



UNIVERSITY OF PADUA
DEPARTMENT OF INFORMATION ENGINEERING

DEPARTMENT OF
INFORMATION
ENGINEERING
UNIVERSITY OF PADOVA



PersDL 2007
10th DELOS Thematic Workshop on
Personalized Access, Profile Management, and Context Awareness in Digital Libraries
Corfu, Greece, 29–30 June 2007

Web Log Mining: A Study of User Sessions

Maristella Agosti
agosti@dei.unipd.it

Giorgio Maria Di Nunzio
dinunzio@dei.unipd.it

Information Management Systems Research Group



Outline

- ▶ Motivations;
- ▶ Approach;
- ▶ Experimental Analysis;
- ▶ Current and Future Works.

Motivations

- ▶ The three main approaches to evaluate an information access service are:
 - ▷ the studies based on test collection analysis, the so-called “*Cranfield approach*”¹,
 - ▷ the user studies, and
 - ▷ the analysis of log data.

- ▶ Web log file analysis began with the purpose to offer to Web site administrators a way to ensure adequate bandwidth and server capacity to their organization.

- ▶ It may offer advices about
 - ▷ a better way to improve the offer of Web content,
 - ▷ information about problems occurred to the users,
 - ▷ and even about problems for the security of the site.

¹C. W. Cleverdon “The Cranfield Tests on Index Languages Devices”. In Readings in Information Retrieval, Morgan Kaufmann Publisher, Inc., San Francisco, California, pp.47–60, 1997.

The European Library

- ▶ Case study: The European Library ², a service set up by the Conference of the European National Librarians (CENL)³.
- ▶ Born to offer access to combined resources (such as books, magazines, and journals – both digital and non-digital) of 47 national libraries of Europe.
- ▶ Analyse the data contained in the logs of their Web servers.

²<http://www.theeuropeanlibrary.org/>

³<http://www.nlib.ee/cenl/>

Scope of the study

- ▶ Evaluate the information access service to give recommendations for developing possible future personalization services.
- ▶ Report on initial findings on a specific aspect that is highly relevant for personalization:
 - ▷ the study of user sessions.
- ▶ The log data used for the present analysis refer to the collections from 27 out of the 47 national libraries that were full partners at the moment of the analysis.

Approach

- ▶ Information extracted from Web has to be managed efficiently for analyses.
- ▶ The proposed solution is based on database management methods which permit the definition of an application which maintains and manages the necessary data⁴.
- ▶ The database and the application which have been developed enable separation of the different entities recorded and facilitate data-mining and on-demand querying of the log data.

⁴M. Agosti, G.M. Di Nunzio and A. Niero “From Web Log Analysis to Web User Profiling” In DELOS Conference 2007. Working Notes. Pisa, Italy, 2007, pp 121–132.

