Available online at www.sciencedirect.com
SCIENCE DIRECT®
Data & Knowledge Engineering 53 (2005) 225–241
DATA & KNOWLEDGE ENGINEERING
www.elsevier.com/locate/datak

Mining interesting knowledge from weblogs: a survey

Federico Michele Facca, Pier Luca Lanzi *

Dipartimento di Elettronica e Informazione, Artificial Intelligence and Robotics Laboratory, Politecnico di Milano, 20133 Milano, Italy

Received 21 August 2003; received in revised form 10 December 2003; accepted 4 August 2004
Available online 11 September 2004

Abstract

Web Usage Mining is that area of Web Mining which deals with the extraction of interesting knowledge from logging information produced by Web servers. In this paper we present a survey of the recent developments in this area that is receiving increasing attention from the Data Mining community.

Keywords: Machine learning; Web Mining

1. Introduction

The World Wide Web is an immense source of data that can come either from the *Web content*, represented by the billions of pages publicly available, or from the *Web usage*, represented by the log information daily collected by all the servers around the world. *Web Mining* [1] is that area of Data Mining which deals with the extraction of interesting knowledge from the World Wide Web. More precisely [2], *Web Content Mining* is that part of Web Mining which focuses on the raw information available in Web pages; source data mainly consist of textual data in Web pages

* Corresponding author.
E-mail addresses: facca@elet.polimi.it (F.M. Facca), lanzi@elet.polimi.it (P.L. Lanzi).

0169-023X/$ - see front matter
doi:10.1016/j.datak.2004.08.001

(e.g., words, but also tags); typical applications are *content-based* categorization and *content-based* ranking of Web pages. *Web Structure Mining* is that part of Web Mining which focuses on the structure of Web sites; source data mainly consist of the structural information present in Web pages (e.g., links to other pages); typical applications are *link-based* categorization of Web pages, ranking of Web pages through a combination of content and structure (e.g. [3]), and reverse engineering of Web site models. *Web Usage Mining* is that part of Web Mining which deals with the extraction of knowledge from server log files; source data mainly consist of the (textual) logs that are collected when users access Web servers and might be represented in standard formats (e.g., Common Log Format [4], Extended Log Format [5], LogML [6]); typical applications are those based on user modeling techniques, such as Web personalization, adaptive Web sites, and user modeling.

The recent years have seen the flourishing of research in the area of Web Mining and specifically of Web Usage Mining. Since the early papers published in the mid 1990s, more than 400 papers on Web Mining have been published; more or less 150 papers, of the overall 400, have been published before 2001; around 50% of these papers regarded Web Usage Mining. The first workshop entirely on this topic, WebKDD, was held in 1999. Since 2000, the published papers on Web Usage Mining are more than 150 showing a dramatic increase in the interest for this area.

There are a number of papers which provide an overview of what has happened in the area of Web Mining since 1996. Ref. [2] defines Web Mining, providing the categorization in Web Content Mining, Web Structure Mining, and Web Usage Mining; then it provides a survey mainly focused on the results in the area of Web Content Mining. Ref. [7] presents a survey of the research in the area of Web Usage Mining with a main focus on the available commercial solutions and on the WebSIFT project [8] (formerly known as Webminer). Recently, [9] has presented an overview of the Soft Computing techniques (e.g., neural networks, fuzzy logic, genetic algorithms, and rough sets) used in Web Mining applications with a specific focus on Web Content Mining; some examples of applications of this technique to Web Usage Mining are also presented.

This paper is a survey of the recent developments in the area of Web Usage Mining. In contrast with [2,7,9], we focus only on Web Usage Mining, specifically on the research results reported in the literature since 2000 and on the software currently available. This survey is based on more than 150 papers published since 2000 on the topic of Web Usage Mining. Since it is not possible to cite them all here we provide an on-line bibliography at [10]. The paper is organized as follows. Initially, we discuss the different types of Web usage data that can be collected from user navigation (Section 2). Then we focus on the preprocessing of collected log data (Section 3) in which Web logs are filtered for different purposes, such as, to sort out uninteresting data (e.g., access from Web spiders), to identify user sessions (e.g., by means of *cookies*), to store data into a relational database, or to provide a structure adequate to the next mining step. Next we overview two topics of Web Usage Mining which provide orthogonal viewpoints: the mining techniques (Section 4) and the applications (Section 5). In Section 6, we discuss the commercial and public software packages currently available for performing Web Usage Mining tasks. In Section 7, we provide a crossreference among the typical Web Usage Mining applications, the techniques employed, and the class of data sources involved. In Section 8, we discuss the privacy issues that arise when using Web Usage Mining applications that can accurately track users behavior. Finally, in Section 9 we present what we believe to be the future research trends in this area.

2. Data sources

Web Usage Mining applications are based on data collected from three main sources: (i) Web servers, (ii) proxy servers, and (iii) Web clients.

The server side. Web servers are surely the richest and the most common source of data. They can collect large amounts of information in their log files and in the log files of the databases they use. These logs usually contain basic information e.g.: name and IP of the remote host, date and time of the request, the request line exactly as it came from the client, etc. This information is usually represented in standard format e.g.: Common Log Format [4], Extended Log Format [5], LogML [4]. Sometimes databases are used instead of text files to store log information so to improve querying of massive log repositories [11,12].

When exploiting log information from Web servers, the major issue is the identification of users' sessions, i.e., how to group all the users' page requests (or click streams) so to clearly identify the paths that users followed during navigation through the web site. This task is usually quite difficult and it depends on the type of information available in log files. The most common approach is to use *cookies* to track down the sequence of users' page requests (see [13] for an overview of cookie standards). If cookies are not available, various heuristics [14] can be employed to reliably identify users' sessions. Note however that, even if cookies are used, it is still impossible to identify the exact navigation paths since the use of the *back* button is not tracked at the server level [15]. Section 3 overviews the techniques currently employed to tackle these problems.

