2-1-2003

Web Mining for Web Personalization

Magdalini Eirinaki
San Jose State University, magdalini.eirinaki@sjsu.edu

M. Vazirgiannis
University of Athens, Greece

Recommended Citation

Magdalini Eirinaki and M. Vazirgiannis. "Web Mining for Web Personalization" *ACM Transactions on Internet Technology* (2003): 1-27. https://doi.org/10.1145/643477.643478

Web Mining for Web Personalization

MAGDALINI EIRINAKI and MICHALIS VAZIRGIANNIS
Athens University of Economics and Business

Web personalization is the process of customizing a Web site to the needs of specific users, taking advantage of the knowledge acquired from the analysis of the user's navigational behavior (usage data) in correlation with other information collected in the Web context, namely, structure, content and user profile data. Due to the explosive growth of the Web, the domain of Web personalization has gained great momentum both in the research and commercial areas. In this article we present a survey of the use of Web mining for Web personalization. More specifically, we introduce the modules that comprise a Web personalization system, emphasizing the Web usage mining module. A review of the most common methods that are used as well as technical issues that occur is given, along with a brief overview of the most popular tools and applications available from software vendors. Moreover, the most important research initiatives in the Web usage mining and personalization areas are presented.

Key Words and Phrases: Web personalization, Web usage mining, user profiling, WWW

1. INTRODUCTION

The continuous growth in the size and use of the World Wide Web imposes new methods of design and development of online information services. Most Web structures are large and complicated and users often miss the goal of their inquiry, or receive ambiguous results when they try to navigate through them. On the other hand, the e-business sector is rapidly evolving and the need for Web marketplaces that anticipate the needs of the customers is more evident than ever.

Therefore, the requirement for predicting user needs in order to improve the usability and user retention of a Web site can be addressed by personalizing it. Web personalization is defined as any action that adapts the information or services provided by a Web site to the needs of a particular user or a set of users, taking advantage of the knowledge gained from the users' navigational behavior and individual interests, in combination with the content and the structure of the Web site. The objective of a Web personalization system is to "provide users with the information they want or need, without expecting from them to ask for it explicitly" [Mulvenna et al. 2000].

At this point, it is necessary to stress the difference between layout customization and personalization. In customization the site can be adjusted to each user's preferences regarding its structure and presentation. Every time a registered user logs in, their customized home page is loaded. This process is performed either manually or semiautomatically. In personalization systems modifications concerning the content or even the structure of a Web site are performed dynamically.

Principal elements of Web personalization include (a) the categorization and preprocessing of Web data, (b) the extraction of correlations between and across different kinds of such data, and (c) the determination of the actions that should be recommended by such a personalization system [Mobasher et al. 2000a].

This research work was partially supported by the IST-2000-31077/I-KnowUMine R&D project funded by the European Union

Authors' address: Department of Informatics, Athens University of Economics and Business, Patision 76, Athens, 10434, Greece; email: {eirinaki,mvazirg}@aueb.gr.

ACM Transactions on Internet Technology, Vol. 3, No. 1, February 2003, Pages 1–27.

Web data are those that can be collected and used in the context of Web personalization. These data are classified in four categories according to Srivastava et al. [2000].

- *Content* data are presented to the end-user appropriately structured. They can be simple text, images, or structured data, such as information retrieved from databases.
- *Structure* data represent the way content is organized. They can be either data entities used within a Web page, such as HTML or XML tags, or data entities used to put a Web site together, such as hyperlinks connecting one page to another.
- *Usage* data represent a Web site's usage, such as a visitor's IP address, time and date of access, complete path (files or directories) accessed, referrers' address, and other attributes that can be included in a Web access log.
- *User profile* data provide information about the users of a Web site. A user profile contains demographic information (such as name, age, country, marital status, education, interests etc.) for each user of a Web site, as well as information about users' interests and preferences. Such information is acquired through registration forms or questionnaires, or can be inferred by analyzing Web usage logs.

The overall process of usage-based Web personalization consists of five modules, which correspond to each step of the process. These are as follows.

- *User profiling:* In the Web domain, user profiling is the process of gathering information specific to each visitor, either explicitly or implicitly. A user profile includes demographic information about the user, her interests and even her behavior when browsing a Web site. This information is exploited in order to customize the content and structure of a Web site to the visitor's specific and individual needs.
- *Log analysis and Web usage mining:* This is the procedure where the information stored in Web server logs is processed by applying data mining techniques in order to (a) extract statistical information and discover interesting usage patterns, (b) cluster the users into groups according to their navigational behavior, and (c) discover potential correlations between Web pages and user groups. This process of extracting information concerning the browsing behavior of the users can be regarded as part of the user profiling process. It is therefore evident that the user profiling and Web usage mining modules overlap.
- *Content management:* This is the process of classifying the content of a Web site in semantic categories in order to make information retrieval and presentation easier for the users. Content management is very important for Web sites whose content is increasing on a daily basis, such as news sites or portals.
- *Web site publishing:* A publishing mechanism is used in order to present the content stored locally in a Web server and/or some information retrieved from other Web resources in a uniform way to the end-user. Different technologies can be used to publish data on the Web.
- *Information acquisition and searching:* In many cases information provided by a Web site is not physically stored in the Web site's server. In the case of a Web portal or vortal (vertical portal), users are interested in information from various Web sources. It remains to the Web site editors to search the Web for content of interest that should consequently be classified into thematic categories. Searching and relevance ranking techniques must be employed both in the process of acquisition of relevant information and in the publishing of the appropriate data to each group of users.

A usage-based Web personalization system utilizes Web data in order to modify a Web site. Site personalization is achieved through the interaction of the aforementioned modules. This survey article is organized as follows. In Section 2, we provide a brief

description of the Web personalization process and illustrate the interaction of these modules in such a system. In the context of this survey we analyze user profiling, as well as log analysis and Web usage mining modules. These modules are described in more detail in Sections 3 and 4, respectively. An analysis of the methods that are used along with relevant technical issues, in addition to an overview of the tools and applications available from software vendors are included. In Section 5 we present the most important research initiatives in the area of Web usage mining and personalization. In Appendices A and B lists of acronyms and abbreviations as well as Web References are presented. In Appendix C there is a tabular comparative presentation of the most representative tools for user profiling and Web usage mining, as well as the most important research initiatives in the area of Web mining and Web personalization.

2. WEB PERSONALIZATION

Web site personalization can be defined as the process of customizing the content and structure of a Web site to the specific and individual needs of each user taking advantage of the user's navigational behavior. The steps of a Web personalization process include: (a) the collection of Web data, (b) the modeling and categorization of these data (preprocessing phase), (c) the analysis of the collected data, and (d) the determination of the actions that should be performed. The ways that are employed in order to analyze the collected data include content-based filtering, collaborative filtering, rule-based filtering, and Web usage mining. The site is personalized through the highlighting of existing hyperlinks, the dynamic insertion of new hyperlinks that seem to be of interest for the current user, or even the creation of new index pages.

Content-based filtering systems are solely based on individual users' preferences. The system tracks each user's behavior and recommends items to them that are similar to items the user liked in the past.

Collaborative filtering systems invite users to rate objects or divulge their preferences and interests and then return information that is predicted to be of interest to them. This is based on the assumption that users with similar behavior (e.g. users that rate similar objects) have analogous interests.

In *rule-based filtering* the users are asked to answer a set of questions. These questions are derived from a decision tree, so as the user proceeds to answer them, what he finally receives as a result (e.g. a list of products) is tailored to his needs. Content-based, rule-based, and collaborative filtering may also be used in combination, for deducing more accurate conclusions.

In this work we focus on *Web usage mining*. This process relies on the application of statistical and data mining methods to the Web log data, resulting in a set of useful patterns that indicate users' navigational behavior. The data mining methods that are employed are: association rule mining, sequential pattern discovery, clustering, and classification. This knowledge is then used from the system in order to personalize the site according to each user's behavior and profile.

The block diagram illustrated in Figure 1 represents the functional architecture of a Web personalization system in terms of the modules and data sources that were described earlier. The content management module processes the Web site's content and classifies it in conceptual categories. The Web site's content can be enhanced with additional information acquired from other Web sources, using advanced search techniques. Given the site map structure and the usage logs, a Web usage miner provides results regarding usage patterns, user behavior, session and user clusters, clickstream information, and so on. Additional information about the individual users can be obtained by the user profiles. Moreover, any information extracted from the Web usage mining process concerning each user's navigational behavior can then be added to her profile. All this information about nodes, links, Web content, typical behaviors, and patterns is conceptually abstracted and classified into semantic categories. Any information extracted from the interrelation between knowledge acquired using usage mining techniques and knowledge acquired from content management will then provide the framework for evaluating possible alternatives for

restructuring the site. A publishing mechanism will perform the site modification, ensuring that each user navigates through the optimal site structure. The available content options for each user will be ranked according to the user's interests.