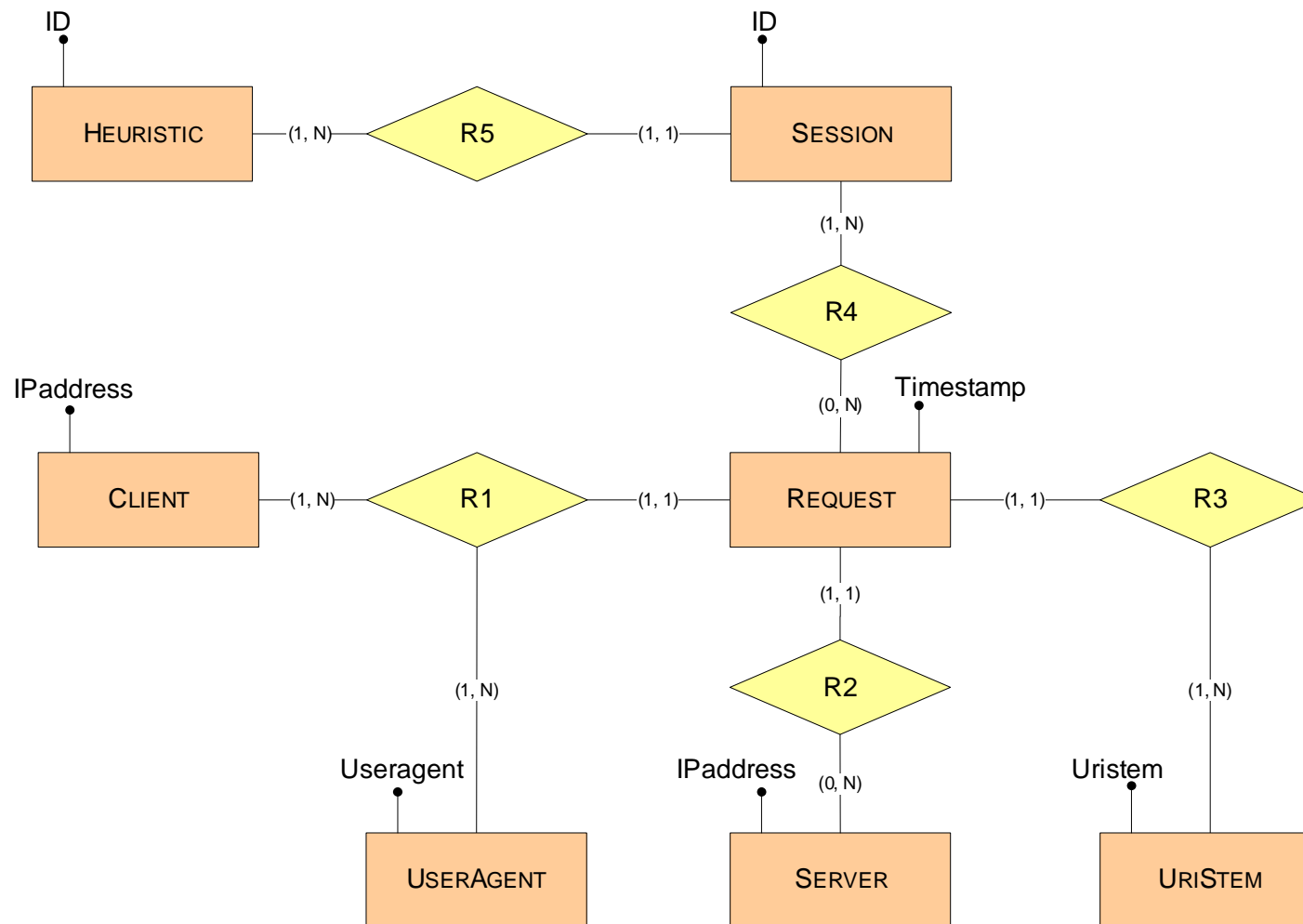
Methodology

- ▶ Usually Log files come in a text file format.
- ▶ The proposed methodology for acquiring data from Web log files identifies two problems:
 - ▷ gathering, and
 - ▷ storing the information.
- ▶ Gathering data with parsers;
- ▶ Storing data with databases.

Gathering Data

- ▶ Web log files have ordered fields to record activities:
 - ▷ *date*: Date, in the form of yyyy-mm-dd.
 - ▷ *time*: Time, in the form of hh:mm:ss.
 - ▷ *s-ip*: The IP of the server.
 - ▷ *cs-method*: The requested action. Usually GET for common users.
 - ▷ *cs-uri-stem*: The URI-Stem of the request.
 - ▷ *cs-uri-query*: The URI-Query, where requested.
 - ▷ *s-port*: The port of the server for the transaction.
 - ▷ *cs-username*: The username for identification of the user.
 - ▷ *c-ip*: The IP address of the client.
 - ▷ *cs(User-Agent)*: User-Agent of the Client. For a standard user this means the browser and other information about operative system.
 - ▷ *cs(Referer)*: The site where the link followed by the user was located.
 - ▷ *sc-status*: HTTP status of the request, that means the response of the server.
 - ▷ *sc-substatus*: The substatus error code.
 - ▷ *sc-win32-status*: The Windows status code.

Storing Data



Approach - HTTP Requests

- ▶ We concentrate our attention here only on deriving the information on user sessions from the analysis of the
 - ▶ HyperText Transfer Protocol (HTTP) requests made by clients,
 - ▶ grouped in sessions,
 - ▶ using a specific heuristic.

