User's behaviour inside a digital library

Marco Scarnò[1]

[1]Inter-University Consortium for SuperComputing, CASPUR, Rome, Italy
(mscarno@caspur.it)

Abstract: CASPUR allows many academic Italian institutions located in the Centre-South of Italy to access more than 7 million of articles through a digital library platform. We analyzed the behaviour of its users by considering their "traces" stored into the web server log file. Using several Web Mining and Data Mining techniques we discovered that there is a gradual and dynamic change in the way how articles are accessed; in particular there is evidence of a Journal browsing increase in comparison to the searching mode. We interpreted such phenomenon by considering that browsing better meets the need 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.

Keywords: Digital Library, Web mining, Web server log file, Data Mining, user's behaviour, search engine, Journal browsing

1. 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;
- 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).

2. 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.
These can be used firstly to verify the number of distinct users that, monthly, accessed the web site periodici.caspur.it (fig. 1); it is possible to observe that there is a growing trend between the 37 months. There could be two possible explanations of this growth: the first is related to the general increasing of the users of the digital library, the second to the user's awareness of the facilities offered by the authentication to the service.

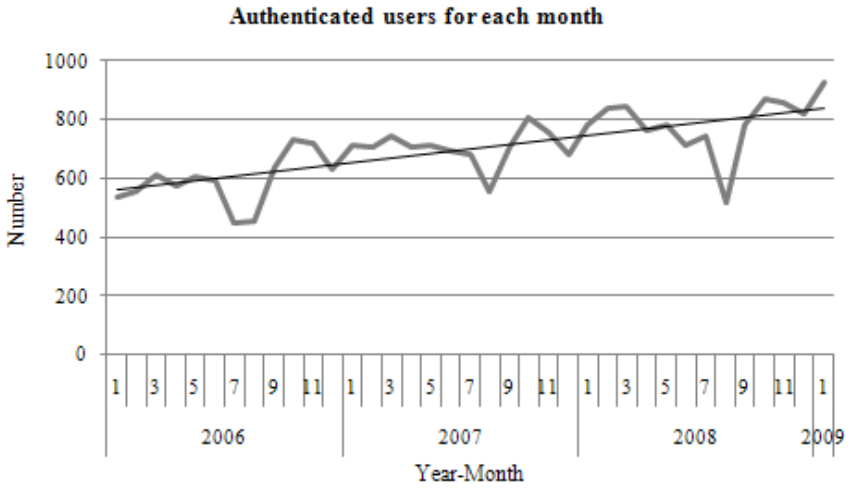

Fig. 1: plot of authenticated users for each month in the period January 2006-January 2009 (with its linear trend)

For what concerns a significant record, this could be identified by considering one of the following situations:

- the user downloaded a full text (represented by a pdf or an html file);
- the user requested a list of articles by using:
    - a "simple search", i.e. a query in which a given term can be searched in one of the fields that are associated to an article (like the title, the abstract, the author name, ISSN, author keywords or journal title);
    - an "advanced search", in which one or more terms can be searched in one or more fields;
    - an "expert search" in which a boolean search expression is entered directly.
- the user "browsed" the content of the digital library by referring to the alphabetical, category or publisher list of journals, then to the volume, the issue and, at last, to the desired full text.

Note that, according to what permitted by the web site (periodici.caspur.it), operations like "bookmark", "keep me informed on new issues" and similar, generate results that belong to the browse action.

Obviously the download action is the consequence of the search one; this could be associated to a sort of "strategy" that represent the "movements" of the users inside the digital libraries, i.e. their behaviors in the information retrieval process.

It is important to observe that the contraposition between browsing and searching is one of the arguments that has become widely discussed in the science and publishing communities after a recent article by Evans (2008). Evans' in-depth research on citations in over 34 million articles and how online availability affects citing patterns; he found that the more issues of a journal that are available online, the fewer numbers of articles in that journal are cited. If the journal is available for free online, it is cited even less. Evans attributes this phenomenon to more searching and less browsing.

Tenopir (2009) observed that the actual numbers of articles found by browsing has not decreased much, even though the percentage of readings found by searching has increased.

It should be noticed that all their results come from “direct surveys”, i.e. on data collected by using questionnaires; an approach based on the web log analysis is, instead, “indirect” because it tries to reconstruct the user’s needs by observing their actions.
Web log analysis requires an information processing tool that can treat huge amount of records and that can help to clean and verify the data. The software used in the framework of this study is ADaMSoft, an Open Source package developed by the Consortium CASPUR. It contains some web mining methods, like the one that can be used to transform a web log into an usable data set.
For what concerns the data cleaning process, after the deletion of all the not significant records, there was the needs to recognize and to delete all those “nonsense” records, like the ones derived by the “double click” actions that result in two records but belong to the same action.
The next analytical step related to the web log analysis implemented in this study was the necessity to identify the different user sessions.

