of the user upon entering The European Library portal for the first time, in particular it seems not to be clear what kind of information can be accessed through this portal. Users are in general ready to search in a Google-like fashion and obtain documents, in terms of links to pages or documents online, in the case of The European Library they are essentially in front of an online public access catalogue which retrieves bibliographic records. Obtaining library catalogue records after a search is a source of confusion which leaves the user unhappy and willing to leave the portal quickly.

**Table 1: Summary of statistics for the time of a user session in minutes calculated in the search logs (between brackets registered user only), HTTP logs (between brackets user who participated in the study), and the time for filling-in the questionnaire.**

| | Search log | HTTP log | Questionnaire |
|---|---|---|---|
| Median | 2.0 (4.0) | 1.3 (30.25) | 31.0 |
| Mean | 6.0 (8.0) | 4.7 (31.80) | 33.0 |

Questionnaires also show that images in particular seem to be very appealing for users; both the "treasures" section, a section which shows high resolution images of ancient documents, and the "exhibition" section, a section which shows pictures of the national libraries buildings, were thoroughly browsed by users even before making any query in the portal. This is an important clue which may suggest that there should be more linking from the images to the catalogue records. The interrelation among the information about users who prefer images and the HTTP log and searches log is still under investigation. In fact, we would like to see if this willingness expressed in the questionnaire is also reflected in user actions: for example, a user who is interested in images clicks more frequently on images or search for documents like maps or paintings; or a user expresses this interest in images but actually does not perform any action in the portal which confirms this interest.

# 5. CONCLUSIONS

The insights gained by analyzing log data together with data from controlled studies are more informative than the results that can be derived by separately analyzing the groups of data. Our studies on logs combined with interviews have shown that the results are more scientifically informative than those obtained when the two types of studies are conducted alone. This encouraging result constitutes the ground on which we are generalizing and formalizing starting from the obtained results. A crucial feature in the future will be making active use also of the information on metadata that are present in the log, because until now no active way of using them has been incorporated in the proposed method.

# 6. ACKNOWLEDGEMENTS

The work has been partially supported by the TELplus Targeted Project for digital libraries, as part of the eContent*plus* Program of the EC, and by the TrebleCLEF Coordination Action, as part of the 7FP of the EC.

# 7. REFERENCES

[1] M. Agosti, editor. *Information access through search engines and digital libraries*. Springer, Berlin, Germany, 2008.

[2] M. Agosti. Log data in digital libraries. In M. Agosti, F. Esposito, and C. Thanos, editors, *IRCDL*, pages 115–122. DELOS: an Association for Digital Libraries, 2008.

[3] M. Agosti, F. Crivellari, and G. M. Di Nunzio. A method for combining and analyzing implicit interaction data and explicit preferences of users. Workshop on Contextual Information Access, Seeking and Retrieval Evaluation (ECIR 2009), April 2009.

[4] P. G. Anick. Using terminological feedback for web search refinement: a log-based study. In *SIGIR*, pages 88–95. ACM, 2003.

[5] S. M. Beitzel, E. C. Jensen, A. Chowdhury, D. A. Grossman, and O. Frieder. Hourly analysis of a very large topically categorized web query log. In M. Sanderson, K. Järvelin, J. Allan, and P. Bruza, editors, *SIGIR*, pages 321–328. ACM, 2004.

[6] N. Buzikashvili. An exploratory web log study of multitasking. In Efthimiadis et al. [12], pages 623–624.

[7] N. Buzikashvili. Sliding window technique for the web log analysis. In Williamson et al. [31], pages 1213–1214.

[8] L. Carr, D. D. Roure, A. Iyengar, C. A. Goble, and M. Dahlin, editors. *Proceedings of the 15th international conference on World Wide Web, WWW 2006, Edinburgh, Scotland, UK, May 23-26, 2006*. ACM, 2006.

[9] J. Chen and T. Cook. Mining contiguous sequential patterns from web logs. In Williamson et al. [31], pages 1177–1178.

[10] S.-L. Chuang, H.-T. Pu, W.-H. Lu, and L.-F. Chien. Auto-construction of a live thesaurus from search term logs for interactive web search. In *SIGIR*, pages 334–336, 2000.

[11] H. Cui, J.-R. Wen, J.-Y. Nie, and W.-Y. Ma. Probabilistic query expansion using query logs. In *WWW 2002*, pages 325–332, 2002.