Apart from Web logs, users' behavior can also be tracked down on the server side by means of TCP/IP packet sniffers. Even in this case the identification of users' sessions is still an issue, but the use of packet sniffers provides some advantages [16]. In fact: (i) data are collected in real time; (ii) information coming from different Web servers can be easily merged together into a unique log; (iii) the use of special buttons (e.g., the *stop* button) can be detected so to collect information usually unavailable in log files. Notwithstanding the many advantages, packet sniffers are rarely used in practice. Packet sniffers raise scalability issues on Web servers with high traffic [16], moreover they cannot access encrypted packets like those used in secure commercial transactions (through the Secure Socket Layer). Unfortunately, this limitation turns out to be quite severe when applying Web Usage Mining to e-businesses [17].

Probably, the best approach for tracking Web usage consists of directly accessing the server application layer, as done in [18]. Unfortunately, this is not always possible. First, there are issue related to the copyright of server applications. Most important, following this approach, Web Usage Mining applications must be tailored for the specific servers and have to take into account the specific tracking requirements.

The proxy side. Many Internet Service Providers (ISPs) give to their customer proxy server services to improve navigation speed through caching. In many respects, collecting navigation data at the proxy level is basically the same as collecting data at the server level. The main difference in this case is that proxy servers collects data of *groups of users* accessing *huge groups* of web servers. Even in this case, session reconstruction is difficult and not all users' navigation paths can be identified. However, when there is no other caching between the proxy server and the clients, the identification of users' sessions is easier.

The client side. Usage data can be tracked also on the client side by using Javascript, Java applets [19], or even modified browsers [20]. These techniques avoid the problems of users' sessions

identification and the problems caused by caching (like the use of the *back* button). In addition, they provide detailed information about actual user behaviors [15]. However, these approaches rely heavily on the users cooperation and raise many issues concerning the privacy laws, which are quite strict (see Section 8).

3. Preprocessing

Data preprocessing has a fundamental role in Web Usage Mining applications. Ref. [21] notices that even if preprocessing techniques are widely used in Web Usage Mining, the literature on this topic is still quite limited, and that the most complete reference on preprocessing [22] dates back to 1999.

The preprocessing of Web logs is usually complex and time demanding. It comprises four different tasks: (i) the data cleaning, (ii) the identification and the reconstruction of users' sessions, (iii) the retrieving of information about page content and structure, and (iv) the data formatting.

Data cleaning. This step consists of removing all the data tracked in Web logs that are useless for mining purposes [23,21] e.g.: requests for graphical page content (e.g., jpg and gif images); requests for any other file which might be included into a web page; or even navigation sessions performed by robots and Web spiders. While requests for graphical contents and files are easy to eliminate, robots and Web spiders navigation patterns must be explicitly identified. This is usually done for instance by referring to the remote hostname, by referring to the user agent, or by checking the access to the *robots.txt* file. However, some robots actually send a false user agent in HTTP request. In these cases, an heuristic based on navigational behavior can be used to separates robot sessions from actual users' sessions (see [24,25]). E.g. in [25] is evidenced that search engine navigational paths are characterized by breadth first navigation in the tree representing the Web site structure and by unassigned referrer (the referrer gives the site that the client reports having been referred from). The heuristic proposed is based on the previous assumption and a classification of navigations. Well known robots' navigational paths are used to train the classifier, and the model obtained is used to classify further navigational sessions even if there is no a priori knowledge about the user agent that generated them.

Session identification and reconstruction. This step consists of (i) identifying the different users' sessions from the usually very poor information available in log files and (ii) reconstructing the users' navigation path within the identified sessions. The complexity of this step can vary a lot depending on the quality and on the quantity of the information available in the Web logs [7].

Most of the problems encountered in this phase are caused by the caching performed either by proxy servers either by browsers. Proxy caching causes a single IP address (the one belonging to the proxy Server) to be associated with different users' sessions, so that it becomes impossible to use IP addresses as users identifiers. This problem can be partially solved by the use of cookies [26], by URL rewriting [27], or by requiring the user to log in when entering the Web site [21]. A cookie is a piece of information sent by a Web server to a Web browser. This information is stored on the user's computer as a text file. Cookies may contain a lot of information about users, among them the one we are interested in is the session identifier. This information can be asked by the Web server every time a user asks for a Web page and stored in the Web log together with the page request. There are situations, however, where cookies will not work. Some browsers, for

example, do not support cookies. Other browsers allow the user to disable cookie support. In such cases, URL rewriting can be used to track the user's session by including the session ID in URLs. URL rewriting involves finding all links that will be written back to the browser, and rewriting them to include the session ID. For example, a link such as <a href="/store/catalog"> can rewritten as <a href="/store/catalog;jsessionid=DA32242SSGE2"> so as to include the session ID information, i.e., DA32242SSGE2. Hence every time a user click on a link in the page, the rewritten form of the URL is sent to the server and stored in the Web log.

Web browser caching is a more complex issue. Logs from Web servers cannot include any information about the use of the *back* button. This can generate inconsistent navigation paths in the users' sessions. However, by using additional information about the Web site structure is still possible to reconstruct a consistent path by means of heuristics. For example as reported in [22] if a page request is made, and this page request is not directly linked to the previous page request, the referrer log can be checked to see from what page the request came from. If the page is in the user's recent history request is possible to assume that the user used the *back* button. And then based on this assumption is possible to reconstruct a complete and consistent navigational path. To solve both proxy and web caching issues, IBM has introduced within SurfAid [28] a Javascript called *Web Bug* which has to be included in each Web page. Every time the Web page is loaded, Web Bug sends a request to the server asking for a 1×1 pixel image; the request is generated with parameters identifying the Web page containing the script and a numeric random parameter; the overall request cannot be cached neither by the proxy neither by the browser but it is logged by the Web server so as to solve caching problems [21,28].

