Centro de Investigación Operativa

TRABAJOS I+D

I-2005-03

Personalisation through inferring user navigation maps from web log files

Federico Botella, Enrique Lazcorreta,
Antonio Fernández-Caballero and
Pascual Gonzalez

March 2005

ISSN 1576-7264
Depósito legal A-646-2000

Centro de Investigación Operativa
Universidad Miguel Hernández de Elche
Avda. de la Universidad s/n
03202 Elche (Alicante)
cio@umh.es

Personalization through Inferring User Navigation Maps from Web Log Files

Federico Botella[1], *Enrique Lazcorreta*[1]

[1] Operations Research Centre
University Miguel Hernandez of Elche,
03202 – Elche, Spain
{federico, enrique}@umh.es

Antonio Fernández-Caballero[2], *Pascual González*[2]

[2] Computer Science Research Institute of Albacete
and Computer Science Department
University of Castilla-La Mancha, 02071 –
Albacete, Spain
{caballer, pgonzalez}@info-ab.uclm.es

Abstract

One of the most challenging areas in human computer interaction is personalization. Nonetheless, usable and well design sites can not be replaced by personalization, which must be used in just measure. In order to avoid dealing with personal data from users and thus not to worry about privacy, we decide to work with anonymous data. Moreover, these data can be found in all web server logs. In this paper, we propose a methodology for simple personalization Amazon.com does with book recommendations based only on anonymous data about user's navigation in the website. Our target is to facilitate the navigation of users by means of suggested links towards the next page that users likely want to go. And this can be accomplished in real time using an intelligent agent system or at user requirements design by running a specific process in the system that offers instantaneously the suggested links.

1 Introduction

A website that could adapt its contents to the needs of the users continues being a challenge for researchers in the present time. For personalizing a website, web developers should know the needs of the users (Nielsen, 1998). And this issue is always difficult to resolve due to the fact that users do not have the same requirements, desires and needs when they visit the same website. If we are able to offer to our visitors personalized information, we achieve to have satisfied users because they could obtain what they are really searching in the minimal time (Van Duyne, Landay & Hong, 2003).

Here we propose a methodology to achieve a simple personalization of a website based on the information gathered along anonymous visits to a web site; all web servers register this kind of data in a log file. The system will be able to present several links of the most visited web pages by all users of the website or by the last visits of the own user in the past. Several techniques of web usage mining will be applied to obtain the information necessary for presenting some suggestions on possible links that a user likely wants to go in the present visit to the website. This proposed methodology will be accomplished by several means: either by implementing an intelligent agent system that will be running on the system in real time and that can be able to make suggestions to the user every time he reaches a new web page; or by implementing a process that only will be called when a user desires to be suggested, normally the first time the user enters to the homepage of a site or a web application. In the second case, an automatic process must be launched periodically (usually at midnight or when the server has the minimal load) to maintain updated the essential data employed to offer a suggested link.

2 Personalization and Web Usage Mining

Internet has become a wide land where anyone could perform a variety of tasks or discover a multitude of services. When designing a website one always must keep in mind that people use tools to accomplish their goals with the highest possible efficacy and efficiency (Constantine & Lockwood, 1999), and a website can be considered a tool where users perform tasks. To personalize a website we need to collect information about users who visit our website, and then try to offer a web page tailored to the needs of our visitors (Nielsen, 1998). The main goal of personalization must be user satisfaction; so the process of personalization never must forget the aim of usability. Nielsen emphasizes the idea that web personalization can not be an excuse for poor design. A web site should have as its primary goal to let users make choices based on their natural intelligence. And this is possible if users find a website with a good design of their pages, their contents and, in general, with a good design of the site, where the principles of usability and accessibility have been taken in account during all the process of the design of the site (Nielsen, 2000).

But personalization deals with gathering and using of personal information, so this is a subtle issue where webmasters should preserve the privacy and confidentiality of the data handled (W3CP3P, 2002). P3P 1.0, developed by the World Wide Web Consortium, emerged as an industry standard providing a simple, automated way for users to gain more control over the use of personal information on web sites they visit. Nine aspects of online privacy are covered by P3P. Five topics detail the data being tracked by the site: Who is collecting this data?, Exactly what information is being collected?, For what purposes?, Which information is being shared with others?, And who are these data recipients?. The remaining four topics explain the site's internal privacy policies: Can users make changes in how their data is used?, How are disputes resolved?, What is the policy for retaining data?, and finally, Where can the detailed policies be found in "human readable" form? Thus, in order to avoid these topics we decide to make use in our proposed methodology of the data registered by all web servers in a special file called web log file, where only anonymous information about the actions of the users in a web site are collected.