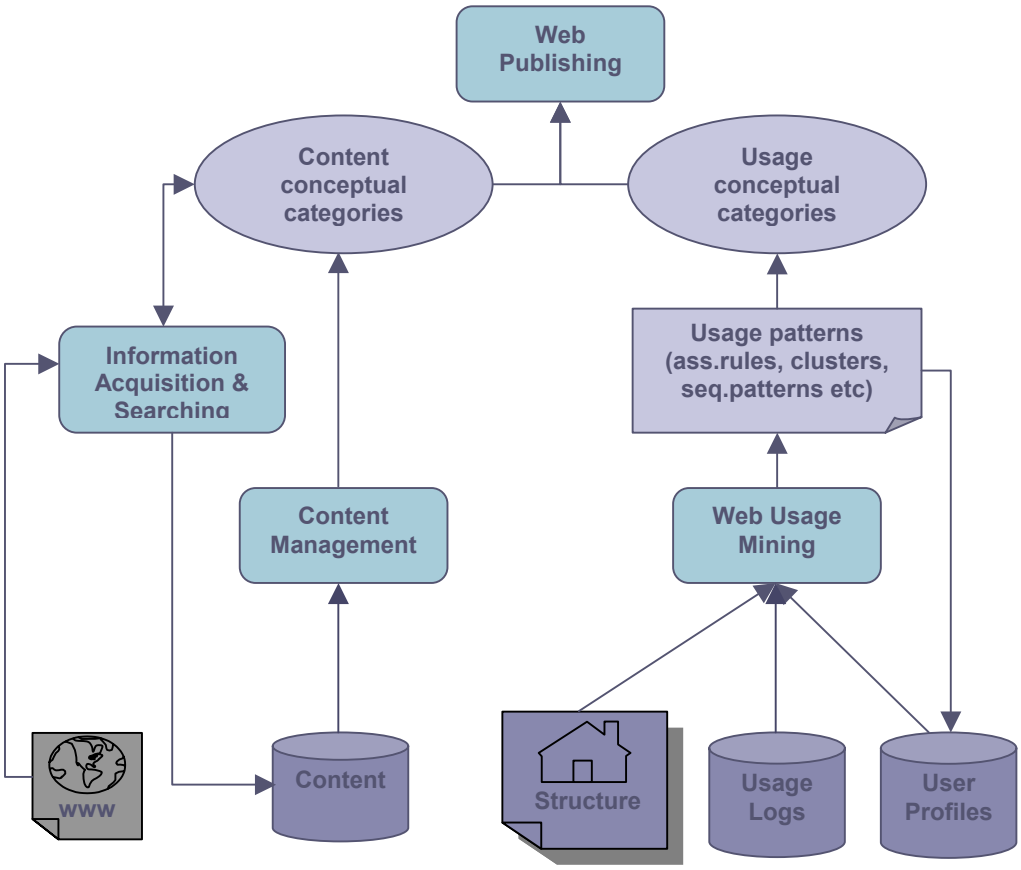

Fig. 1: Modules of a Web personalization system.

3. USER PROFILING

In order to personalize a Web site, the system should be able to distinguish between different users or groups of users. This process is called user profiling and its objective is the creation of an information base that contains the preferences, characteristics, and activities of the users. In the Web domain and especially in e-commerce, user profiling has been developed significantly because Internet technologies provide easier means of collecting information about the users of a Web site, which in the case of e-business sites are potential customers.

A user profile can be either *static,* when the information it contains is never or rarely altered (e.g., demographic information), or *dynamic* when the user profile's data change frequently. Such information is obtained either explicitly, using online registration forms and questionnaires resulting in static user profiles, or implicitly, by recording the navigational behavior and/or the preferences of each user, resulting in dynamic user profiles. In the latter case, there are two further options: either regarding each user as a member of a group and creating aggregate user profiles, or addressing any changes to each user individually. When addressing the users as a group, the method used is the creation of aggregate user profiles based on rules and patterns

extracted by applying Web usage mining techniques to Web server logs. Using this knowledge, the Web site can be appropriately customized. This case is discussed in detail in Section 4, therefore it won't be further analyzed here.

In the following sections, we provide a description of several methods for implicit and explicit collection of user profile data. Privacy issues that arise in the user profiling process are discussed, and an overview of available tools and user profiling applications is presented.

3.1 Data Collection

A way of uniquely identifying a visitor through a session is by using cookies. W3C [WCA] defines cookie as "the data sent by a Web server to a Web client, stored locally by the client and sent back to the server on subsequent requests." In other words, a cookie is simply an HTTP header that consists of a text-only string, which is inserted into the memory of a browser. It is used to uniquely identify a user during Web interactions within a site and contains data parameters that allow the remote HTML server to keep a record of the user identity, and what actions he takes at the remote Web site.

The contents of a cookie file depend on the Web site that is being visited. In general, information about the visitor's identification is stored, along with password information. Additional information such as credit card details, if one is used during a transaction, as well as details concerning the visitor's activities at the Web site, for example, which pages were visited, which purchases were made, or which advertisements were selected, can also be included. Often, cookies point back to more detailed customer information stored at the Web server.

Another way of uniquely identifying users through a Web transaction is by using *identd*, an identification protocol specified in RFC 1413 [RFC] that provides a means to determine the identity of a user of a particular TCP connection. Given a TCP port number pair, it returns a character string, which identifies the owner of that connection (the client) on the Web server's system.

Finally, a user can be identified making the assumption that each IP corresponds to one user. In some cases, IP addresses are resolved into domain names that are registered to a person or a company, thus more specific information is gathered.

As already mentioned, user profiling information can be explicitly obtained by using online registration forms requesting information about the visitor, such as name, age, sex, likes, and dislikes. Such information is stored in a database, and each time the user logs on the site, it is retrieved and updated according to the visitor's browsing and purchasing behavior.

All of the aforementioned techniques for profiling users have certain drawbacks. First of all, in the case where a system depends on cookies for gathering user information, there exists the possibility of the user having turned off cookie support on his browser. Other problems that may occur when using cookies technology are the fact that because a cookie file is stored locally in the user's computer, the user might delete it and when she revisits a Web site will be regarded as a new visitor. Furthermore, if no additional information is provided (e.g., some logon id), there occurs an identification problem if more than one user browses the Web using the same computer. A similar problem occurs when using identd, inasmuch as the client should be configured in a mode that permits plaintext transfer of ids. A potential problem in identifying users using IP address resolving, is that in most cases this address is that of the ISP, and that does not suffice for specifying the user's location. On the other hand, when gathering user information through registration forms or questionnaires, many users submit false information about themselves and their interests resulting in the creation of misleading profiles.

3.2 Privacy Issues

The most important issue that should be encountered during the user profiling process is privacy violation. Many users are reluctant to give away personal information either implicitly as mentioned before, or explicitly, being hesitant to visit Web sites that use

cookies (if they are aware of their existence) or avoiding disclosure of personal data in registration forms. In both cases, the user loses anonymity and is aware that all of his actions will be recorded and used, in many cases without his consent. In addition, even if a user has agreed to supply personal information to a site, through cookie technology such information can be exchanged between sites, resulting in its disclosure without the user's permission.

P3P (Platform for Privacy Preferences) is a W3C proposed recommendation [P3P] that suggests an infrastructure for the privacy of data interchange. This standard enables Web sites to express their privacy practices in a standardized format that can be automatically retrieved and interpreted by user agents. Therefore, the process of reading privacy policies will be simplified for the users, because key information about what data are collected by a Web site can be automatically conveyed to a user, and discrepancies between a site's practices and the user's preferences concerning the disclosure of personal data will be automatically flagged. P3P, however, does not provide a mechanism for ensuring that sites actually act according to their policies.

3.3 Tools and Applications

In this section we present some of the most popular Web sites that use methods such as decision tree guides, collaborative filtering, and cookies in order to profile users and create customized Web pages. Additionally, a brief description of the most important tools available for user profiling is given. An overview along with products references is provided in Appendix Table AI.