Because the HTTP protocol is stateless, it is virtually impossible to determine when a user actually leaves the Web site in order to determine when a session should be considered finished. This problem is referred to as *sessionization*. Ref. [14] described and compared three heuristics for the identification of sessions termination; two were based on the time between users' page requests, one was based on information about the referrer. Ref. [29] proposed an adaptive time out heuristic. Ref. [22] proposed a technique to infer the timeout threshold for the specific Web site. Other authors proposed different thresholds for time oriented heuristics based on empiric experiments. The most commonly used timeout threshold is 25.5 min (or near values) which was proposed in [20].

Content and structure retrieving. The vast majority of Web Usage Mining applications use the visited URLs as the main source of information for mining purposes. URLs are however a poor source of information since, for instance, they do not convey any information about the actual page content. Ref. [22] has been the first to employ content based information to enrich the Web log data. Ref. [22] introduced an additional categorization step in which Web pages are classified according to their content type; this additional information is then exploited during the mining of Web logs. If an adequate classification is not known in advance, Web Structure Mining techniques can be employed to develop one. As in search engines, Web pages are classified according to their semantic areas by means of Web Content Mining techniques; this classification information can then be used to enrich information extracted from logs. For instance, [30] proposes to use Semantic Web for Web Usage Mining: Web pages are mapped onto ontologies to add meaning to the frequently observed paths. Given a page in the Web site, we must be able to extract domain-level structured objects as semantic entities contained in this page. This task may involve the automatic extraction and classification of objects of different types into classes based on the

underlying domain ontologies. The domain ontologies themselves may be pre-specified, or may be learned automatically from available training data [31]. Given this capability, the transaction data can be transformed into a representation which incorporates complex semantic entities accessed by users during a visit to the site. Ref. [32] introduces concept-based paths as an alternative to the usual user navigation paths; concept-based path are a high level generalization of usual path in which common concepts are extracted by means of intersection of raw user paths and similarity measures. Ref. [33] proposes the use of information scent to improve the results of user modeling. The idea of information scent is borrowed from Web Content Mining and Web Structure Mining. Information scent [33] is defined as the "*imperfect*, *subjective perception of the value and cost of information sources obtained from proximal cues*, *such as Web links*, *or icons representing the content sources*". Ref. [34] presents experimental results showing that proper preprocessing cannot be performed without the use of additional information about the content and structure of the Web site, and that this information greatly improves the effectiveness of pattern analysis processes.

Data formatting. This is the final step of preprocessing. Once the previous phases have been completed, data are properly formatted before applying mining techniques. Ref. [35] stores data extracted from Web logs into a relational database using a click fact schema, so as to provide better support to log querying finalized to frequent pattern mining. Ref. [11] introduces a method based on signature tree to index log stored in databases for efficient pattern queries. A tree structure named WAP-tree is also introduced in [36] to register access sequence to Web pages, this structure is optimized to exploit the sequence mining algorithm developed by the same authors [36]. Ref. [37] stores log data in another tree structure, the FBP-tree, to improve sequence pattern discovery. Ref. [38] uses a cube-like structure to store session information, to improve the extraction of *cube slices* used by clustering techniques.

4. Techniques

Most of the commercial applications of Web Usage Mining exploit consolidated statistical analysis techniques. In contrast, research in this area is mainly focused on the development of knowledge discovery techniques specifically designed for the analysis of Web usage data. Most of this research effort focuses on three main paradigms: association rules, sequential patterns, and clustering (see [39] for a detailed description of these techniques).

Association rules. These are probably the most elementary data mining technique and, at the same time, the most used technique in Web Usage Mining. Association rules are implications of the form $X \Rightarrow Y$ where the rule *body* X and the rule *head* Y are set of *items* within a set of *transactions*. The rule $X \Rightarrow Y$ states that the transactions which contain the items in X are *likely* to contain also the items in Y. When applied to Web Usage Mining, association rules are used to find associations among Web pages that frequently appear together in users' sessions. The typical result has the form:

"*A.html*, *B.html* $\Rightarrow$ *C.html*"

which states that if a user has visited page *A.html* and page *B.html*, it is very likely that in the same session the same user has also visited page *C.html*. This type of result is for instance produced by [12] and [40] by using a modification of the Apriori algorithm [39]. Ref. [41] proposes and

evaluates measures of interest to evaluate the association rules mined from Web usage data. Ref. [42] exploits a mixed technique of association rules and fuzzy logic to extract *fuzzy association rules* from Web logs.

Sequential patterns are used to discover frequent subsequences among large amount of sequential data. In Web Usage Mining, sequential patterns are exploited to find sequential navigation patterns that appear in users' sessions frequently. The typical sequential pattern has the following form [43]: the 70% of users who *first* visited *A.html* and *then* visited *B.html* afterwards, have also accessed page *C.html* in the same session. Sequential patterns might appear *syntactically* similar to association rules; in fact algorithms to extract association rules can also be used for sequential pattern mining. However, sequential patterns include the notion of time, i.e., at which point of the sequence a certain event happened. In the above example, pages *A*, *B*, and *C* appears sequentially, one after another, in the user sessions; in the previous example on association rules, information about the event sequence is not considered.

There are essentially two class of algorithms that are used to extract sequential patterns: one includes methods based on association rule mining; the other one includes methods based on the use of tree structures and Markov chains to represent navigation patterns. Some well-known algorithms for mining association rules have been modified to extract sequential patterns. For instance, [41,44] used *AprioriAll* and *GSP*, two extensions of the *Apriori* algorithm for association rules mining [39]. Ref. [36] argues that algorithms for association rule mining (e.g., Apriori) are not efficient when applied to long sequential patterns, which is an important drawback when working with Web logs. Accordingly, [36] proposes an alternative algorithm in which tree structures (WAP-tree) are used to represent navigation patterns. The algorithm (WAP-mine) [36] and the data structure (WAP-tree), specifically tailored for mining Web access patterns, WAP-mine outperforms other Apriori-like algorithms [36] like GSP. Tree structures are also used in [37]. Ref. [44] provides a comparison of different three sequential pattern algorithms applied to Web Usage Mining. The comparison includes (i) *PSP+*, an evolution of GSP, based on candidate generation and test heuristics, (ii) *FreeSpan* [45], based on the integration of frequent sequence mining and frequent pattern mining, and the newly proposed (iii) *PrefixSpan* [46] that uses an approach based on data projection. The results of the comparison [44] show that PrefixSpan outperforms the other two algorithms and offers very good performance even on long sequences. Ref. [47] proposes an hybrid method: data are stored in a database according to a so-called *Click Fact Schema*; an Hypertext Probabilistic Grammar (HPG) is generated by querying the databases; HPGs represent transitions among Web pages through a model which shares many similarities with Markov chains. The frequent sequential patterns are mined through a breadth first search over the hypertext probabilistic grammar. HPGs were first proposed in [48], and later improved in [47] where some scalability issues of the original proposal have been solved.