The branch of Web Mining that deals with web log files is called Web Usage Mining, which consists of the application of data mining techniques to the usage of web resources, as recorded in web server logs (Scime, 2004; Srivastava, Cooley, Deshpande & Tan, 2000). Web usage mining is the process of extracting navigational and usage patterns of users by means of the analysis of the web log file (W3CLog, 1996). The web server logs are analyzed by using data mining techniques for extracting statistical information and discovering interesting usage patterns, gathering the users into groups related with their navigational behavior, and finally, discovering relationships between pages and user groups. In the preprocessing phase the main problems of the web usage mining are located, before data mining techniques are applied (Cooley, Mobashe, & Srivastava, 1999). Web servers register one record in the web log file for each request of any object that a browser makes to the web server, i.e., by each hit. All data associated to user's navigation in the website are included in the web log file. These files can be analyzed in several ways. The usual components of a web log file are date, time, time-taken, bytes, IP address, data transmission protocol, status and URI.

Web log files also contain entries not useful for data mining techniques, such as the requests to image files or error codes, which are superfluous to defining the user profile. The first step is to remove these irrelevant entries of the log file. Next, data transactions are identified to find similar user sessions that will be grouped according to an established principle. Once the web log file has been preprocessed and the user sessions have been delimited, we have several types of access discovering that can be accomplished by the analyst: association rules mining (Agrawal & Srikant, 1994), sequential pattern discovery (Srikant & Agrawal, 1996), clustering and classification (Ng & Han, 1994). The simplest method that can be applied to data of a web log file is statistical analysis.

Association rule mining consists of finding associations, patterns and correlations between sets of items. This technique can be used to find correlations between pages that are accessed jointly during a user session. These rules can discover the possible relationship between pages that are viewed together even if they are not directly connected. So the analyst could find relations between pages with similar accesses or between users groups with common interests. That can be used as a guide for web site restructuring, i.e., by adding links that interconnect pages often viewed together.

Sequential pattern discovery reveals patterns of access of the users with the concept of time sequence. In a web domain, a pattern might be a web page or a set of pages accessed immediately after another set of pages. In this manner, we can find several tendencies of users and define several predictions about patterns of visits.

With clustering we group elements with similar characteristics. In web mining, we can discern between user clusters and page clusters. In user clustering we make groups of users with similar ways of navigating in the website. In page clustering we make groups of pages with appearing similarity according the point of view of the user.

Finally, the process of classification maps a data item into one of several predetermined classes. In web mining, classes are used to denote different user profiles and classification is performed using selected features that describe each user's category. The most common classification algorithms are decision trees, Bayesian classifiers and neural networks.

3 Constructing user navigational maps

First of all, we analyze the anonymous user sessions in the web log files of a server to define the navigational map of a website, which it is not always provided by the webmaster of the site that one analyzes. For getting the navigational map, we first construct the personal use navigational maps of each particular user that visits our website. From these use maps we will infer the complete use navigational map, which probably will not coincide exactly with the structure navigational map constructed by the webmaster.

3.1 Preprocessing

Anonymous user sessions registered in the web log file will be extracted the more trustworthy as possible. To define a user session, the clickstreams registered in the web log file for each user will be extracted.

Previously, the web log file is cleaned of irrelevant records for getting a log file with only meaningful data that let us to locate the user sessions. All records of requests to image files (such as .jpg, .gif or .png), to media files (as .wav or .swf) and to script files (as .js or .inc) will be removed from the log file (See Figure 1). This way, only records of "visitable" pages requested by users will remain in the log. In this cleaned web log file, the set of clickstreams registered for each user will be found. A clickstream is the set of web pages requested by one user in the same session. All requests from the same IP address in the interval of time that a session is opened are considered to be made by the same user; in spite of the fact that this IP address could correspond to various users that are using the same machine, but this will lead to different user sessions.

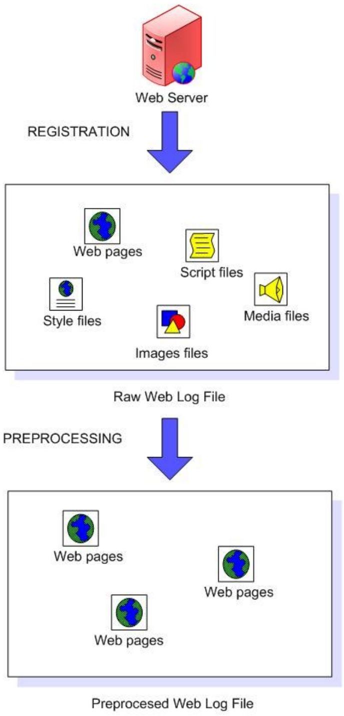

Figure 1: Preprocessing of a Web Log File

3.2 Defining user sessions

When all clickstreams are obtained, the next step is to define the user sessions, so that the permanence time in each requested web page is calculated (see Figure 2). We consider a user session to be the set of web pages visited by the same user from the first time he accesses to the site until he abandons it. We use the parameter time between two pages that are visited consecutively (T_i) to register a user session. We will use values of this parameter T_i in the range varying from 5 minutes to 3 hours. These values let us experimenting with different types of user sessions, depending of the type of the website to analyze.