- ▶ A request represents the data of the HTTP request that are recorded in the Web log files.

```
2005-11-30 23:00:37 192.87.31.35 GET /index.htm - 80 - 152.xxx.xxx.xxx Mozilla/4.0+(compatible; ...
```

Approach - Sessions

- ▶ A session is a particular set of requests made in a certain interval of time by the same client.

```
2005-11-30 23:00:37 192.87.31.35 GET /index.htm - 80 - 152.xxx.xxx.xxx Mozilla/4.0+(compatible; ...
2005-11-30 23:00:38 192.87.31.35 GET /portal/index.htm - 80 - 152.xxx.xxx.xxx Mozilla/4.0+( ...
2005-11-30 23:00:38 192.87.31.35 GET /portal/scripts/Hashtable.js - 80 - 152.xxx.xxx.xxx Mozilla/4.0+ .
2005-11-30 23:00:44 192.87.31.35 GET /portal/scripts/Session.js - 80 - 152.xxx.xxx.xxx Mozilla/4.0+ ...
2005-11-30 23:00:46 192.87.31.35 GET /portal/scripts/Query.js - 80 - 152.xxx.xxx.xxx Mozilla/4.0+ ...
2005-11-30 23:00:47 192.87.31.35 GET /portal/scripts/Search.js - 80 - 152.xxx.xxx.xxx Mozilla/4.0+ ...
```

- ▶ Organizing the HTTP requests in a single session permits to have a better view of the actions performed by visitors.
- ▶ Sessions are found through empirical rules when information about sessions are not available.

Approach - Sessions' Reconstruction

- ▶ “Session reconstruction” may be used in order to map the list of activities performed by every single user to the visitors of the site.
- ▶ Possible choices:
 - ▷ the IP address and the user-agent are the same of the requests already inserted in the session⁵,
 - ▷ the request is done less than fifteen minutes after the last request inserted⁶.

⁵D. Nicholas, P. Huntington, A. Watkinson “Scholarly journal usage: the results of deep log analysis”, *Journal of Documentation* Vol. 61 No. 2, 2005.

⁶B. Berendt, B. Mobasher, M. Nakagawa, M. Spiliopoulou “The Impact of Site Structure and User Environment on Session Reconstruction in Web Usage Analysis”, *WEBKDD 2002*, LNAI 2703, pp 159-179, 2003.

Experimental Analyses

- ▶ Experimental analysis was performed on a sample of The European Library Web log files of eleven months
 - ▷ from 31st October 2005,
 - ▷ to 25th September 2006.
- ▶ The structure of the log file record is conform to the W3C Extended Log File Format⁷.
- ▶ The analyses, that are presented in the following, cover software tools such as operating systems and browsers used by clients, sessions in terms of daily distribution, and time intervals per number of HTTP requests.
- ▶ The numbers we are reporting include all the requests and sessions, even those ones which can belong to automatic crawlers and spiders.

⁷<http://www.w3.org/TR/WD-logfile.html>.

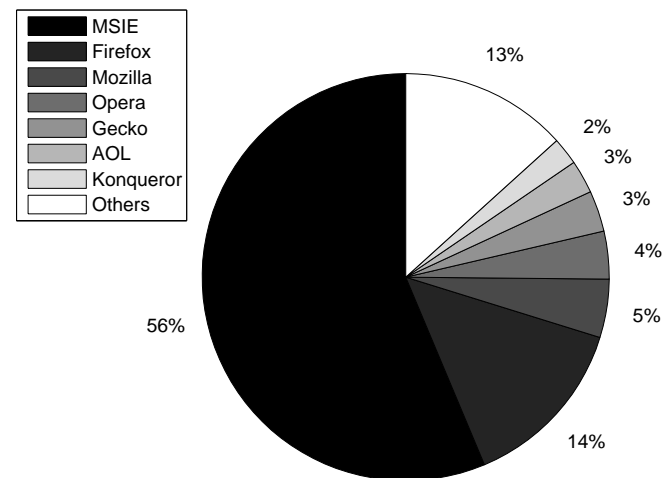
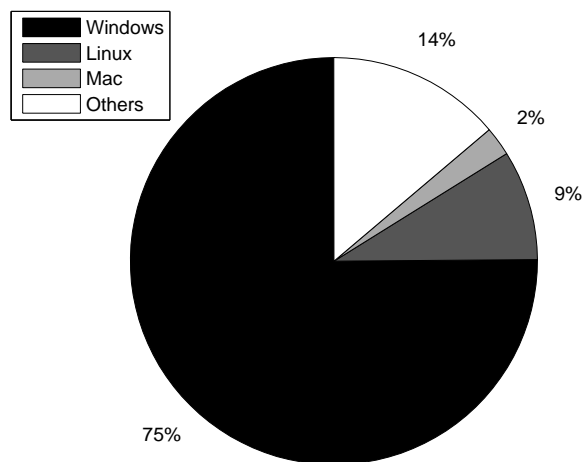
Experimental Analyses - HTTP Requests