Clustering techniques look for groups of similar items among large amount of data based on a general idea of *distance function* which computes the similarity between groups. Clustering has been widely used in Web Usage Mining to group together similar sessions [19,49,38,32]. Ref. [50] was the first to suggest that the focus of Web Usage Mining should be shifted from single user sessions to group of user sessions; Ref. [50] was also the first to apply clustering for identifying such cluster of similar sessions. Ref. [32] proposes similarity graph in conjunction with the time spent on Web pages to estimate group similarity in concept-based clustering. Ref. [51] uses *sequence alignment* to measure similarity, while [50] exploits belief functions. Ref. [52] uses Genetic

Algorithms [53] to improve the results of clustering through user feedback. Ref. [54] couples Fuzzy Artificial Immune System and clustering techniques to improve the users' profiles obtained through clustering. In [55], clustering is based both on user navigation patterns, on additional information about the user, and on a mixture of hidden Markov models. Ref. [49] applies multi-modal clustering, a technique which build clusters by using multiple information data features. Ref. [56] presents an application of matrix clustering to Web usage data. Ref. [57] combines association rule mining and clustering into a method called *association rule hypergraph partitioning*. First, association rules are used to extract frequent patterns from user sessions; then the frequent patterns are used to build a graph in which: (i) nodes are the visited Web pages, (ii) edges connect two or more nodes if there is a frequent pattern which contains the pages represented by the nodes; (iii) edges are weighted depending on the relevance of patterns connecting the nodes. Note that this defines an hypergraph since an edge can connect more than two nodes. The hypergraph is recursively partitioned in clusters to identify interesting group of users' behaviors.

5. Applications

The general goal of Web Usage Mining is to gather interesting information about users navigation patterns (i.e., to characterize Web users). This information can be exploited later to improve the Web site from the users' viewpoint. The results produced by the mining of Web logs can used for various purposes [7]: (i) to personalize the delivery of Web content; (ii) to improve user navigation through pre-fetching and caching; (iii) to improve Web design; or in e-commerce sites (iv) to improve the customer satisfaction.

Personalization of web content. Web Usage Mining techniques can be used to provide personalized Web user experience. For instance, it is possible to anticipate the user behavior in real time by comparing the current navigation pattern with typical patterns which were extracted from past Web log. In this area, recommendation systems are the most common application; their aim is to recommend interesting links to products which could be interesting to users [58,42,59,60]. Personalized Site Maps [61] are an example of recommendation system for links (see also [43]). Ref. [62] proposed an adaptive technique to reorganize the product catalog according to the forecasted user profile. An approach to integrate domain ontologies into the personalization process based on Web Usage Mining is proposed in [63], including an algorithm to construct domain-level aggregate profiles from a collection of semantic objects extracted from user transactions. A survey on existing commercial recommendation systems, implemented in e-commerce Web sites, is presented in [64].

Pre-fetching and caching. The results produced by Web Usage Mining can be exploited to improve the performance of Web servers and Web-based applications. Typically, Web Usage Mining can be used to develop proper pre-fetching and caching strategies so as to reduce the server response time, as done in [65–67,40,68].

Support to the design. Usability is one of the major issue in the design and implementation of Web sites. The results produced by Web Usage Mining techniques can provide guidelines for improving the design of Web applications. Ref. [69] uses stratograms to evaluate the organization and the efficiency of Web sites from the users' viewpoint. Ref. [70] exploits Web Usage Mining techniques to suggest proper modifications to Web sites. Adaptive Web sites represents a further

step. In this case, the content and the structure of the Web site can be dynamically reorganized according to the data mined from the users' behavior [71,72].

E-commerce. Mining business intelligence from Web usage data is dramatically important for e-commerce Web-based companies. Customer Relationship Management (CRM) can have an effective advantage from the use of Web Usage Mining techniques. In this case, the focus is on business specific issues such as: customer attraction, customer retention, cross sales, and customer departure [73,18,74].

6. Software

There are many commercial tools which perform analysis on log data collected from Web servers. Most of these tools are based on statistical analysis techniques, while only a few products actually exploit Data Mining techniques. With respect to Web Mining commercial tools, it is worth noting that since the review made in [7], the number of existing products almost doubled. Companies which sold Web Usage Mining products in the past, have disappeared (e.g., Andromeda's Aria); others have been bought by other companies. In most cases, Web Usage Mining tools have become part of integrated Customer Relation Management (CRM) solutions for e-commerce (e.g., [75] and [76]). In some cases, Web Usage Mining tools are simple Web log analyzers (e.g., [77–79]). One software developed in a research environment, WUM [80], appears to be at an interesting maturity level; WUM has currently reached the version 7.0.