3. Session detection

The users of a digital library access to these with the aim of reading one or more articles; Tenopir, (2009, cit.) observed that the most frequent principal purpose of reading is research (48.5% of readings), followed by teaching (22.5%), writing (articles, etc., 10.8%), and current awareness/keeping up (8.0%).
The process of arriving to those articles that will be read is determined by the digital library appearance that, in this case, permit to browse its content or to use the search engine.
According to Swanson (1977), searching can be viewed as a trial-and-error process in which a query is a guess about the attributes a desired document is expected to have and the response of the system is then used to correct the initial guess for another try.
That way the users gradually refine both their queries and their goals. Spink et al. (1998) referred to this process as a “successive search phenomenon” and defined it as: "The process of repeatedly searching over time in relation to a specific, but possibly an evolving information problem."
As a consequence, when users interact with a search engine in order to achieve their goals, they produce a sequence of queries able of being recorded and subsequently analyzed.
Hence, a session from a search engine point of view can be:

- the whole sequence of queries issued by one user during one single day;
- the sequence of queries issued by one user since s/he starts the browser until s/he quits;
- a sequence of queries with no more than a few minutes of inactivity between them.

There is not a general consensus about the “session” concept in the literature; see Gayo-Avello (2009). The first clearly stated definition of this concept referred to a search engines is possibly that of Silverstein (1999): "A session is a series of queries by a single user made within a small range of time; a session is meant to capture a single user’s attempt to fill a single information need."
Under this assumption it is clear that could be identified more than one session in a single day for each authenticated user.
Unfortunately specific session identifiers were not available for the search engine and session boundary detection was, therefore, not as obvious.
Methods for session boundary detection have been proposed based on content

analysis of queries or timing characteristics of queries. Subject analysis can be performed automatically or manually. Manual analysis can be impractical for large datasets, whereas automatic approaches may be unreliable for short queries, and might not take into account whether users engage in multiple search topics in a given session.

The second method for session boundary detection, which relies on timing characteristics, considers temporal cut-off points or probabilistic characteristics of the datasets. The main drawback of this method is that the boundary detection method usually does not take into account the subject content of queries, relying more on temporal patterns of query submission.

Murray et al. (2006), for example, relied on timing characteristics that assume a minimum of 20 queries per authenticated user from which large gaps in inter-query times.

General cut-off values based on qualitative assessment of query sets to delineate sessions have been more widely used. Spink and Jansen (2004) concluded that most Web search sessions last about 15 min, with a substantial percentage lasting less than 5 min (p. 121). Similarly Goker (2002) suggested that an optimal session boundary interval was 11-15 min.

The method for session boundary determination used in the present study considered the distribution of the time between two temporally adjacent queries associated with the same user (in fig. 2 is the distribution of such time as evaluated on the considered data).

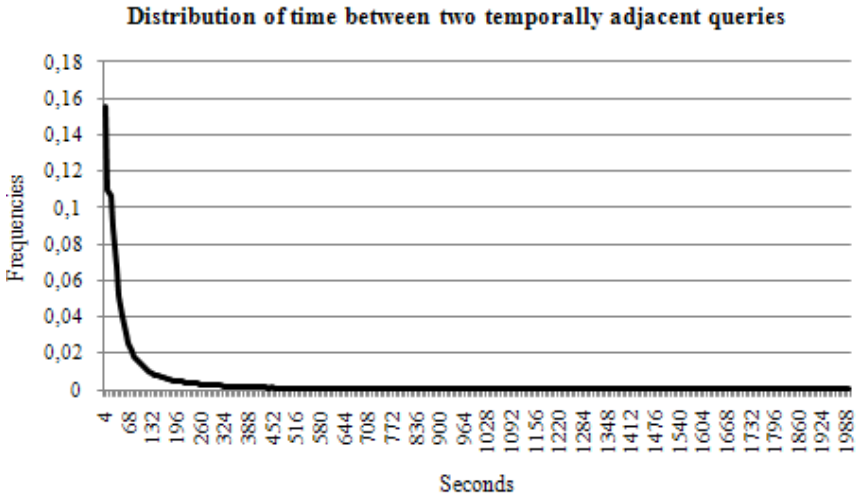

Fig. 2: distribution of time between two temporally adjacent queries for all the authenticated user in the period January 2006-January 2009.

Although there was a gradual decline in the distribution of the times between the queries, there was not a clear cut-off point. It is interesting to observe that in the 83% of the cases a query follows the previous one in less than 3 minutes, while in 93% of the cases the inter-query time is less than 11 minutes.

An arbitrary assignment of a cut-off point based on past studies could not take into account the different behaviour between each user, i.e. the possibility that one user is “faster” than another in making queries.