[12] E. N. Efthimiadis, S. T. Dumais, D. Hawking, and K. Järvelin, editors. *SIGIR 2006: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, 2006*. ACM, 2006.

[13] W. Gao, C. Niu, J.-Y. Nie, M. Zhou, J. Hu, K.-F. Wong, and H.-W. Hon. Cross-lingual query suggestion using query logs of different languages. In Kraaij et al. [19], pages 463–470.

[14] M. A. Gonçalves, G. Panchanathan, U. Ravindranathan, A. Krowne, E. A. Fox, F. Jagodzinski, and L. N. Cassel. The xml log standard for digital libraries: Analysis, evolution, and deployment. In *JCDL*, pages 312–314. IEEE Computer Society, 2003.

[15] C. Grimes, D. Tang, and D. M. Russell. Query logs alone are not enough. In E. Amitay and C. G. M. J. Teevan, editors, *Query Log Analysis: Social And Technological Challenges. A workshop at the 16th International World Wide Web Conference (WWW 2007)*, May 2007.

[16] R. Hu, W. Chen, P. Bai, Y. Lu, Z. Chen, and Q. Yang. Web query translation via web log mining. In S.-H. Myaeng, D. W. Oard, F. Sebastiani, T.-S. Chua, and M.-K. Leong, editors, *SIGIR*, pages 749–750. ACM, 2008.

[17] P. Ingwersen and K. Järvelin. *The Turn*. Springer, The Netherlands, 2005.

[18] T. Koch, A. Ardö, and K. Golub. Browsing and searching behavior in the renardus web service a study based on log analysis. In H. Chen, H. D. Wactlar, C. chih Chen, E.-P. Lim, and M. G. Christel, editors, *JCDL*, page 378. ACM, 2004.

[19] W. Kraaij, A. P. de Vries, C. L. A. Clarke, N. Fuhr, and N. Kando, editors. *SIGIR 2007: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, July 23-27, 2007*. ACM, 2007.

[20] M. Maslov, A. Golovko, I. Segalovich, and P. Braslavski. Extracting news-related queries from web query log. In Carr et al. [8], pages 931–932.

[21] J. C. Miller, G. Rae, and F. Schaefer. Modifications of kleinberg's hits algorithm using matrix exponentiation and weblog records. In W. B. Croft, D. J. Harper, D. H. Kraft, and J. Zobel, editors, *SIGIR*, pages 444–445. ACM, 2001.

[22] J. Parikh and S. Kapur. Unity: relevance feedback using user query logs. In Efthimiadis et al. [12], pages 689–690.

[23] N. Pharo and K. Järvelin. The SST method: a tool for analysing Web information search processes. *Information Processing & Management*, 40(4):633–654, July 2004.

[24] S. Sekine and H. Suzuki. Acquiring ontological knowledge from query logs. In Williamson et al. [31], pages 1223–1224.

[25] X. Shi and C. C. Yang. Mining related queries from search engine query logs. In Carr et al. [8], pages 943–944.

[26] R. Srikant and Y. Yang. Mining web logs to improve website organization. In *WWW 2001*, pages 430–437, 2001.

[27] Y. Sun, K. Xie, N. Liu, S. Yan, B. Zhang, and Z. Chen. Causal relation of queries from temporal logs. In Williamson et al. [31], pages 1141–1142.

[28] J. Teevan, E. Adar, R. Jones, and M. A. S. Potts. Information re-retrieval: repeat queries in yahoo's logs. In Kraaij et al. [19], pages 151–158.

[29] X. Wang and C. Zhai. Learn from web search logs to organize search results. In Kraaij et al. [19], pages 87–94.

[30] R. W. White, C. L. A. Clarke, and S. Cucerzan. Comparing query logs and pseudo-relevance feedbackfor web-search query refinement. In Kraaij et al. [19], pages 831–832.

[31] C. L. Williamson, M. E. Zurko, P. F. Patel-Schneider, and P. J. Shenoy, editors. *Proceedings of the 16th International Conference on World Wide Web, WWW 2007, Banff, Alberta, Canada, 2007*. ACM, 2007.

[32] Z. Zhang and O. Nasraoui. Mining search engine query logs for query recommendation. In Carr et al. [8], pages 1039–1040.