Accrue [76] offers a complete suite of software packages dedicated to Web analytics. The most advanced, *Accrue G2* [81], offers a hybrid OLAP technology, based on IBM DB2, that allows advanced information extraction and integration from different sources (e.g. CRM data, Web server logs, transaction logs, etc.). *Accrue G2* appears to be well suited for large companies since it is scalable and can be widely personalized. *Accrue Insight 5* [82] is oriented towards Web analytics for e-business and offers solutions to maximize the effectiveness of content, e-commerce, advertising, and affiliate programs. *Pilot HitList* [83], acquired by Accrue through the acquisition of *PilotSoftware*, offers an efficient Web analytics software for medium size companies. All these products provide data mining tools for user profiling. Lumio's *Recognition* [84] is a complete e-business solution for Web analytics. *Re:cognition* provides different approaches to track users behavior: (i) cookies and Web logs, from the Web server side, (ii) Javascript and Java Applets, from the client side, and even more advanced techniques including (iii) packet sniffer, proxy server logging, etc. *NetTracker* [78], by Sane Solutions, is a family of tools for Web analytics. The advanced editions are suited for e-business analysis and allow integrations with CRM solutions. Sane also offers a service for remote Web analysis based on NetTracker solution. *Elytics Analysis Suite* [85] integrates Web log data with data from the client side and combines them with different user metrics. *E.piphany E.6* [86] is a complete CRM solution that includes tools for Web log and transaction log analysis. NetIQ's *WebTrends Log Analyzer Series* [77] provides Web traffic reporting for the small business. Thanks to the join with NetGenesis, SPSS offers a complete Web analytic solution that integrates NetGenesis Web analytics [87] and SPSS Clementine in *SPSS Predictive Web Analytics* [88]. *WebSideStory* offers *HitBox* [75], a suite of products that includes different solutions for tracking users behavior, ranging from simple statistics to very complete solutions for big companies including e-business oriented analysis. Ad-hoc Javascript

code. IBM provides a family of on-line services called *Surfaid Analytics* [28] that allow the filtering and integration of Web log, commerce server, registration and other data into a database of profiles. These profiles are subsequently used for producing standardized reports as well as performing OLAP operations. Data can be collected both from Web logs and using Javascript based on techniques developed by IBM. Quest's *Funnel Web Analyzer* [79] is available both with a free license and with a commercial license. The former version offers basic statistical reports, the latter one offers advanced features based on data mining techniques. *WebHound* [89] is the Web analytic tool developed by SAS. It extracts information from Web logs and performs click-stream analysis. SAS offers also an on-line Web analytic service, SAS IntelliVisor [90], developed for e-business oriented analysis of Web sites. Different versions of the service for three different market areas are available: Financial Services, Pharmaceuticals, and Retail. *Megaputer WebAnalyst* [91] is a complete application server that integrates with existing Web servers to provide advanced Web usage analysis. The application server acts as a sort of proxy between the clients and the actual Web server. The application server instead of the Web server intercepts the page requests and after a processing step it forwards the page requests to the actual Web server. The page forwarded from the server is again intercepted by the application server before being sent to the client who issued the page request. Prudsys' *ECOMMINER* [92] is oriented toward *e-business* and integrates data from Web logs with server side transactional data. It has been developed for two specific *e-commerce* platforms: *InterShop* [93] and *Logisma Business Webstore* [94].

7. From techniques to applications

When facing a Web Usage Mining task, the main issue is how to determine which technique is most suited for the current problem setting. As in most of the application domains, also in Web Usage Mining there is not a unique answer for such a question. In the previous two sections we have overviewed the most used techniques and the main application settings in Web Usage Mining research. Generally any technique can be suitable in any of the four applications area we discussed: (i) Personalization of Web Content; (ii) Pre-fetching and Caching; (iii) Support to Desing; (iv) E-commerce. For instance, in [95] a divisive hierarchical clustering technique is used to group Web site users according to their interests and then to personalize the Web Content according to the group to which the user belongs. In [21] both sequential patterns and clustering techniques are applied to Personalization of Web Content. Clustering is used to group together similar user session, while Sequential patterns by means of Markov chains are used to predict the users behavior. Ref. [96] uses association rules for recommender systems. In [40] association rules are used to address the problem of Caching and Pre-fetching. To address the same problem [67] uses a classifier based on sequential patterns. Ref. [43] uses sequential patterns to create dynamic adaptive Web sites. In [70] classifiers are used to classify Web pages according to users' navigation and then classified pages are used to reorganize the Web site structure. Ref. [73] presents how different data mining techniques—namely association rules, sequential patterns, classification and clustering—can be used for customer segmentation and profiling in e-commerce applications. In Table 1 we summarize the applications and the techniques used in the reviewed researches so as to provide a birdseye view of the field and a short guide about how the techniques are coupled to the

Table 1
Web Usage Mining: a reference between applications, techniques, and data sources

Paper	Application	Technique	Data source
[21]	Personalization	Clustering, Association rules, Relational Markov Models	Web Server
[63]	Personalization	Clustering	Web Server
[15]	Personalization		Client
[71]	Personalization	Fuzzy Clustering	Web Server
[95]	Personalization	Clustering	Web Server
[96]	Personalization	Association rules	Web Server
[57]	Personalization	Association rules, Clustering	Web Server
[50]	Personalization	Clustering	Web Server
[97]	Personalization	Clustering, Sequential patterns	Web Server
[98]	Caching	Classifiers, Association rules, Sequential patterns	Proxy Server
[66]	Caching	Association rules	Web Server
[67]	Caching	Association rules, Classifiers	Web Server
[99]	Caching	Association rules	Proxy Server
[68]	Caching	Markov Models	Web Server
[100]	Caching	Association rules	Web Server
[101]	Design	Markov Models	Web Server
[69]	Design	Classifiers, Sequential patterns	Web Server
[70]	Design	Classifiers	Web Server
[102]	Design	Sequential patterns	Web Server
[103]	Design	Sequential patterns	Web Server
[104]	Design	Markov Models	Web Server
[18]	E-commerce		Web Server
[73]	E-commerce	Classifiers, Association rules, Sequential patterns	Web Server
[105]	E-commerce	Clustering	Web Server
[51]	E-commerce	Fuzzy Logic, Clustering, Genetic Algorithms	Web Server

application domains. It can be noted that there is no strict correlation between techniques and application domain. Most of the approaches are used in most of the domains.