- ▶ A total of 25,881,469 of HTTP requests were recorded in the log files of the eleven months.
- ▶ The distribution of HTTP methods which are present in the log files is the following

HTTP method	total number
CONNECT	2
LINK	6
PROPFIND	760
PUT	3,640
OPTIONS	3,779
HEAD	33,770
POST	844,058
GET	24,995,454
Total	25,881,469

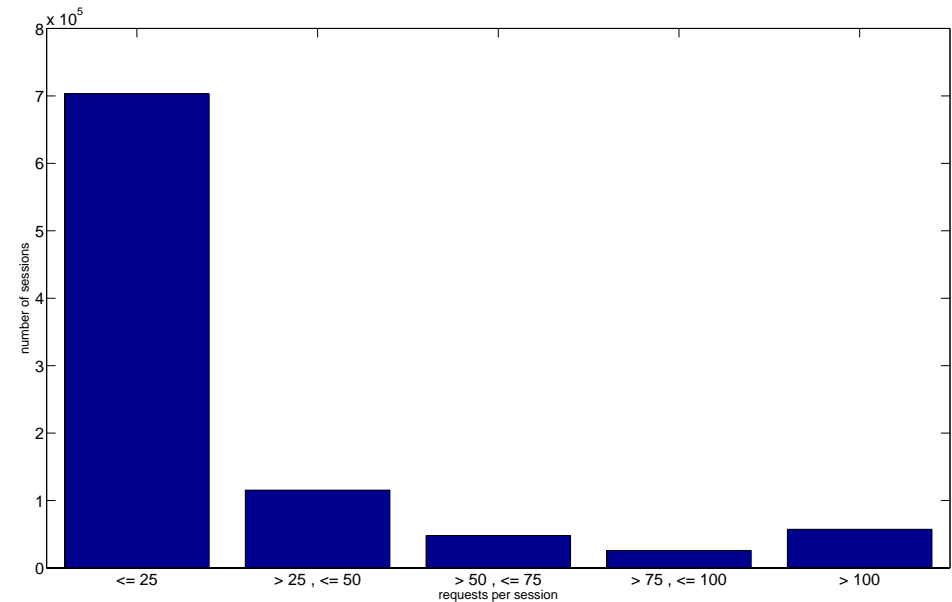
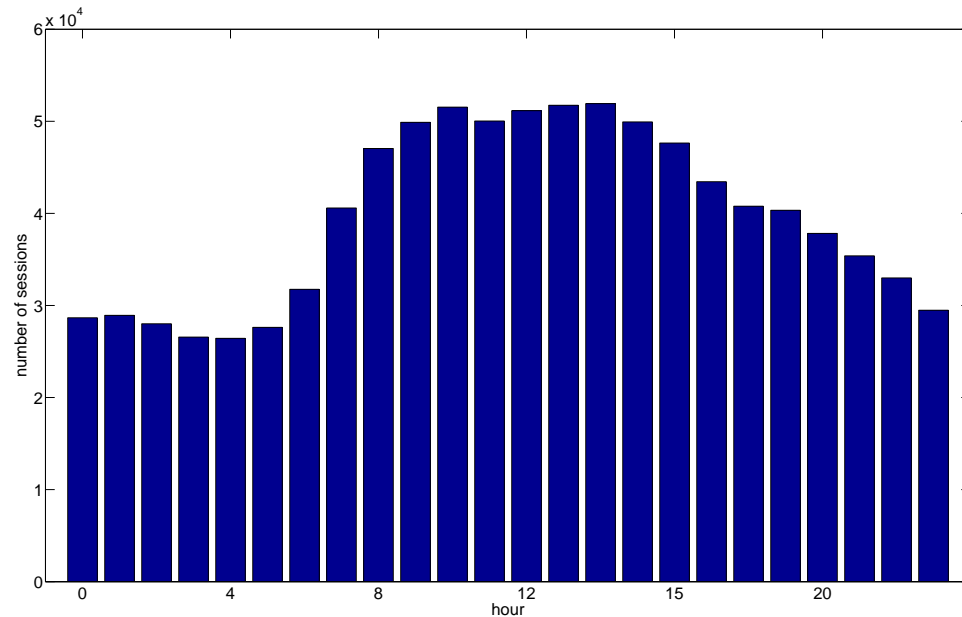
Experimental Analyses - Clients, OS, Browsers