Popular Web sites such as Yahoo! [YAH], Excite [EXC], or Microsoft Network [MSN] allow users to customize home pages based on their selections of available content, using information supplied by the users and cookies thereafter. In that way, each time the user logs on the site, what she sees is a page containing information addressed to her interests.

Rule-based filtering is used from online retailers such as Dell [DEL] and Apple Computer [APP], giving users the ability to easily customize product configurations before ordering. As far as recommendation systems are concerned, the most popular example is Amazon.com [AMA]. The system analyzes past purchases and posts suggestions on the shopper's customized recommendations page. Users who haven't made a purchase before can rate books and see listings of books they might like. The same approach, based on user ratings, is used in many similar online shops, such as CDNOW [CDN].

Another interesting approach is that of Food.com [FOO]. Users are not required to fill in any form to order food from a specific nearby restaurant. Customization happens automatically as users give the necessary information for a food delivery or pickup, because zip code data provide the necessary information for suggesting nearby restaurants.

Commercial Web sites, including many search engines such as Alta-Vista [ALT] or Lycos [LYC], have associations with commercial marketing companies such as DoubleClick Inc. [DCL]. These sites use cookies to monitor their visitors' activities, and any information collected is stored as a profile in DoubleClick's database. DoubleClick then uses this profile information to decide which advertisements or services should be offered to each user when he visits one of the affiliated DoubleClick sites. Of course, this information is collected and stored without the users' knowledge and more importantly, consent.

There are several systems available for creating user profiles. They vary according to the user profiling method that they utilize. These include (a) Broadvision's One-To-One [BRO], a high-end marketing tool designed to let sites recognize customers and display relevant products and services (customers include Kodak Picture Network, and US West); (b) Net Perception's GroupLens [NPE], a collaborative filtering solution requiring other users to actively or passively rate content (clients include Amazon.com and Musicmaker); (c) Open Sesame's Learn Sesame [OSE], a cookie-based product (clients include Ericsson and Toronto Dominion Bank); (d) the early leader in collaborative filtering Firefly Passport [MSF], developed by MIT Media Lab and now

owned by Microsoft (clients include Yahoo, Ziff-Davis, and Barnes & Noble); (e) Macromedia's LikeMinds Preference Server [MIC], another collaborative filtering system that examines users' behavior and finds other users with similar behaviors in order to create a prediction or product recommendation (clients include Cinemax-HBO's Movie Matchmaker and Columbia House's Total E entertainment site); (f) Neuromedia's NeuroStudio [NME], an intelligent-agent software that allows Webmasters to give users the option to create customized page layouts, using either cookies or user log-in (customers include Intel and Y2K Links Database site); and (g) Apple's WebObjects [APP], a set of development tools that allow customized data design (clients include The Apple Store and Cybermeals) [Dean 1998].

3.4 Conclusions

User profiling is the process of collecting information about the characteristics, preferences, and activities of a Web site's visitors. This can be accomplished either explicitly or implicitly. Explicit collection of user profile data is performed through the use of online registration forms, questionnaires, and the like. The methods that are applied for implicit collection of user profile data vary from the use of cookies or similar technologies to the analysis of the users' navigational behavior that can be performed using Web usage mining techniques.

It is evident that in order to personalize a Web site, user profiling is essential. However, all the techniques that are used for this purpose have some drawbacks. The users' privacy violation is the most important issue that should be addressed. P3P is a standard that enables Web sites to express their privacy practices in a standardized format that can be automatically retrieved and interpreted by user agents. In that way, the process of reading the privacy statements of the Web sites becomes simpler, however, P3P does not provide a guarantee that these sites act according to these declared policies.

The extraction of information concerning the navigational behavior of Web site visitors is the objective of Web usage mining. Nevertheless this process can also be regarded as part of the creation of user profiles; it is therefore evident that those two modules overlap and are fundamental in the Web personalization process.

4. LOG ANALYSIS AND WEB USAGE MINING

The purpose of Web usage mining is to reveal the knowledge hidden in the log files of a Web server. By applying statistical and data mining methods to the Web log data, interesting patterns concerning the users' navigational behavior can be identified, such as user and page clusters, as well as possible correlations between Web pages and user groups.

The Web usage mining process can be regarded as a three-phase process, consisting of the data preparation, pattern discovery and pattern analysis phases [Srivastava et al. 2000]. In the first phase, Web log data are preprocessed in order to identify users, sessions, pageviews, and so on. In the second phase, statistical methods, as well as data mining methods (such as association rules, sequential pattern discovery, clustering, and classification) are applied in order to detect interesting patterns. These patterns are stored so that they can be further analyzed in the third phase of the Web usage mining process.

A description of the fields included in a log entry of a Web usage log follows, along with a set of definitions of Web data abstractions, such as Web site, user, session, pageviews, and clickstreams. Technical issues, concerning data preparation are discussed. A more detailed analysis of the methods employed in the Web usage mining process, including simple log analysis is presented. Finally, a brief overview of the commercially available tools and applications specializing in log analysis or Web usage mining is given.

4.1 Web Log

Each access to a Web page is recorded in the access log of the Web server that hosts it. The entries of a Web log file consist of fields that follow a predefined format. The fields of the common log format are:

remotehost rfc931 authuser date "request" status bytes

where *remotehost* is the remote hostname or IP number if the DNS hostname is not available; *rfc931* the remote log name of the user; *authuser,* the username with which the user has authenticated himself, available when using password-protected WWW pages; *date,* the date and time of the request; *"request",* the request line exactly as it came from the client (the file, the name, and the method used to retrieve it); *status,* the HTTP status code returned to the client, indicating whether the file was successfully retrieved and if not, what error message was returned; and *bytes,* the content-length of the documents transferred. If any of the fields cannot be determined a minus sign (-) is placed in this field.

Lately, W3C [W3Clog] presented an improved format for Web server log files, called the "extended" log file format, partially motivated by the need to support the collection of data for demographic analysis and for log summaries. This format permits customized log files to be recorded in a format readable by generic analysis tools. The main extension to the common log format is that a number of fields are added to it. The most important are: *referrer,* which is the URL the client was visiting before requesting that URL, *user_agent,* which is the software the client claims to be using, and *cookie,* in the case where the site visited uses cookies.

In general, extended log format consists of a list of prefixes such as *c* (client), *s* (server), *r* (remote), *cs* (client to server), *sc* (server to client), *sr* (server to remote server, used by proxies), *rs* (remote server to server, used by proxies), *x* (application-specific identifier), and a list of identifiers such as *date, time, ip, dns, bytes, cached* (records whether a cache hit occurred)*, status, comment* (comment returned with status code)*, method, uri, uri-stem* and *uri-query.* Using a combination of some of the aforementioned prefixes and identifiers, additional information such as referrers' IPs, or keywords used in search engines can be stored.

4.2 Web Data Abstractions

In the Web domain, several abstractions are mentioned, concerning Web usage, content, and structure. The W3C Web Characterization Activity [WCA] has published a draft establishing precise semantics for concepts such as Web site, user, user sessions, server sessions, pageviews, and clickstreams.

A *Web site* is defined as a collection of interlinked Web pages, including a host page, residing at the same network location. A *user* is defined to be the principal using a client to interactively retrieve and render resources or resource manifestations. In the Web context, a user is an individual that is accessing files from a Web server, using a browser. A *user session* is defined as a delimited set of user clicks across one or more Web servers. A *server session* is defined as a collection of user clicks to a single Web server during a user session. It is also called a *visit.* A *pageview* is defined as the visual rendering of a Web page in a specific environment at a specific point in time. In other words, a pageview consists of several items, such as frames, text, graphics, and scripts that construct a single Web page. A *clickstream* is a sequential series of pageview requests, made from a single user.

4.3 Data Preprocessing

There are some important technical issues that must be taken into consideration during this phase in the context of the Web personalization process, because it is necessary for Web log data to be prepared and preprocessed in order to use them in the consequent phases of the process. An extensive description of data preparation and preprocessing methods can be found in Cooley et al. [1999a]. In the sequel, we provide a brief overview of the most important ones.

The first issue in the *preprocessing* phase is *data preparation*. Depending on the application, Web log data may need to be cleaned from entries involving pages that returned an error or graphics file accesses. In some cases such information might be useful, but in others such data should be eliminated from a log file. Furthermore, crawler activity can be filtered out, because such entries do not provide useful information about the site's usability. Another problem to be met has to do with caching. Accesses to cached pages are not recorded in the Web log, therefore such information is missed. Caching is heavily dependent on the client-side technologies used and therefore cannot be dealt with easily. In such cases, cached pages can usually be inferred using the referring information from the logs. Moreover, a useful aspect is to perform *pageview identification*, determining which page file accesses contribute to a single pageview. Again such a decision is application-oriented.