8. Privacy issues

Web Usage Mining tools integrate different data sources (Web logs, cookies data, as well as personal data) to accurately track users behavior. This raises the issue of users privacy, a topic that is currently highly relevant for the whole data mining area. The European Union, the United States, and other countries are publishing very strict laws about privacy [106]. In the Web Mining context, [7] was the first to discuss users privacy as a *relevant* and *sensitive* issue. But in general, this topic is rarely discussed in research papers proposing advanced Web Usage Mining techniques. In [64,74] privacy concerns linked to Web Personalization are discussed. One of the main proposal to deal with privacy issues in the Web area is the Platform for Privacy Preferences or P3P [107]. The purpose of the P3P standard is to enable Web sites to express their privacy practices in a standardized format that can be automatically retrieved and interpreted by user agents. Note however, that P3P does not solve the privacy issue completely since it does not provide any mechanism to ensure that visited Web sites will actually act according to their declared policies. In addition, P3P

does not address the issue of using data mining techniques over users data. User profiling is very important for many e-business applications, being probably the most relevant area in Web Usage Mining. To solve the related privacy issues, researchers have also tackled the problem as the possibility to develop effective user models *without* accessing precise information available in individual data records so as to not violating users' privacy. This approach was first presented in [108,109] with a specific focus on decision tree techniques; later [110,111] extended the focus to Association rules. However, no papers have yet proposed approaches or solutions that would take into account the privacy from Web Usage Mining viewpoint.

9. Future trends

There are a number of open issues in Web Usage Mining area. In many practical applications, due to the introduction of stricter laws, privacy respect represents a big challenge. Anyway we believe that the most interesting research area deals with integration of semantics within Web site design so to improve the results of Web Usage Mining applications. Indeed, with the growing interest in the notion of Semantic Web, an increasing number of sites use structured semantics and domain ontologies as part of the site design, creation, and content delivery. The primary challenge for the next-generation of personalization systems will regard the integration of semantic knowledge from domain ontologies into the various parts of the process, including the data preparation, pattern discovery, and recommendation phases. Efforts in this direction are likely to be the most fruitful in the creation of much more effective Web Usage Mining and personalization systems that are consistent with emergence and proliferation of the Semantic Web. We believe that the approach presented in [112] is one of the most promising in the ongoing researches. The authors present a framework that integrates the design and the development of Web applications with the analysis of Web usage. The approach proposed is based on the adoption of the Web Modeling Language (WebML), and its supporting CASE tool WebRatio, for the design and the development of Web applications. The use of an application conceptual schema expressed in WebML, and the integration of WebRatio with advanced logging techniques allow the development of Web applications that produce rich Web logs containing information not commonly available with other approaches.

10. Summary

We presented a survey of the recent developments in the area of Web Usage Mining. The survey is based on the more than 150 papers published since 2000 on this topic. As it was not possible to cite all the papers here, we refer the interested reader to the Web Mining on-line bibliography hosted on the cInQ project Web site [10].

Acknowledgements

This work has been supported by the *consortium on discovering knowledge with Inductive Queries* (*cInQ*) [10], a project funded by the Future and Emerging Technologies arm of the IST Pro-

gramme (Contract no. IST-2000-26469). The authors wish to thank Stefano Ceri for the inspiration and discussions that made this work possible, Maristella Matera for several helpful discussions, and Mario Verdicchio for several helpful comments.

References

[1] O. Etzioni, The world-wide Web: quagmire or gold mine? Communications of the ACM 39 (11) (1996) 65–68.
[2] R. Kosala, H. Blockeel, Web mining research: a survey, SIGKDD: SIGKDD explorations: newsletter of the special interest group (SIG) on knowledge discovery & data mining, ACM 2 (1) (2000) 1–15.
[3] S. Brin, L. Page, The anatomy of a large-scale hypertextual Web search engine, Computer Networks and ISDN Systems 30 (1–7) (1998) 107–117.
[4] Configuration file of W3C httpd, http://www.w3.org/Daemon/User/Config/ (1995).
[5] W3C Extended Log File Format, http://www.w3.org/TR/WD-logfile.html (1996).
[6] J.R. Punin, M.S. Krishnamoorthy, M.J. Zaki, Logml: Log markup language for web usage mining, in: R. Kohavi, B. Masand, M. Spiliopoulou, J. Srivastava (Eds.), WEBKDD 2001—Mining Web Log Data Across All Customers Touch Points, Third International Workshop, San Francisco, CA, USA, August 26, 2001. Revised Papers, vol. 2356 of Lecture Notes in Computer Science, Springer, 2002, pp. 88–112.
[7] J. Srivastava, R. Cooley, M. Deshpande, P.-N. Tan, Web usage mining: discovery and applications of usage patterns from web data, SIGKDD Explorations 1 (2) (2000) 12–23.
[8] The WebSIFT project, http://www.cs.umn.edu/research/websift/ (2003).
[9] S. Pal, V. Talwar, P. Mitra, Web Mining in soft computing framework: relevance, state of the art and future directions, IEEE Transactions on Neural Networks 13 (5) (2002) 1163–1177.
[10] Consortium on discovering knowledge with Inductive Queries (clnQ). Project funded by the European Commission under the Information Society Technologies Programme (1998–2002) Future and Emerging Technologies arm. Contract no. IST-2000-26469., http://www.cinq-project.org. Bibliography on Web Usage Mining available at http://www.cinq-project.org/intranet/polimi/.
[11] A. Nanopoulos, M. Zakrzewicz, T. Morzy, Y. Manolopoulos, Indexing web access-logs for pattern queries, in: fourth ACM CIKM International Workshop on Web Information and Data Management (WIDM'02), 2002.
[12] K.P. Joshi, A. Joshi, Y. Yesha, On using a warehouse to analyze web logs, Distributed and Parallel Databases 13 (2) (2003) 161–180.
[13] D.M. Kristol, Http cookies: standards, privacy, and politics, ACM Transactions on Internet Technology (TOIT) 1 (2) (2001) 151–198.
[14] B. Berendt, B. Mobasher, M. Nakagawa, M. Spiliopoulou, The impact of site structure and user environment on session reconstruction in web usage analysis, in: Proceedings of the 4th WebKDD 2002 Workshop, at the ACM-SIGKDD Conference on Knowledge Discovery in Databases (KDD'2002), 2002.
[15] K.D. Fenstermacher, M. Ginsburg, Mining client-side activity for personalization, in: Fourth IEEE International Workshop on Advanced Issues of E-Commerce and Web-Based Information Systems (WECWIS'02), 2002, pp. 205–212.
[16] Pilot Software, Web site analysis, Going Beyond Traffic Analysis http://www.marketwave.com/products-solutions/hitlist.html (2002).
[17] S. Ansari, R. Kohavi, L. Mason, Z. Zheng, Integrating e-commerce and data mining: Architecture and challenges, in: WEBKDD 2000—Web Mining for E-Commerce—Challenges and Opportunities, Second International Workshop, 2000.
[18] S. Ansari, R. Kohavi, L. Mason, Z. Zheng, Integrating e-commerce and data mining: Architecture and challenges, in: N. Cercone, T.Y. Lin, X. Wu (Eds.), Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM 2001), IEEE Computer Society, 2001.
[19] C. Shahabi, F. Banaei-Kashani, A framework for efficient and anonymous web usage mining based on client-side tracking, in: R. Kohavi, B. Masand, M. Spiliopoulou, J. Srivastava (Eds.), WEBKDD 2001—Mining Web Log Data Across All Customers Touch Points, Third International Workshop, San Francisco, CA, USA, August 26, 2001. Revised papers, vol. 2356 of Lecture Notes in Computer Science, Springer, 2002, pp. 113–144.