- ▶ During this period, according to the chosen heuristic,
 - ▶ 949,643 sessions were reconstructed with an average of ~ 27 accesses per session;
 - ▶ 285,125 different pairs IP address and user-agent were found.
- ▶ Operating systems (left) and browsers (right) used by clients



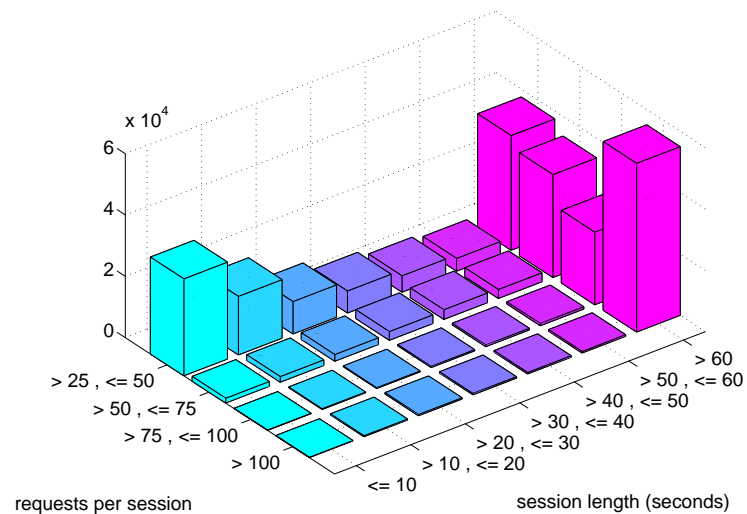
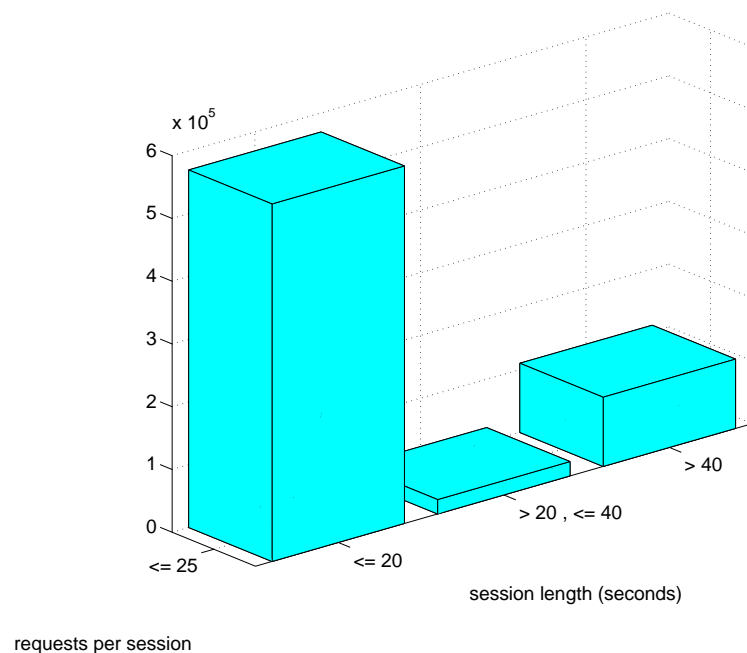
Experimental Analyses - Times, Sessions' Lengths

- ▶ Number of sessions per hour of day, Web server set on CET (left).
- ▶ Number of sessions per HTTP request intervals (right).