Most important of all is the *user identification* issue. There are several ways to identify individual visitors. The most obvious solution is to assume that each IP address (or each IP address/client agent pair) identifies a single visitor. Nonetheless, this is not very accurate because, for example, a visitor may access the Web from different computers, or many users may use the same IP address (if a proxy is used). A further assumption can then be made, that consecutive accesses from the same host during a certain time interval come from the same user. More accurate approaches for a priori identification of unique visitors are the use of cookies or similar mechanisms or the requirement for user registration. However, a potential problem in using such methods might be the reluctance of users to share personal information.

Assuming a user is identified, the next step is to perform *session identification*, by dividing the clickstream of each user into sessions. The usual solution in this case is to set a minimum timeout and assume that consecutive accesses within it belong to the same session, or set a maximum timeout, where two consecutive accesses that exceed it belong to different sessions.

4.4 Log Analysis

Log analysis tools (also called traffic analysis tools) take as input raw Web data and process them in order to extract statistical information. Such information includes statistics for the site activity (such as total number of visits, average number of hits, successful/failed/redirected/cached hits, average view time, and average length of a path through a site), diagnostic statistics (such as server errors, and page not found errors), server statistics (such as top pages visited, entry/exit pages, and single access pages), referrers statistics (such as top referring sites, search engines, and keywords), user demographics (such as top geographical location, and most active countries/cities/organizations), client statistics (visitor's Web browser, operating system, and cookies), and so on. Some tools also perform clickstream analysis, which refers to identifying paths through the site followed by individual visitors by grouping together consecutive hits from the same IP, or include limited low-level error analysis, such as detecting unauthorized entry points or finding the most common invalid URL. These statistics are usually output to reports and can also be displayed as diagrams.

This information is used by administrators for improving the system performance, facilitating the site modification task, and providing support for marketing decisions [Srivastava et al. 2000]. However, most advanced Web mining systems further process this information to extract more complex observations that convey knowledge, utilizing data mining techniques such as association rules and sequential pattern discovery, clustering, and classification. These techniques are described in more detail in the next paragraph.

4.5 Web Usage Mining

Log analysis is regarded as the simplest method used in the Web usage mining process. The purpose of Web usage mining is to apply statistical and data mining techniques to the preprocessed Web log data, in order to discover useful patterns. As mentioned before, the most common and simple method that can be applied to such data is statistical analysis. More advanced data mining methods and algorithms tailored

appropriately for use in the Web domain include association rules, sequential pattern discovery, clustering, and classification.

Association rule mining is a technique for finding frequent patterns, associations, and correlations among sets of items. Association rules are used in order to reveal correlations between pages accessed together during a server session. Such rules indicate the possible relationship between pages that are often viewed together even if they are not directly connected, and can reveal associations between groups of users with specific interests. Aside from being exploited for business applications, such observations also can be used as a guide for Web site restructuring, for example, by adding links that interconnect pages often viewed together, or as a way to improve the system's performance through prefetching Web data.

Sequential pattern discovery is an extension of association rules mining in that it reveals patterns of cooccurrence incorporating the notion of time sequence. In the Web domain such a pattern might be a Web page or a set of pages accessed immediately after another set of pages. Using this approach, useful users' trends can be discovered, and predictions concerning visit patterns can be made.

Clustering is used to group together items that have similar characteristics. In the context of Web mining, we can distinguish two cases, user clusters and page clusters. Page clustering identifies groups of pages that seem to be conceptually related according to the users' perception. User clustering results in groups of users that seem to behave similarly when navigating through a Web site. Such knowledge is used in e-commerce in order to perform market segmentation but is also helpful when the objective is to personalize a Web site.

Classification is a process that maps a data item into one of several predetermined classes. In the Web domain classes usually represent different user profiles and classification is performed using selected features that describe each user's category. The most common classification algorithms are decision trees, naïve Bayesian classifier, neural networks, and so on.

After discovering patterns from usage data, a further analysis has to be conducted. The exact methodology that should be followed depends on the technique previously used. The most common ways of analyzing such patterns are either by using a query mechanism on a database where the results are stored, or by loading the results into a data cube and then performing OLAP operations. Additionally, visualization techniques are used for an easier interpretation of the results. Using these results in association with content and structure information concerning the Web site there can be extracted useful knowledge for modifying the site according to the correlation between user and content groups.

4.6 Tools and Applications

It is evident that Web usage mining is a powerful tool for corporations that invest in the e-business sector. The application of Web usage mining techniques to data gathered from customers' online activity helps them to acquire business intelligence by providing high-level knowledge in the form of rules and patterns that describe consumer navigational and purchasing behavior [Buchner and Mulvenna 1998]. Thus consumer profiles and market segmentation can be achieved giving these companies a competitive advantage. Even in the case of smaller organizations or individuals, the outcome of log analysis and Web usage mining can help them improve the performance of their systems, identify their Web site's visitors, and even customize their Web site making it more efficient and user-friendly. Therefore, there exists a large variety of Web analytics' products, ranging from free traffic analysis tools to integrated CRM (Customer Relationship Management) solutions. The most important of these are presented here. An overview along with products references is included in Appendix Table AII.

Log analysis is the first step in Web usage mining and is performed by all the commercially available systems. The majority of the public and shareware tools are log/traffic analyzers and their functionality is limited to simply producing a set of statistical reports. Some publicly available applications are Analog [ANA], WebLogs

[CAP], WebLog [AWS], Ststat [STS], Follow 2 [MNO], and WUM [WUM]. All of them provide the end-user with a set of statistical reports, and some of them, such as WebLog and Follow 2, also track user sessions by presenting specific information about each individual visitor. WUM is a more advanced application, because it emerged from a research project. It can be characterized rather as a sequence miner, being also appropriate for sequential pattern discovery inasmuch as it is supported by a mining query language, MINT. There are also some shareware products and services available, such as Web Trends [WTR], Funnel Web [QUE], Net Tracker [NTR], Mach 5 Faststats Analyzer [MAH], Sawmill [SAW], SurfStats Log Analyzer [SUR], Happy Log [HAP], Webfeedback [LIE], and WebLog Manager Pro [MON]. Some of them such as Happy Log and Webfeedback have limited capabilities, providing only reports for general statistics and server statistics. The rest of the products offer more advanced functionality, however they don't make use of data mining techniques (except for some companies that also offer high-end systems or services). These products target individuals or small enterprises, which want an inexpensive solution in order to track and analyze the traffic on their (usually) single Web server.

More advanced features are offered by integrated solutions, which usually provide some data mining algorithms along with data warehousing services, output to reports, charts and diagrams or even providing recommendations in natural language. Most of them are parts of integrated CRM products that help a company gather business intelligence by combining the knowledge mined from Web logs with customer information collected from other sources such as registration information, operational data (CRM and ERP), demographics, and the like. Such systems are either packaged solutions that are installed by the company, or services that outsource analytics to the companies (Application Service Providers, ASPs). The latter is a sector that is becoming increasingly popular because ASPs offer faster implementation times and are less expensive than relevant software solutions. ASPs usually gather data at the client side, through the user's browser.

Some of the most well-known software systems are those of Accrue [ACC], Elytics [ELY], E.piphany [EPI], Lumio [LUM], NCR [NCR], NetGenesis [NGE], Net Perceptions [NPE], Quest [QUE], Sane solutions [SAN], SAS [SAS], and WebTrends [WTR]; on the other hand Coremetrics [COR], IBM Global Services [IGS], Personify [PER], WebSideStory [WSS], and WebTrends offer Web analytics as outsourced services. Such products/services also analyze e-commerce events such as products bought and advertisement click-through rates providing key performance indicators to the marketers of a company. Most of them include OLAP engines.

Software Providers – Integrated CRM Solutions. Accrue's HitList targets the midmarket by delivering sophisticated Web site monitoring and Insight 5 addresses to larger, more complex sites, enabling enterprises to monitor marketing campaigns, retain visitors, and determine browse-to-buy ratios.

Elytics software integrates Web log data with data from the client side and combines them with user metrics such as screen size, preferred language, and so on, providing a hybrid system that combines the advantages of both software systems and ASPs.

E.piphany's Enterprise Insight includes tools for analyzing Web and commerce server logs. It also provides the functionality for sharing data with E.piphany's personalization engine and integrating Web visitor information with customer data from other operational systems. Enterprise Insight can be used alone or as the analytical component of the E.5 System.