To this purpose the strategy used in this study evaluated, for each user and for each day, the maximum time that occurred between two adjacent queries; in this way there will be many values of this statistic, for each user, that refers to the days in which they used the service. Then this statistic was finally

synthesized by considering its minimum value. The result is a cut-off time value "personalized" for each user. In particular the mean value, for all the users, of the evaluated cut-off is equal to 330 seconds (5 minutes and 30 seconds). At this point it was possible to describe the behaviours of the users from the moment in which they start their search.

5. User's behaviour

The number of the identified sessions and of the full texts viewed (with their linear trends) in the months between January 2006 and January 2008 is displayed in fig. 3.

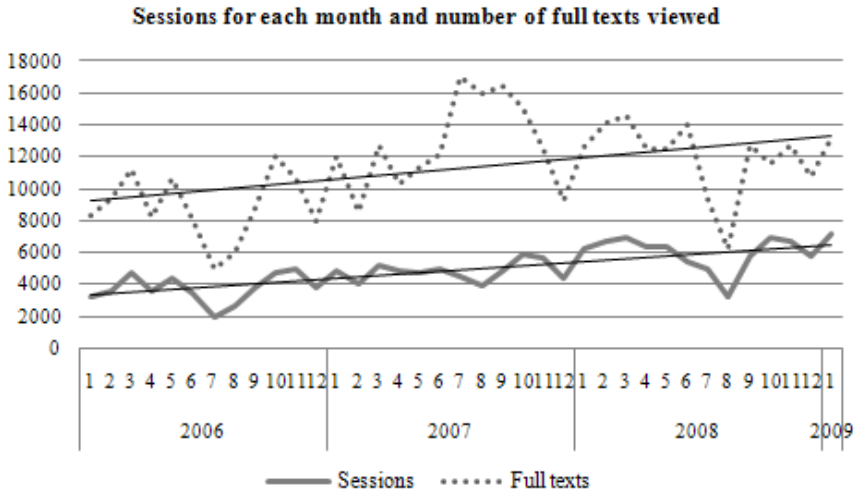

Fig. 3: plot of number of sessions and of full texts for each month in the period January 2006-January 2009 (with their linear trends)

There were a monthly mean of about 5 thousand (4889) sessions in the 37 months, to which corresponds a mean of 11 thousand (11259) of full texts viewed (this implies that for each session were viewed a mean of 2.3 full texts). For what concerns the first step of the "strategy" adopted by the users to interact with the digital library, the results showed that:

- 41% of times they used the *simple search*;
- 22% of times they used the *advanced search*;
- 1% of times they used the *expert search*;
- 35% of times they *browsed* the DL.

For what concerns the way in which the strategy evolves inside a session, it is possible to consider the *transition table* reported in table 1, where the rows are related to the previous step and the column to the following ones.

	Simple search	Adv. search	Exp. search	Browsing	Total
Simple search	0.91	0.04	0.00	0.05	221804
Adv. Search	0.07	0.83	0.01	0.09	150886
Exp. Search	0.07	0.09	0.79	0.05	7533
Browsing	0.06	0.07	0.00	0.87	206190

Tab. 1: transition table between one step (row) and the following ones (columns); note that inside the table the values represent the row relative frequencies, while the column total contains the real total values

It is interesting to see the first step along each month, in order also to verify its trend (fig. 4); in this case it could be used the ratio between the number of the four considered strategies with the number of session. This in order to avoid "false effects" coming from the growth of the sessions during the period.

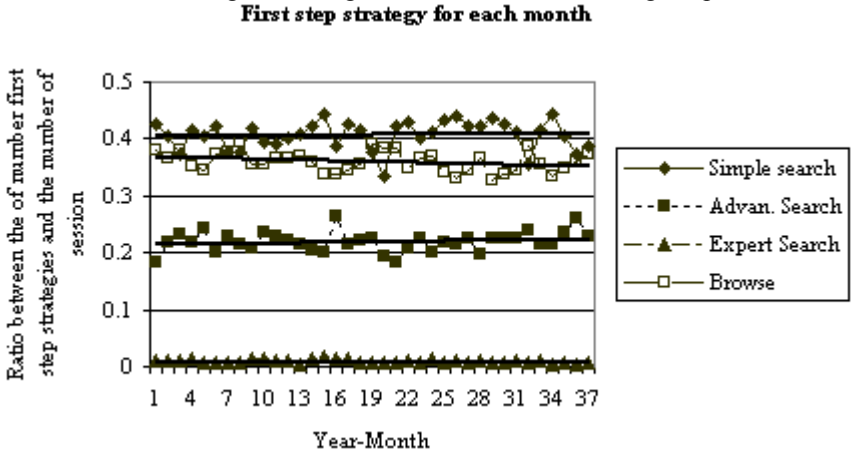

Fig. 4: plot of first step strategy for each month in the period January 2006-January 2009 (with their linear trends); note that with *1* it is indicated the first month in the period, with *2* the second one, etc.