[20] L.D. Catledge, J.E. Pitkow, Characterizing browsing strategies in the World-Wide Web, Computer Networks and ISDN Systems 27 (6) (1995) 1065–1073.
[21] C.R. Anderson, A machine learning approach to web personalization, Ph.D. thesis, University of Washington, 2002.
[22] R. Cooley, B. Mobasher, J. Srivastava, Data preparation for mining world wide web browsing patterns, Knowledge and Information Systems 1 (1) (1999) 5–32.
[23] B. Diebold, M. Kaufmann, Usage-based visualization of web localities, in: Australian symposium on information visualisation, 2001, pp. 159–164.
[24] P.-N. Tan, V. Kumar, Modeling of web robot navigational patterns, in: WEBKDD 2000—Web Mining for E-Commerce—Challenges and Opportunities, Second International Workshop, 2000.
[25] P.-N. Tan, V. Kumar, Discovery of web robot sessions based on their navigational patterns, Data Mining and Knowledge Discovery 6 (1) (2002) 9–35.
[26] R. Cooley, Web usage mining: discovery and application of interesting patterns from web data, Ph.D. thesis, University of Minnesota, 2000.
[27] B. Mobasher, R. Cooley, J. Srivastava, Automatic personalization based on web usage mining, Communications of the ACM 43 (8) (2000) 142–151.
[28] IBM, SurfAid Analytics http://surfaid.dfw.ibm.com (2003).
[29] M. Chen, A.S. LaPaugh, J.P. Singh, Predicting category accesses for a user in a structured information space, in: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval, 2002, pp. 65–72.
[30] G. Stumme, A. Hotho, B. Berendt, Usage mining for and on the semantic web, in: National Science Foundation Workshop on Next Generation Data Mining, 2002.
[31] M. Craven, D. DiPasquo, D. Freitag, A. McCallum, T. Mitchell, K. Nigam, S. Slattery, Learning to construct knowledge bases from the world wide web, Artificial Intelligence 118 (1–2) (2000) 69–113.
[32] A. Banerjee, J. Ghosh, Clickstream clustering using weighted longest common subsequences, in: Proceedings of the Web Mining Workshop at the 1st SIAM Conference on Data Mining, 2001.
[33] E.H. Chi, P. Pirolli, K. Chen, J.E. Pitkow, Using information scent to model user information needs and actions and the web, in: Proceedings of ACM CHI 2002 Conference on Human Factors in Computing Systems, ACM Press, 2001, pp. 490–497.
[34] R. Cooley, The use of web structure and content to identify subjectively interesting web usage patterns, ACM Transactions on Internet Technology (TOIT) 3 (2) (2003) 93–116.
[35] J. Andersen, A. Giversen, A.H. Jensen, R.S. Larsen, T.B. Pedersen, J. Skyt, Analyzing clickstreams using subsessions, in: International Workshop on Data Warehousing and OLAP (DOLAP 2000), 2000.
[36] J. Pei, J. Han, B. Mortazavi-asl, H. Zhu, Mining access patterns efficiently from web logs, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2000, pp. 396–407.
[37] E. Menasalvas, S. Millan, J. Pena, M. Hadjimichael, O. Marban, Subsessions: a granular approach to click path analysis, in: Proceedings of FUZZ-IEEE Fuzzy Sets and Systems Conference, at the World Congress on Computational Intelligence, Honolulu, HI, 12–17 May 2002.
[38] J.Z. Huang, M. Ng, W.-K. Ching, J. Ng, D. Cheung, A cube model and cluster analysis for web access sessions, in: R. Kohavi, B. Masand, M. Spiliopoulou, J. Srivastava (Eds.), WEBKDD 2001—Mining Web Log Data Across All Customers Touch Points, Third International Workshop, San Francisco, CA, USA, August 26, 2001. Revised papers, vol. 2356 of Lecture Notes in Computer Science, Springer, 2002, pp. 48–67.
[39] J. Han, M. Kamber, Data Mining Concepts and Techniques, Morgan Kaufmann, 2001.
[40] A. Nanopoulos, D. Katsaros, Y. Manolopoulos, Exploiting web log mining for web cache enhancement, in: R. Kohavi, B. Masand, M. Spiliopoulou, J. Srivastava (Eds.), WEBKDD 2001—Mining Web Log Data Across All Customers Touch Points, Third International Workshop, San Francisco, CA, USA, August 26, 2001. Revised papers, vol. 2356 of Lecture Notes in Computer Science, Springer, 2002, pp. 68–87.
[41] X. Huang, N. Cercone, A. An, Comparison of interestingness functions for learning web usage patterns, in: Proceedings of the Eleventh International Conference on Information and Knowledge Management, ACM Press, 2002, pp. 617–620.
[42] S.S.C. Wong, S. Pal, Mining fuzzy association rules for web access case adaptation, in: Workshop on Soft Computing in Case-Based Reasoning, International Conference on Case-Based Reasoning (ICCBR'Ol), 2001.