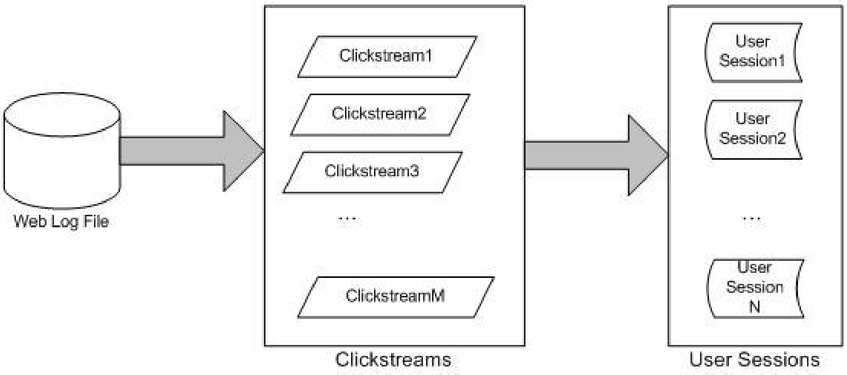

Figure 2: Extracting user session from clickstreams of a web log file

The user sessions are registered as vectors as defined in Equation 1:

$$\left(sesID, p_1, u_1, p_2, u_2, ..., p_{n-1}, u_{n-1}, p_n\right) \tag{1}$$

where *sesID* is the session identifier, u_i represents the time that a user remains in the page p_i, for $1 \leq i \leq n$-1, whilst n is the total number of visited pages in the same session. In a first approach, we consider that a session is closed if more than 30 minutes have passed since the last request of a page by the user. So, time u_n is not considered in the definition of the vector, and this time will be discarded, because we will never know how long the user session is exactly, unless the user would stay in a web application with a controlled environment and with a button or link to close the session. Even then, the user always could close directly the window, so we never could know the real length of the session.

3.3 Generating personal use navigational maps

Once all sessions have been extracted of the web log file, we will generate the personal use navigational maps for each session. In these personal use maps, the permanence time of each user in each visited page during that session and the number of hits reached in each page are annotated.

A personal use navigational map of a session is defined as a directed graph where nodes represent the web pages and edges represent the links from one page to another one. We denote by *WP* the set of web pages visited by a user in the same session that will define the personal use navigational map. If the user has visited *m* pages in that session, we have that:

$$WP = \{wp_1, wp_2, .., wp_m\} \tag{2}$$

where p_i is the *i*-th page visited by the user. We denote a link between the *i*-th page and the *j*-th page by h_{ij}. Each link h_{ij} will be pondered by the number of hits reached in the *j*-th page from the *i*-th page.

$$G_P = (P, H) \mid \left(P = \{(p_i, u_i) \mid 1 \leq i \leq n-1\}\right) \wedge \left(H = \{h_{ij} \mid 1 \leq i, j \leq n-1\}\right) \tag{3}$$

Thereby the personal graphs G_P will be composed by the set of nodes defined by pairs (p_i, u_i) where p_i represents the *i*-th page visited by the user and u_i represents the permanence time of the user in that page, and by the set of edges defined by the links h_{ij}, as shown in Equation 3.

3.4 Constructing user's personal use navigational maps

Next, the personal use navigational maps of each concrete user (or concrete IP address) are calculated (see Figure 3). Remember that we are dealing with anonymous users, so a user is defined only by an IP address. In our Intranet the IP addresses are usually fixed IPs, and in few subnetworks the IP addresses are leased by a DHCP server with a leased time of one week, so the users usually have the same IP during several months. In the case that these requirements have not been satisfied, this methodology should be applied in an identified user environment. The personal use map of one user is composed by the mathematical union of all personal use navigational maps of each session for this user and for all sessions registered for such a user.