The average number of articles that the users "browsed" in the resulting pages obtained with a search step was equals to 30.
It is, also, interesting to analyze the trend of the total search steps and of the browsing one, for each month (fig. 5).

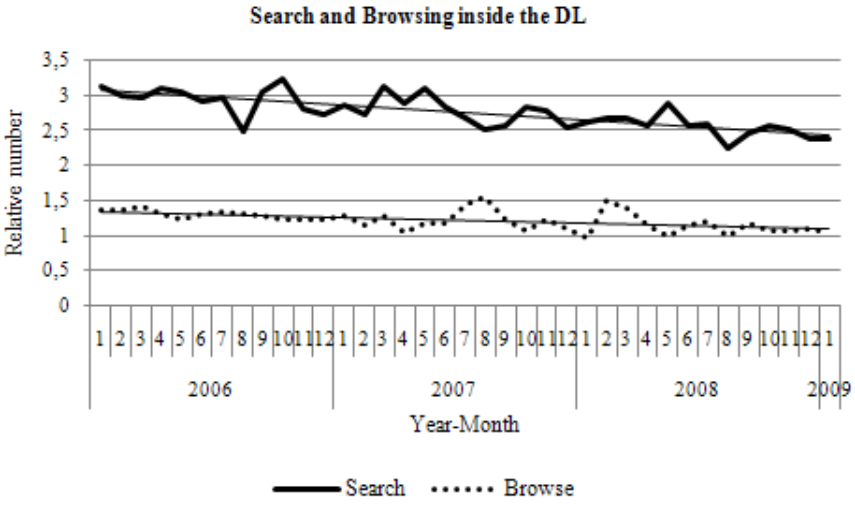

Fig. 5: number of search and browse steps inside the digital library for each month (ratio with the number of sessions)

4. Conclusions

During the Information Retrieval process that can be referred to a digital library, the users mostly prefer to search than to browse its content.
In fact they usually start the interaction with the DL by interacting with the search engine; inside each session (i.e. a single user's attempt to fill a single information need) there are an average of 2.3 full texts viewed on a mean of 30

that were "browsed" as the results of the search actions.
By considering the different kind of strategies offered by the search engine (simple, advanced and expert), the users prefer the simple one; anyway they tend to repeat the previously used search action (maybe refining it).
For what concerns the comparison between the "search" and the "browse" action, there is an evidence that shows how these are decreasing over the considered months, but the Journal browsing has a lower trend than the searching mode. This could be justified by considering that the studied sample of users tends to become aware of the possibilities given by the service, that help them to build "an environment" in which they can better find what they need (using, for example, personalized link to Journals, to previously saved searches, etc.).

References

Evans J. A., (2008). Electronic Publication and the Narrowing of Science and Scholarship. *Science 321*, no. 5887, 395--399.

Gayo-Avello D. (2009), A survey on session detection methods in query logs and a proposal for future evaluation, *Information Sciences*, Vol. 179, issue 12, 1822—1843

Goker, A., & He, D. (2002). Analysing Web search logs to determine session boundaries for user-oriented learning, *Proceedings of the international conference on adaptive hypermedia and adaptive Web-based systems*, London: Springer-Verlag, 319--322

Jansen B.J., et al., (1998). Real life information retrieval: a study of user queries on the Web, *ACM SIGIR Forum*, Volume: 32, 5--17

Jansen B.J., Spink A., (2000). Methodological approach in discovering user search patterns through Web log analysis, *ACM SIGIR Forum*, Vol. 32, 5—17

Murray G. C., Lin, A., Chowdhury A. (2006). Identification of user sessions with hierarchical agglomerative clustering, *Proceedings of the ASIS&T annual meeting* [CD-ROM]. Medford, NJ: Information Today, Inc.

Silverstein C., et al. (1999). Analysis of a very large Web search engine query log, *ACM SIGIR Forum*, Vol.: 33, 6--12

Spink A. et al. (1998). Modelling users' successive searches in digital environments, D-Lib Magazine

Spink A., et al., (2001). Searching the Web: The public and their queries, *Journal of the American Society for Information Science and Technology*, Vol.52, 226—234

Spink, A., Jansen B. J. (2004). Web search: Public searching of the Web, *Dordrecht: Kluwer eds.*

Swanson D.R. (1977). Information retrieval as a trial-and-error process, *Library Quarterly*, Vol.: 47, 128--148

Tenopir C, King D. W., Edwards S., Wu L. (2009). Electronic Journals and Changes in Scholarly Article Seeking and Reading Patterns, *Aslib Proceedings: New Information Perspectives*, 61 (1), 5-32

Wolfram D., Xie H., (2002). Traditional IR for web users: a context for general audience digital libraries, *Information Processing and Management*, Vol. 38, 627--648