Lumio's Re:cognition product suite provides the IT infrastructure required to enhance the effectiveness of e-businesses with complementary products supporting behavioral data collection, analysis of data, creation of knowledge storage of data and knowledge, deployment of knowledge in real-time, and measurement of key performance indicators to continuously monitor the quality of the interactions with customers. All six products in the Re:cognition suite, namely, Re:collect, Re:store, Re:search, Re:order, Re:action, and Re:view are standards-compliant.

NCR's E-Business Teradata @ctive Warehouse includes utilities that convert Web log, registration, demographic, and operational data. It also provides OLAP tools and verticalized reporting software. In addition, it can serve as a back-end to personalization engines.

NetGenesis' E-Metrics Solutions Suite provides a set of business performance indicators such as recency, frequency, monetary value, and duration that enable enterprises to evaluate their Web sites. This combination of customer, financial, and Web site metrics is called E-Metrics and can be delivered in the form of reports produced using the InfraLens reporting software. This reporting can be personalized for each user separately.

Net Perceptions' E-commerce Analyst examines visitor patterns to find correlations between customers and products. The whole process consists of four routines, namely, data preparation, data transformation, data mining, and analytics, enabling the optimization of cross-selling.

Funnel Web products, provided by Quest, enable the creation of sessions extracted from Web server logs and the production of reports that describe visitor information and behavior. The software supports multiple languages and the ability of report customization.

NetTracker, provided by Sane Solutions, is a powerful tool for Web analytics. Its editions that target the low and midmarket allow the storage of detailed data instead of summary data that other relevant products provide. Its high-end edition allows the integration of Web log data with operational data from CRM and ERP systems.

SAS is the provider of a set of tools and applications that enable analytical CRM, personnel management, data warehousing, and data mining. Its Web analytics solution is WebHound, which extracts information from Web logs, performing clickstream analysis. Engage Profile Server generates anonymous profiles of visitors, enabling the personalization of services. E-Discovery is an integrated CRM solution. It enables the integration of clickstream data with purchasing, customer service, demographic, and psychographic data about the company's customers.

WebTrends was one of the first suppliers of inexpensive and popular Web analyzers, with its WebTrends Log Analyzer software, an application that analyzes single-server sites. Its more advanced product is Commerce Trends, a platform that sessionizes Web log data, loads it into an RDB, and allows the creation of standardized or customized reports. WebTrends offers its software solutions as hosted solutions via its WebTrends Live ASP Service.

ASPs. Coremetrics' eLuminate is a service that receives data from JavaScript embedded in Web pages using cookies for identifying visitors and produces reports about campaign and merchandising effectiveness of the enterprise.

Surfaid Analytics, the Web analytics service provided by IBM focuses on supplying OLAP and data mining capabilities in addition to structured reports. The system filters log data and creates sessions by reconstructing every visitor's path through the site by combining information such as the visitor's IP address, timestamps, user agent strings, and cookies. These data are then stored in a relational "cube" and IBM's clustering software identifies visitor segments.

Personify's Central is an ASP model that offers the same services as Profit Platform, the software provided by Personify for Web analytics. Thus it enables the filtering and integration of Web log, commerce server, registration, and other data into a database of profiles that are then used for producing standardized reports as well as performing OLAP operations.

WebSideStory was one of the first Web analytics ASPs. Its HitBox solution consists of code embedded in the clients' Web pages that sends data to their servers for further analysis. Depending on the edition the services provided range from simple Web statistics to deeper statistical analysis and extended features.

4.7 Conclusions

Web usage mining is the process of applying statistical and data mining methods to Web log data in order to extract useful patterns concerning the users' navigational behavior, user and page clusters, as well as possible correlations between Web pages and user groups.

The discovered rules and patterns can then be used for improving the system's performance or for making modifications to the Web site. The information included in the Web logs can also be integrated with customer data collected from CRM and ERP systems, in order to gather business intelligence.

Several issues must be taken into consideration, including decisions to be made during data filtering and processing, user and session identification, and pageview identification. Another important issue is the choice of the data mining methods that should be used.

Web usage mining lately has been used in combination with other technologies, such as user profiling and in some cases content mining, in order to provide a more integrated view of the usage of a Web site, and make personalization more effective.

5. RESEARCH INITIATIVES

Recently, many research projects are dealing with Web usage mining and Web personalization areas. Most of the efforts focus on extracting useful patterns and rules using data mining techniques in order to understand the users' navigational behavior, so that decisions concerning site restructuring or modification can then be made by humans. In several cases, a recommendation engine helps the user navigate through a site. Some of the more advanced systems provide much more functionality, introducing the notion of adaptive Web sites and providing means of dynamically changing a site's structure. All research efforts combine more than one of the aforementioned methods in Web personalization, namely, user profiling, Web usage mining techniques, content management and publishing mechanisms. In the sequel we provide a brief description of the most important research efforts in the Web mining and personalization domain. A summarized overview of the research initiatives and the Web personalization domains they investigate is presented in Appendix Table AIII.

One of the earliest attempts to take advantage of the information that can be gained through exploring a visitor's navigation through a Web site resulted in Letizia [Lieberman 1995], a client-site agent that monitors the user's browsing behavior and searches for potentially interesting pages for recommendations. The agent looks ahead at the neighboring pages using a best-first search augmented by heuristics inferring user interest, inasmuch as they're derived from the user's navigational behavior, and offers suggestions.

An approach for automatically classifying a Web site's visitors according to their access patterns is presented in the work of Yan et al. [1996]. The model they propose consists of two modules; an offline module that performs cluster analysis on the Web logs and an online module aiming at dynamic link generation. Every user is assigned to a single cluster based on his current traversal patterns. The authors have implemented the offline module (Analog) and have given a brief description of the way the online module should function.

One of the most popular systems from the early days of Web usage mining is WebWatcher [Joachims et al. 1997]. The idea is to create a tour guide agent that provides navigation hints to the user through a given Web collection, based on its knowledge of the user's interests, the location and relevance of various items in the location, as well as the way in which other users have interacted with the collection in the past. The system starts by profiling the user, acquiring information about her interests. Each time the user requests a page, this information is routed through a proxy server in order to easily track the user session across the Web site and any links believed to be of interest for the user are highlighted. Its strategy for giving advice is learned from feedback from earlier tours. A similar system is the Personal WebWatcher [Mladenic 1999], which is structured to specialize for a particular user, modeling his

interests. It solely records the addresses of pages requested by the user and highlights interesting hyperlinks without involving the user in its learning process, asking for keywords or opinions about pages as WebWatcher does.

Chen et al. [1996] introduce the "maximal forward reference" concept in order to characterize user episodes for the mining of traversal patterns. Their work is based on statistically dominant paths and association rules discovery, and a maximal forward reference is defined as the sequence of pages requested by a user up to the last page before backtracking. The SpeedTracer project [Wu et al. 1998] is built on the work proposed by Chen et al. [1996]. SpeedTracer uses the referrer page and the URL of the requested page as a traversal step and reconstructs the user traversal paths for session identification. Each identified user session is mapped into a transaction and then data mining techniques are applied in order to discover the most frequent user traversal paths and the most frequently visited groups of pages.

A different approach is adopted by Zaiane et al. [1998]. The authors combine the OLAP and data mining techniques and a multidimensional data cube, to extract interactively implicit knowledge. Their WebLogMiner system after filtering the data contained in the Web log, transforms them into a relational database. In the next phase a data cube is built, each dimension representing a field with all possible values described by attributes. OLAP technology is then used in combination with data mining techniques for prediction, classification, and time-series analysis of Web log data. Huang et al. [2001] also propose the use of a cube model that explicitly identifies Web access sessions, maintains the order of the session's components and uses multiple attributes to describe the Web pages visited. Borges and Levene [1999] model the set of user navigation sessions as a hypertext probabilistic grammar whose higher probability generated strings correspond to the user's preferred trails. Shahabi et al. [1997] propose the use of a client-side agent that captures the client's behavior creating a profile. Their system then creates clusters of users with similar interests.

Joshi et al. [2000; Krishnapuram et al. 2001; Nasraoui et al. 2000] introduce the notion of uncertainty in Web usage mining, discovering clusters of user session profiles using robust fuzzy algorithms. In their approach, a user or a page can be assigned to more than one cluster. After preprocessing the log data, they create a dissimilarity matrix that is used by the fuzzy algorithms presented in order to cluster typical user sessions. To achieve this, they introduce a similarity measure that takes into account both the individual URLs in a Web session, as well as the structure of the site.

