Chapter 67
User's Behaviour inside a Digital Library

Marco Scarnò
Inter-University Consortium for SuperComputing, CASPUR, Italy

ABSTRACT

CASPUR allows many academic Italian institutions located in the Centre-South of Italy to access more than 7 million articles through a digital library platform. The behaviour of its users were analyzed by considering their "traces", which are stored in the web server log file. Using several web mining and data mining techniques the author discovered a gradual and dynamic change in the way articles are accessed. In particular there is evidence of a journal browsing increase in comparison to the searching mode. Such phenomenon were interpreted using the idea that browsing better meets the needs of users when they want to keep abreast about the latest advances in their scientific field, in comparison to a more generic searching inside the digital library.

INTRODUCTION

The CASPUR Consortium was established on June 5th, 1992; its name comes from the acronym: Inter-University Consortium for the Application of Super-Computing for Universities and Research. The Consortium headquarter is in Rome, Italy.

CASPUR is a no-profit Organization; it is financed by MIUR (the Ministry for Education, Universities and Research) and by associated Universities (mainly located in the Centre-South of Italy).

CASPUR main purposes are:

- To manage a center capable of guarantee a high quality and high-powered processing service;
- To promote the use of the most advanced information processing systems;
- To become a center of excellence available to the national university and research network and to MIUR, with the aim of spreading the culture of information and communication technology;

DOI: 10.4018/978-1-4666-2455-9.ch067

- Developing research programs aiming at a more effective and innovative usage of information and communication technology, in collaboration with other organizations and enterprises;

In the field of virtual newspaper and periodical library, CASPUR allows many users (mainly coming from academic Italian institutions) to access to over 5200 academic and scientific full-text periodicals and over 7.5 millions pdf articles (last update: January 2009).

Journals are available dating the nineties; they cover all fields and are issued by different publishers and professional societies, including, for example, the American Chemical Society, Blackwell Publishing, Elsevier Science, Institute of Physics Publishing, Kluwer Academic Publisher, Springer.

This service is accessible from a web site (periodici.caspur.it) and its main advantage consists in the possibility of allowing research (also personalized) in different fields (author, title, keyword or full-text words) within the entire series. In this way users can refer to a title list arranged by publisher, class or alphabetical order, made possible by an homogeneous interface based on web-usability criteria.

Users access to the service and to the research function through a web client; this access is restricted to authorized Institutes and Universities through a procedure that checks the IP address or by considering a username and a password, which would allow the access to the virtual library from anywhere.

The virtual library service is based on Science Server software, and supplied by three Linux servers, indistinguishable by the final user. UltraATA disk strips (on 2 Gbps fiberchannel interface) form the disk space on which software, metadata and the indexes' database are installed, for a total of 14 TB. Of these, 8 TB are dedicated to the online system, and the others are a copy of it, necessary to the whole system data backup

The idea of this study is to describe the behavior of the users by considering and analyzing their traces stored into the web server log file. The analysis of such logs can provide an insight about searching behavior on digital library and about Information Retrieval. It has to be noticed that the first in-depth studies on query logs date back to the late 1990s; see, for example, Jansen (1998, 2000) and Spink (2001). But, for what concerns the use of these files in a digital libraries context, there are less studies; see Wolfram (2002).

MATERIALS AND METHODS

Data were collected by considering the web logs coming from those users that accessed to the digital library using a username and password; this facilitates the need to identify all the distinct search sessions (see further in the next paragraph).

In particular data are referred to the time interval between January 2006 and January 2009; the records belonging to these 37 months and contained into the web log are more than 10 millions. Note that the users that accessed to the service with such type of authentication are only a small part (less than the 10%) of the total ones.

Such huge amount of data can be significant reduced by considering that each record in the log file represents an "object" that was returned to the user browser after a query; this object can be, for example, a full text, another html page with the resulting articles obtained by a search request. But a record can, also, stores the information that an image file was passed to the client browser (this is very common when the result of the query is an html page that contains logos, buttons represented by using jpeg or gif files).

So the really significant records are only a small part of the original ones; in order to represent the real interactions between the users (i.e., their queries) and the service answers, the resulting data set has about 1.2 millions of total records.