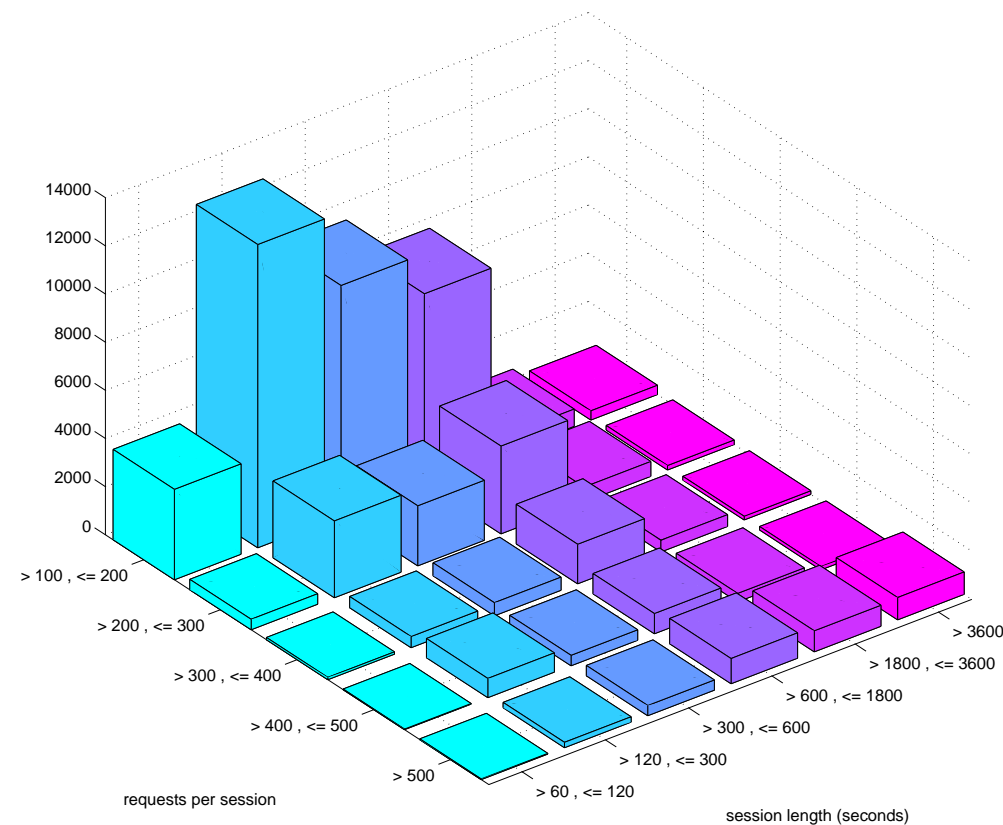
Experimental Analyses - Sessions Breakdown

- ▶ Breakdown sessions according to both the number of requests and the length of the session.
- ▶ Sessions with less than or equal 25 requests (left), more than 25 (right).
- ▶ 16% of sessions last more than 60 seconds regardless of the number requests per session.
- ▶ 12% of the sessions contain more than 50 requests.



Experimental Analyses - Focus on Lengthy Sessions

- ▶ An analysis of the sessions with more than 100 requests has been computed separately.
- ▶ The majority of sessions with a high number of requests last from 2 to 30 minutes.



Conclusions

- ▶ Preliminary analysis of eleven months of The European Library Web log data, according to a methodology for gathering and mining information from Web log files based on a DataBase Management System (DBMS) application.
- ▶ Report on initial findings about the study of user sessions which have been reconstructed by means of heuristic methods, since no personal data was available to track each user.
- ▶ Heuristics used to identify users and sessions suggested that authentication would be required since it would allow Web servers to identify users, track their requests, and more importantly create more accurate profiles to tailor specific needs.
- ▶ Authentication would also help to mitigate the problem concerning crawlers accesses, granting access to some sections of the Web site only to registered users, blocking crawlers using faked user agents.

Current and Future Works

- ▶ As a follow up of the cooperation with The European Library, the Office of The European Library has implemented the changes suggested by this work.
 - ▷ The use of cookies in the HTTP server logging system (September 2006)
 - ▷ An user authentication procedure has been established (since August 2006).
- ▶ A comparison with sessions' reconstruction using the heuristic and using cookies should give hints about the use of heuristics.
- ▶ A more accurate profile for each user, studying nationality, language, and chosen collections.