Cooley et al. [1999b; Srivastava et al. 2000] define Web usage mining as a three-phase process, consisting of preprocessing, pattern discovery, and pattern analysis. Their prototype system, WebSIFT, first performs intelligent cleansing and preprocessing for identifying users, server sessions, and inferring cached page references through the use of the referrer field, and also performs content and structure preprocessing [Cooley et al. 1999a]. Pattern discovery is accomplished through the use of general statistic algorithms and data mining techniques such as association rules, sequential pattern analysis, clustering, and classification. The results are then analyzed through a simple knowledge query mechanism, a visualization tool, or the information filter, that makes use of the preprocessed content and structure information to automatically filter the results of the knowledge discovery algorithms.

Masseglia et al. [1999a,b] apply data mining techniques such as association rules and sequential pattern discovery on Web log files and then use them to customize the server hypertext organization dynamically. They regard Web usage mining as a two-phase process, consisting of the preprocessing phase where all irrelevant data are removed and log file entries are clustered based on time considerations, and the Web mining phase where data mining techniques are applied. The prototype system, WebTool, also provides a visual query language in order to improve the mining process. A generator of dynamic links uses the rules generated from sequential patterns or association rules, and each time the navigation pattern of a visitor matches a rule, the hypertext organization is dynamically modified. In a recent work [Masseglia et al. 2000], the problem of incremental Web usage mining is addressed. Using the ISEWUM method, they handle the problem of mining user patterns when new

transactions are added to the Web log file by only considering user patterns obtained by an earlier mining.

Buchner and Mulvenna [1998] present a knowledge discovery process in order to discover marketing intelligence from Web data. They propose an environment that combines existing online analytical mining, as well as Web usage mining approaches and incorporates marketing expertise. For this purpose, a generic Web log data hypercube is defined. In a more recent work Buchner et al. [1999] introduce the data mining algorithm MiDAS for discovering sequential patterns from Web log files, in order to perceive behavioral marketing intelligence. In this work, domain knowledge is described as flexible navigation templates that specify navigational behavior, as network structures for the capture of Web site topologies, as well as concept hierarchies and syntactic constraints.

Spiliopoulou et al. [Spiliopoulou and Faulstich 1998; Spiliopoulou et al. 1999; Spiliopoulou 2000] have designed MINT, another mining language for the implementation of WUM, a sequence mining system for the specification, discovery, and visualization of interesting navigation patterns. The Web log is preprocessed and an "aggregate materialized view" of the Web log is stored. In the data preparation phase, except for log data filtering and completion, user sessions are identified using timeout mechanisms. The path each user follows is called a "trail". Because many users access the same pages in the same order (creating similar trails), an "aggregate tree" is constructed by merging trails with the same prefix. This tree is called an "aggregated log" and navigation patterns of interest can be extracted using MINT. This language supports the specification of criteria of statistical, structural, and textual features.

Berendt [2000, 2001] has implemented STRATDYN, an add-on module that extends WUM's capabilities by identifying the differences between navigation patterns and exploiting the site's semantics in the visualization of the results. In this approach, concept hierarchies are used as the basic method of grouping Web pages together. The accessed pages or paths are abstracted, because Web pages are treated as instances of a higher-level concept, based on page content, or by the kind of service requested. An "interval-based coarsening" technique is used in order to mine Web usage at different levels of abstraction using basic and coarsened stratograms for the visualization of the results.

Coenen et al. [2000] propose a framework for self-adaptive Web sites, taking into account the site structure except for the site usage. The authors underline the distinction between strategic changes, referring to the adaptations that have important influence on the original site structure, and tactical changes, referring to the adaptations that leave the site structure unaffected. The proposed approach is based on the fact that the methods used in Web usage mining produce recommendations including links that don't exist in the original site structure, resulting in the violation of the beliefs of the site designer and the possibility of making the visitor get lost following conceptual but not real links. Therefore, they suggest that any strategic adaptations based on the discovery of frequent item sets, sequences, and clusters should be made offline and the site structure should be revised. On the other hand, as far as the tactical adaptations are concerned, an algorithm for making online recommendations leaving the site structure unaffected is proposed.

Perkowitz and Etzioni [1998, 1999, 2000a] were the first to define the notion of adaptive Web sites as sites that semiautomatically improve their organization and presentation by learning from visitor access patterns [Perkowitz and Etzioni 1997]. The system they propose semiautomatically modifies a Web site, allowing only nondestructive transformations. Therefore, nothing is deleted or altered; instead, new index pages containing collections of links to related but currently unlinked pages are added to the Web site. The authors propose PageGather, an algorithm that uses a clustering methodology to discover Web pages visited together and to place them in the same group. In a more recent work [Perkowitz and Etzioni 2000b], they move from the statistical cluster-mining algorithm PageGather to IndexFinder, which fuses statistical and logical information to synthesize index pages. In this latter work, they formalize the problem of index page synthesis as a conceptual clustering problem and try to discover coherent and cohesive link sets that can be represented to a human Webmaster

as candidate index pages. The difference is that information is also derived from the site's structure and content. Therefore, IndexFinder combines the statistical patterns gleaned form the log file with logical descriptions of the contents of each Web page in order to create index pages.

Cingil et al. [2000] describe an architecture that provides a broader view of personalization, through the use of various W3C standards. They describe how standards such as XML, RDF, and P3P can be used to create personalization applications. In this architecture, a log of the user's navigation history is created as a "user agent" at the client site gathers clickstream information about the user. This information is kept in an XML file, creating a user profile that reflects the user's interests and preferences. Privacy of the user is preserved through P3P. On the server side statistical modeling is run on user profiles to match up visitors that seem to have similar interests and preferences so that the most likely content or products can be recommended to a user based on these similarities. The user profile is exploited by the user agent to discover resources on the Internet that may be of interest to the user as well as obtaining personalized information from the resources. When the metadata of the resources are expressed in RDF, it will be a lot easier for agents to discover the resources on the Web that match the user profiles. Until then, metadata tags of HTML are used in the proposed system.

The most advanced system is the WebPersonalizer, proposed by Mobasher et al. [1999, 2000a]. WebPersonalizer provides a framework for mining Web log files to discover knowledge for the provision of recommendations to current users based on their browsing similarities to previous users. It relies solely on anonymous usage data provided by logs and the hypertext structure of a site. After data gathering and preprocessing (converting the usage, content, and structure information contained in the various data sources into various data abstractions), data mining techniques such as association rules, sequential pattern discovery, clustering, and classification are applied, in order to discover interesting usage patterns. The results are then used for the creation of aggregated usage profiles, in order to create decision rules. The recommendation engine matches each user's activity against these profiles and provides him with a list of recommended hypertext links.

This framework has been recently extended [Mobasher et al. 2000b,c] to incorporate content profiles into the recommendation process as a way to enhance the effectiveness of personalization actions. Usage and content profiles are represented as weighted collections of pageview records. The content profiles represent different ways in which pages with partly similar content can be grouped. The overall goal is to create a uniform representation for both content and usage profiles in order to integrate them more easily. The system is divided into two modules; the offline, which is comprised of data preparation and specific Web mining tasks, and the online component, which is a real-time recommendation engine.

6. CONCLUSIONS

Web personalization is the process of customizing the content and structure of a Web site to the specific and individual needs of each user, without requiring them to ask for it explicitly. This can be achieved by taking advantage of the user's navigational behavior, as revealed through the processing of Web usage logs, as well as the user's characteristics and interests. Such information can be further analyzed in association with the content of a Web site, resulting in improvement of the system performance, users' retention, and/or site modification.

The overall process of Web personalization consists of five modules, namely: user profiling, log analysis and Web usage mining, information acquisition, content management, and Web site publishing.

User profiling is the process of gathering information specific to each visitor to a Web site either implicitly, using the information hidden in the Web logs or technologies such as cookies, or explicitly, using registration forms, questionnaires, and the like. Such information can be demographic, personal, or even information concerning the user's navigational behavior. However, many of the methods used in

user profiling raise some privacy issues concerning the disclosure of the user's personal data, therefore they are not recommended. Because user profiling seems essential in the process of Web personalization, a legal and more accurate way of acquiring such information is needed. P3P is an emerging standard recommended by W3C that provides a technical mechanism that enables users to be informed about privacy policies before they release personal information and gives them control over the disclosure of their personal data.