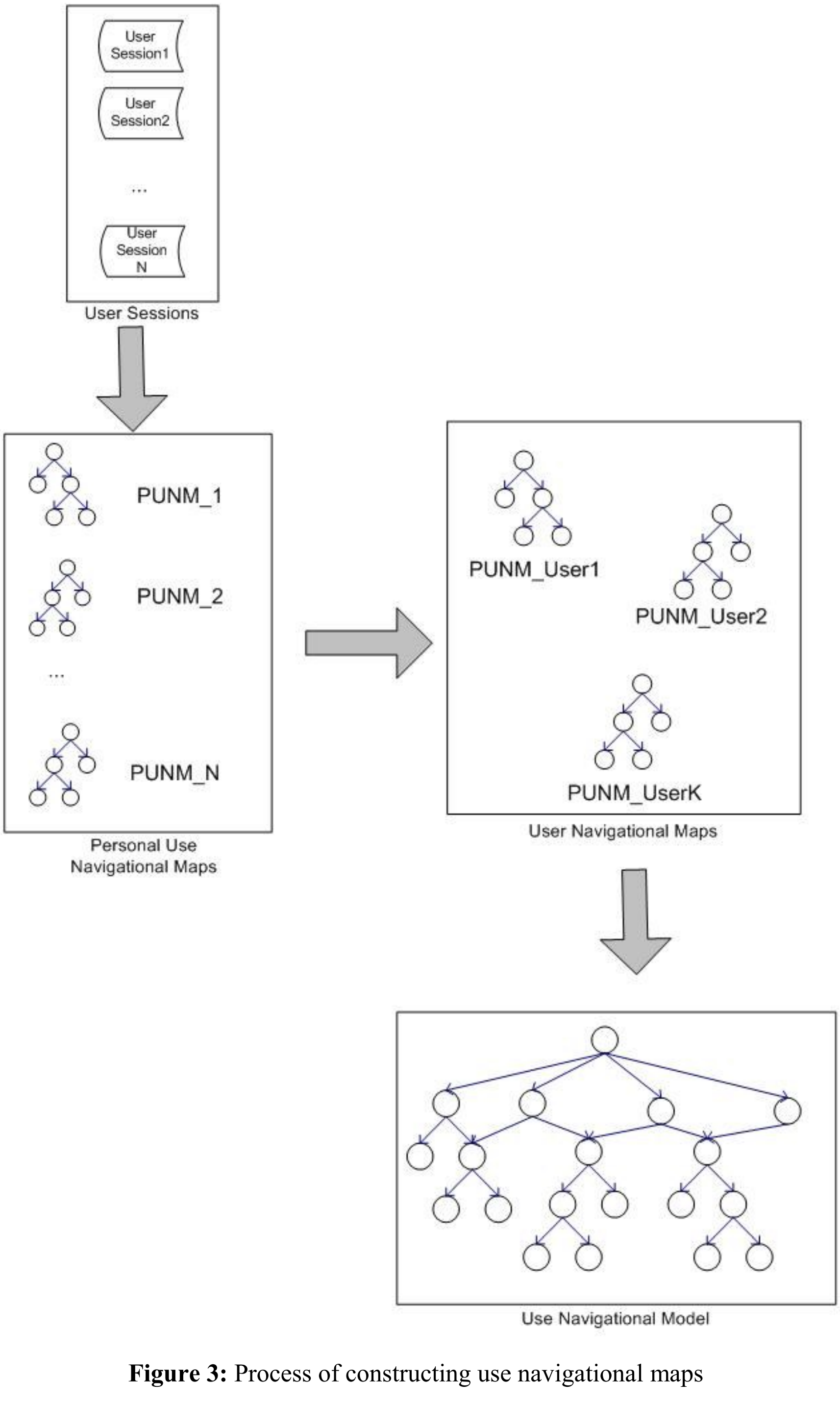

Figure 3: Process of constructing use navigational maps

3.5 Inferring Use Navigational Model

From the set of personal use navigational maps we will build the use navigational model of the complete web site. Analogously, this navigational model is defined by a directed graph. We denote by WS the set of web pages that define the navigational map of our complete web site. If our complete web site is composed of m pages, we have that

$$WS = \{ws_1, ws_2, .., ws_m\} \tag{4}$$

Now we denote a link between the *i*-th page and the *j*-th page by l_{ij}. Each link l_{ij} will be pondered by the total number of hits reached in the *j*-th page from the *i*-th page by all visitors of the web site.