The main component of a Web personalization system is the usage miner. Log analysis and Web usage mining is the procedure where the information stored in the Web server logs is processed by applying statistical and data mining techniques such as clustering, association rules discovery, classification, and sequential pattern discovery, in order to reveal useful patterns that can be further analyzed. Such patterns differ according to the method and the input data used, and can be user and page clusters, usage patterns, and correlations between user groups and Web pages. Those patterns can then be stored in a database or a data cube and query mechanisms or OLAP operations can be performed in combination with visualization techniques. The most important phase of Web usage mining is data filtering and preprocessing. In that phase, Web log data should be cleaned or enhanced, and user, session, and pageview identification should be performed.

Web personalization is a domain that has been recently gaining great momentum not only in the research area, where many research teams have addressed this problem from different perspectives, but also in the industrial area, where there exists a variety of tools and applications addressing one or more modules of the personalization process. Enterprises expect that by exploiting the information hidden in their Web server logs they may discover the interactions between their Web site visitors and the products offered through their Web site. Using such information, they can optimize their site in order to increase sales and ensure customer retention. Apart from Web usage mining, user profiling techniques are also employed in order to form a complete customer profile. Lately, there has been an effort to incorporate Web content in the recommendation process, in order to enhance the effectiveness of personalization. However, a solution that efficiently combines techniques used in user profiling, Web usage mining, content acquisition, and management as well as Web publishing has not yet been proposed.

APPENDIX A. ACRONYMS AND ABBREVIATIONS

ASP	Application Service Provider
CRM	Customer Relationship Manager
ERP	Enterprise Resource Planning
HTML	Hypertext Markup Language
HTTP	Hypertext Transport Protocol
IP	Internet Protocol
ISP	Internet Service Provider
OLAP	OnLine Analytical Processing
P3P	Platform for Privacy Preferences
RDF	Resource Description Framework
URL	Uniform Resource Locator
W3C	World Wide Web Consortium
XML	eXtensible Markup Language

APPENDIX B. WEB REFERENCES

[ACC] Accrue http://www.accrue.com
[ALT] Alta-Vista http://www.altavista.com
[AMA] amazon.com http://www.amazon.com
[ANA] Analog http://www.analog.cx
[APP] Apple Computer http://www.apple.com
[AWS] AWSD.com http://awsd.com

[BRO] Broadvision http://www.broadvision.com
[CAP] capecom http://www.cape.com
[CDN] CDNOW http://www.cdnow.com
[COR] Coremetrics http://www.coremetrics.com
[DCL] DoubleClick Inc. http://www.doubleclick.com
[DEL] Dell http://www.dell.com
[ELY] Elytics http://www.elytics.com
[EPI] E.piphany http://www.epiphany.com
[EXC] Excite http://www.excite.com
[FOO] Food.com http://www.food.com
[HAP] Axolot http://www.axolot.com
[IGS] IBM Global Services http://Surfaid.dfw.ibm.com
[LIE] Liebhart Systems http://www.liebhart.com
[LUM] Lumio http://www.lumio.com
[LYC] Lycos http://www.lycos.com
[MAC] Macromedia http://www.macromedia.com
[MAH] Mach 5 http://Mach5.com
[MNO] mnot http://www.mnot.net
[MON] Monocle Solutions http://www.monocle-solutions.com
[MSF] Microsoft http://www.microsoft.com
[MSN] Microsoft Network http://www.msn.com
[NCR] NCR Corporation http://www.ncr.com
[NGE] NetGenesis http://www.netgen.com
[NME] Neuromedia http://www.neuromedia.com
[NPE] NetPerceptions http:///www.netperceptions.com
[NTR] Net Tracker http://www.sane.com
[OSE] Open Sesame http://Sesame.com
[PER] Personify http://www.personify.com
[QUE] Quest http://www.quest.com
[STS] Ststat http://awsd.com/scripts/weblog
[SAN] Sane solutions http://www.sane.com
[SAS] SAS http://www.sas.com
[SAW] Sawmill http://www.flowerfire.com
[SUR] SurfStats http://www.surfstats.com
[WSS] WebSideStory http://www.Websidestory.com
[WTR] Web Trends http://www.webtrends.com
[WUM] WUM http://wum.wiwi.hu-berlin.de
[YAH] Yahoo! http://www.yahoo.com

APPENDIX C – COMMERCIAL AND RESEARCH TOOLS AND APPLICATIONS

Table AI. User Profiling Tools

Vendor	Product Name	Collaborative Filtering	Page Customization	Cookies	User Registration
BroadVision [BRO]	One-To-One		*		
Macromedia [MAC]	LikeMinds	*			
Microsoft [MSF]	Firefly Passport	*			*
NetPerceptions [NPE]	Group Lens	*			
Neuromedia [NME]	NeuroStudio		*	*	*
OpenSesame [OSE]	Learn Sesame		*	*	

Table AII. Log Analyzers and Web Usage Miners

Vendor	Product Name	Data Source	ASP	Software	Complete CRM Solution
Analog [ANA]	Analog	Server (Web logs)		* (Freeware log analyser)	
Accrue [ACC]	HitList, Insight 5	Server (Web logs)			*
Coremetrics [COR]	Eluminate	Client	*		
Elytics [ELY]	Elytics Analysis Suite	Client, server, other enterprise systems		* (Hybrid, incorporating ASP technology)	
E.piphany [EPI]	Enterprise Insight, E.5	Server (Web logs), operational data (ERP)			*
Follow [MNO]	Follow 2	Server (Web logs)		* (Freeware log analyser)	
IBM Global Services [IGS]	Surfaid Analytics	Client, server	*		
Lumio [LUM]	Re: cognition suite	Server			*
NCR Corporation [NCR]	E-business Teradata @ctive Warehouse	Server (Web logs), registration data, operational data (CRM, ERP) etc.			*
NetGenesis [NGE]	NetGenesis 5 E-Metrics Solutions	Server (Web logs, packet sniffers, server plug-ins)			*
NetPerceptions [NPE]	E-commerce Analyst	Server (Web logs)		*	
Personify [PER]	Profit Platform (s/w), Central (ASP)	Server (Web logs), commerce server data, registration data	*	*	
Quest [QUE]	Funnel Web	Server (Web logs)		*	
Sane solutions [SAN]	NetTracker	Server (Web logs)		*	
SAS [SAS]	WebHound, e-Discovery, Engage ProfileServer	Server (Web logs), operational data, demographic data			*
WebSideStory [WSS]	HitBox	Client (browser)	*		
WebTrends [WTR]	WebTrends Log Analyzer, Commerce Trends, Web Trends Live (ASP)	Server (Web logs), client (ASP solution)	*	*	* (High-end product)

Table AIII. Research Initiatives

Project Name	Data Source	User Profiling	Web Usage Mining	Content Management	Publishing Mechanism
Letizia [Lieberman 1995]	Client		*	(*)	
WebWatcher [Joachims et al. 1997]	Proxy	*	*	(*)	
Analog [Yan et al. 1996]	Server		*		* (Suggested)
SpeedTracer [Wu et al. 1998]	Server		*		
WebLogMiner [Zaiane et al. 1998]	Server		*		
Borges and Levene [1999]	Server		*		
Shahabi et al. [1997]	Client	*	*		
Joshi et al. [2000; Krishnapuram et al. 2001; Nasraoui et al. 2000]	Server	*	*		
WebSIFT [Cooley et al. 1999b,a; Srivastava et al. 2000]	Server		*	(*)	
WebTool [Masseglia et al. 1999a,b, 2000]	Server		*		* (Suggested)
Buchner et al. [Buchner and Mulvenna 1998; Buchner et al. 1999]	Server	*	*	*	
WUM [Spiliopoulou and Faulstich 1998; Spiliopoulou et al. 1999; Spiliopoulou 2000]	Server		*		
STRATDYN [Berendt 2000, 2001]	Server		*	*	
Coenen et al. [2000]	Server		*		* (Suggested)
Adaptive Web Sites [Perkowitz and Etzioni 1999, 2000]	Server		*	*	*
Cingil et al. [2000]	Client		*	*	* (Suggested)
WebPersonalizer [Mobasher et al. 1999, 2000a]	Server	*	*		*
Mobasher et al. [2000b,c]	Server	*	*	*	*

ACKNOWLEDGEMENTS

We are grateful to Alkis Polyzotis and Stratis Valavanis for taking time to read carefully drafts of this paper and provide us with valuable comments.

REFERENCES

BERENDT, B. 2000. Web usage mining, site semantics, and the support of navigation. In *Proceedings of the Workshop WEBKDD'2000 Web Mining for E-Commerce—Challenges and Opportunities, Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Boston, August).

BERENDT, B. 2001. Understanding Web usage at different levels of abstraction: Coarsening and visualizing sequences. In *Proceedings of the Workshop WEBKDD 2001 Mining Log Data Across All Customer TouchPoints, Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, August).

BORGES, J. AND LEVENE, M. 1999. Data mining of user navigation patterns. In*Web Usage Analysis and User Profiling*, Lecture Notes in Computer Science, vol. 1836, Springer-Verlag New York, 92–111.

BUCHNER, A. AND MULVENNA, M. D. 1998. Discovering Internet marketing intelligence through online analytical Web usage mining. *SIGMOD Rec. 27,* 4, 54–61.

BUCHNER, A.G., BAUMGARTEN, M., ANAND, S.S., MULVENNA, M..D., AND HUGHES, J.G. 1999. Navigation pattern discovery from Internet data. In *Proceedings of theWeb Usage Analysis and User Profiling Workshop (WEBKDD'99), Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Diego, August), 25–30.

CHEN, M. S., PARK, J. S., AND YU, P. S. 1996. Data mining for path traversal patterns in a web environment. In *Proceedings of the Sixteenth International Conference on Distributed Computing Systems* (May), 385–392.

CINGIL, I., DOGAC, A., AND AZGIN, A. 2000. A broader approach to personalization. *Commun. ACM, 43*, 8 (August), 136–141.

COENEN, F., SWINNEN, G., VANHOOF, K., AND WETS, G. 2000. A framework for self adaptive websites: Tactical versus strategic changes. In *proceedings of WEBKDD'2000 Web Mining for ECommerce—Challenges and Opportunities, Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Boston, August).

COOLEY, R., MOBASHER, B., AND SRIVASTAVA, J. 1999a. Data preparation for mining world wide web browsing patterns. *Knowl. Inf. Syst., 1*, 1 (Feb.).

COOLEY, R., TAN, P-N., AND SRIVASTAVA, J. 1999b. WebSIFT: The web site information filter system. In *Proceedings of the Web Usage Analysis and User Profiling Workshop (WEBKDD'99), Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Boston, August).

DEAN, R. 1998. Personalizing your web site. Available at http://Webbuilder.netscape.com/Web-building/pages/Business/Personal/index.html.

HUANG, Z., NG, J., CHEUNG, D.W., NG, M.K., AND CHING, W-K. 2001. A cube model for web access sessions and cluster analysis. In *Proceedings of the Mining Log Data Across All Customer TouchPoints Workshop (WEBKDD'01) Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, August).

JOACHIMS, T., FREITAG, D., AND MITCHELL, T. 1997. WebWatcher: A tour guide for the world wide web. In *Proceedings of IJCAI97* (August).

JOSHI, A., JOSHI, K., AND KRISHNAPURAM, R. 2000. On mining web access logs. In *Proceedings of the 2000 ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery*, 63–69.

KRISHNAPURAM, R., JOSHI, A., NASRAOUI, O., AND YI, L. 2003. Low-complexity fuzzy relational clustering algorithms for web mining, IEEE Trans. Fuzzy Syst. *9,* 4, 596-607.

LIEBERMAN, H. 1995. Letizia: An agent that assists web browsing. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence* (Montreal).

MASSEGLIA, F., PONCELET, P., AND CICCHETTI, R. 1999a. WebTool: An integrated framework for data mining. In *Proceedings of the Ninth International Conference on Database and Expert Systems Applications (DEXA'99)* (Florence, Italy, August), 892–901.

MASSEGLIA, F., PONCELET, P., AND TEISSEIRE, M. 1999b. Using data mining techniques on web access logs to dynamically improve hypertext structure. In *ACM SigWeb Lett., 8*, 3 (Oct.) 13–19.

MASSEGLIA, F., PONCELET, P., AND TEISSEIRE, M. 2000. Web usage mining: How to efficiently manage new transactions and new customers. In *Proceedings of the Fourth European Conference on Principles of Data Mining and Knowledge Discovery (PKDD'00)* (Lyon, France, Sept.).

MLADENIC, D. 1999. Machine learning used by personal webwatcher. In *Proceedings Of ACAI-99 Workshop on Machine Learning and Intelligent Agents* (Chania, Greece, July).

MOBASHER, B., COOLEY, R., AND SRIVASTAVA, J. 1999. Creating adaptive web sites through usagebased clustering of URLs. In *Proceedings of the 1999 IEEE Knowledge and Data Engineering Exchange Workshop (KDEX'99)* (Nov.).

MOBASHER, B., COOLEY, R., AND SRIVASTAVA, J. 2000a. Automatic personalization based on web usage mining. *Commun. ACM, 43,* 8 (August), 142–151.

MOBASHER, B., DAI, H., LUO, T., SUNG, Y., AND ZHU, J. 2000b. Discovery of aggregate usage profiles for web personalization. In *Proceedings of the Web Mining for E-Commerce Workshop (WEBKDD'2000), Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Boston, August).

MOBASHER, B., DAI, H., LUO, T., SUNG, Y., AND ZHU, J. 2000c. Integrating web usage and content mining for more effective personalization. In *Proceedings of the International Conference on Ecommerce and Web Technologies (ECWeb2000)*. (Greenwich, UK, Sept.).

MULVENNA, M. D., ANAND, S. S., AND BUCHNER, A. G. 2000. Personalization on the net using web mining. *Commun. ACM, 43*, 8 (August), 123–125.

NASRAOUI, O., FRIGUI, H., KRISHNAPURAM, R., AND JOSHI, A. 2000. Extracting web user profiles using relational competitive fuzzy clustering. *Int. J. Arti. Intell. Tools 9*, 4.

P3P. Platform for Privacy Preferences Project. Available at http://www.w3.org/P3P.

PERKOWITZ, M. AND ETZIONI, O. 1997. Adaptive web sites: An AI challenge. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence* (Nagoya, Japan).

PERKOWITZ, M. AND ETZIONI, O. 1998. Adaptive web sites: Automatically synthesizing web pages. In *Proceedings of the Fifteenth National Conference on Artificial Intelligence* (Madison, WI, July).

PERKOWITZ, M. AND ETZIONI, O. 1999. Adaptive web sites: Conceptual cluster mining. In *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence (IJCAI99)* (Stockholm).

PERKOWITZ, M. AND ETZIONI, O. 2000a. Towards adaptive web sites: Conceptual framework and case study. In *Artif. Intell. 118*, 1–2, 245–275.

PERKOWITZ, M. AND ETZIONI, O. 2000b. Adaptive web sites. *Commun. ACM, 43*, 8 (August), 152–158.

RFC. Identification Protocol. Available at http://www.rfc-editor.org/rfc/rfc1413.txt.

SHAHABI, C., ZARKESH, A. M., ADIBI, J., AND SHAH, V. 1997. Knowledge discovery for users web-page navigation. In *Workshop on Research Issues in Data Engineering* (Birmingham, UK).

SPILIOPOULOU, M. 2000. Web usage mining for web site evaluation. *Commun. ACM 43*, 8 (August), 127–134.

SPILIOPOULOU, M. AND FAULSTICH, L. C. 1998. WUM: A web utilization miner. In *Proceedings of the International Workshop on the Web and Databases* (Valencia, March).

SPILIOPOULOU, M., FAULSTICH, L. C., AND WILKLER, K. 1999. A data miner analyzing the navigational behavior of web users. In *Proceedings of the Workshop on Machine Learning in User Modelling of the ACAI99* (Chania, Greece, July).

SRIVASTAVA, J., COOLEY, R., DESHPANDE, M., AND TAN, P-N. 2000. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations 1*, 2 (Jan.), 12–23.

W3CLOG. Extended log file format. Available at http://www.w3.org/TR/WD-logfile.html.

WCA. Web characterization terminology & definitions. Available at http://www.w3.org/1999/05/WCA-terms/.

WU, K-L., YU, P. S., AND BALLMAN, A. 1998. SpeedTracer: A web usage mining and analysis tool. *IBM Syst. J. 37*, 1.

YAN, T. W., JACOBSEN, M., GARCIA-MOLLINA, H., AND DAYAL, U. 1996. From user access patterns to dynamic hypertext linking. *In Fifth International World Wide Web Conference (WWW5)* (Paris).

ZAIANE, O. R., XIN, M., AND HAN, J. 1998. Discovering web access patterns and trends by applying OLAP and data mining technology on web logs. In *Proceedings of Advances in Digital Libraries Conference (ADL'98)* (Santa Barbara, CA, April).