Thus the navigational model of the complete web site is represented by a directed graph G_C (see Equation 5) whose nodes are defined by pairs (p_i,t_i), where p_i represents the page *i* visited by any user and t_i represents the permanence time of all users of our website in that page, and whose edges are defined by links l_{ij}, which represent the pondered weight of the total number of hits reached in the *j*-th page from the *i*-th page for all visitors.

$$G_C = (S, L) \mid (S = \{(s_i, t_i) \mid 1 \le i \le n-1\}) \wedge (L = \{l_{ij} \mid 1 \le i, j \le n-1\}) \tag{5}$$

4 Personalization of the web site

Both personal use navigational maps and use navigational map will permit to suggest to the users of the web site, either the last links visited by the user with the longest permanence time on each page during the previous session (these will be the pages where the user performs the usual tasks in the web site), or the links to pages visited by this user with the longest permanence time in the overall set of his sessions. Furthermore the method can suggest to each user the links to pages with the longest permanence time for all users of our web site.

At the beginning the system will suggest links to pages with the longest permanence time at no more than two clicks from the actual page. If there are two pages with the same permanence time, we will use the pondered weights of links to determine which page must be suggested to the user, by adding the number of hits received by the candidate page from the actual page (max two hops).

The suggesting area of the web page can be disconnected from the system, so the process of searching the suggested pages, by means of a link or a button, will be started by the user. It will be possible to have an agent running in our system; so, when any user arrives to a new page, the *N* pages more visited by all users of the web site, or during all visits of a given user, will be calculated in order to be shown afterwards.

Depending on the type of users, normally we will find that suggested links by all users of the web site will coincide with all links suggested by the historic of visits of a concrete user, even more when this user is an usual user of the web site. The first time any user arrives to the web site, obviously, only suggested links from navigational use model of the complete site will be offered. To start to suggest links to a user, it will be necessary to have at least five visits registered in the web log file, in order to analyze the historic of personal use maps of the same user with minimal guarantees to suggest links that this user wants to go.

5 Conclusions

Personalization is a promising area where researchers are still dedicating their efforts, especially in the field of e-commerce and e-learning. But the best rule of personalization is simplicity: if we want that users can customize our web site, we should present some few characteristics or items so users don't waste their time. Next we can perform some additional personalization of the web site based on these and other data. But we can achieve a simple way of personalization based only on anonymous data collected in web log files.

We have presented a method that enables to make personalization of a web site by suggesting users where they are likely to want to go in their next click. The proposed methodology only uses anonymous data from the web server log to define the personal use navigational maps of each user of the web site and the complete use navigational map. So the first can be used for suggesting or reminding of links to users, and the second can be used moreover to aid the webmaster to improve the structure of the site, because we could find relations between pages that users are employing and the webmaster did not take into account.

Finally, personal use navigational maps of each user could be employed to discover user profiles that could lead us to apply other data mining techniques for clustering and classification.

Acknowledgements

This work has been partially funded by Spanish PBC-03-003 grant supported by Junta de Comunidades de Castilla-La Mancha and by R.R.1256/04 grant supported by University Miguel Hernández of Elche – Bancaja.

References

Agrawal, R. & Srikant, R. (1994). Fast Algorithms for Mining Association Rules. In Proc.20th Int.Conf.Very Large Data Bases pp. 487-499. Santiago, Chile.

Constantine, L. L. & Lockwood, L. A. D. (1999). Software for use: A Practical Guide to the Models and Methods of Usage-Centered Design. Addison-Wesley.

Cooley, R., Mobasher, B., & Srivastava, J. (1999). Data Preparation for Mining World Wide Web Browsing Patterns. Knowledge and Information Systems, 1, 5-32.

Ng, R. T. & Han, J. (1994). Efficient and Effective Clustering Methods for Spatial Data Mining. In 20th International Conference on Very Large Data Bases, Santiago, Chile pp. 144-155.

Nielsen, J. (2000). Designing Web Usability. Prentice Hall.

Nielsen, J. (1998). Alertbox: Personalization is Over-Rated. From http://www.useit.com/alertbox/981004.html

Scime, A. (2004). Web Mining: Applications and Techniques. Idea Group Publishing.

Srikant, R. & Agrawal, R. (1996). Mining sequential patterns: Generalizations and Performance Improvements. In Proc.5th Int.Conf.Extending Database Technology pp. 3-17.

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P.-N. (2000). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. SIGKDD Explorations, 2000, 1, 12-23.

Van Duyne, D. K., Landay, J. A., & Hong, J. I. (2003). The design of sites: Patterns, Principles and Processes for crafting a customer-centered web experience. Addison-Wesley.

W3CLog (1996). Extended log file format. From http://www.w3.org/TR/WD-logfile.html

W3CP3P (2002). The Platform for Privacy Preferences 1.0 (P3P1.0) Specification. From http://www.w3.org/TR/P3P/