[43] E.S. Nan Niu, M. El-Ramly, Understanding web usage for dynamic web-site adaptation: A case study, in: Proceedings of the Fourth International Workshop on Web Site Evolution (WSE'02), IEEE, 2002, pp. 53–64.
[44] B. Mortazavi-Asl, Discovering and mining user web-page traversal patterns, Master's thesis, Simon Fraser University, 2001.
[45] J. Han, J. Pei, B. Mortazavi-Asl, Q. Chen, U. Dayal, M. Hsu, FreeSpan: frequent pattern-projected sequential pattern mining, in: Proceedings of the Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD'2000), Boston, MA, 2000.
[46] J. Pei, J. Han, B. Mortazavi-Asl, H. Pinto, Q. Chen, U. Dayal, M.-C. Hsu, Mining sequential patterns by pattern-growth: the PrefixSpan Approach, IEEE Transactions on Knowledge and Data Engineering, in press.
[47] S.E. Jespersen, J. Thorhauge, T.B. Pedersen, A hybrid approach to web usage mining, in: Proceedings of the 4th International Conference on Data Warehousing and Knowledge Discovery, Springer-Verlag, 2002, pp. 73–82.
[48] J. Borges, A data mining model to capture UserWeb navigation patterns, Ph.D. thesis, Department of Computer Science, University College London, 2000.
[49] J. Heer, E.H. Chi, Mining the structure of user activity using cluster stability, in: Proceedings of the Workshop on Web Analytics, Second SIAM Conference on Data Mining, ACM Press, 2002.
[50] Y. Xie, V.V. Phoha, Web user clustering from access log using belief function, in: Proceedings of the First International Conference on Knowledge Capture (K-CAP 2001), ACM Press, 2001, pp. 202–208.
[51] B. Hay, G. Wets, K. Vanhoof, Clustering navigation patterns on a website using a sequence alignment method. In: Intelligent Techniques for Web Personalization: IJCAI 2001, 17th Int. Joint Conf. on Artificial Intelligence, August 4, 2001, Seattle, WA, USA, pp. 1–6.
[52] C. Shahabi, Y.-S. Chen, Improving user profiles for e-commerce by genetic algorithms, E-Commerce and Intelligent Methods Studies in Fuzziness and Soft Computing 105 (8) (2002).
[53] J.H. Holland, Adaptation in Natural and Artificial Systems, University of Michigan Press, Ann Arbor, 1975, republished by the MIT press, 1992.
[54] O. Nasraoui, F. Gonzalez, D. Dasgupta, The fuzzy artificial immune system: Motivations, basic concepts, and application to clustering and web profiling, in: Proceedings of the World Congress on Computational Intelligence (WCCI) and IEEE International Conference on Fuzzy Systems, 2002, pp. 711–716.
[55] A. Ypma, T. Heskes, Clustering web surfers with mixtures of hidden markov models, in: Proceedings of the 14th Belgian–Dutch Conference on AI (BNAIC'02), 2002.
[56] S. Oyanagi, K. Kubota, A. Nakase, Application of matrix clustering to web log analysis and access prediction, in: WEBKDD 2001—Mining Web Log Data Across All Customers Touch Points, Third International Workshop, 2001.
[57] B. Mobasher, H. Dai, M. Tao, Discovery and evaluation of aggregate usage profiles for web personalization, Data Mining and Knowledge Discovery 6 (2002) 61–82.
[58] G. Adomavicius, A. Tuzhilin, Extending recommender systems: A multidimensional approach. Workshop on Intelligent Techniques for Web Personalization, IJCAI 2001, Seattle, WA, USA.
[59] D. VanderMeer, K. Dutta, A. Datta, Enabling scalable online personalization on the web, in: Proceedings of the 2nd ACM E-Commerce Conference (EC'00), ACM Press, 2000, pp. 185–196.
[60] B. Mobasher, H. Dai, T. Luo, M. Nakagawa, Effective personalization based on association rule discovery from web usage data, Web Information and Data Management (2001) 9–15.
[61] F. Toolan, N. Kushmerick, Mining web logs for personalized site maps. In: Third Int. Conf. on Web Information Systems Engineering (WISE 02), Workshop on Mining for Enhanced Web Search. Singapore, December 11, 2002, pp. 232–237.
[62] H.-Y. Paik, B. Benatallah, R. Hamadi, Dynamic restructuring of e-catalog communities based on user interaction patterns, World Wide Web 5 (4) (2002) 325–366.
[63] H.K. Dai, B. Mobasher, Using ontologies to discover domain-level web usage profiles, in: Proceedings of the 2nd Semantic Web Mining Workshop at ECML/PKDD 2002, Helsinki, Finland, August 2002.
[64] J.B. Schafer, J.A. Konstan, J. Riedl, E-commerce recommendation applications, Data Mining and Knowledge Discovery 5 (1–2) (2001) 115–153.
[65] C.-Y. Chang, M.-S. Chen, A new cache replacement algorithm for the integration of web caching and prefetching, in: Proceedings of the Eleventh International Conference on Information and Knowledge Management, ACM Press, 2002, pp. 632–634.

[66] B. Lan, S. Bressan, B.C. Ooi, K.-L. Tan, Rule-assisted prefetching in web-server caching, in: Proceedings of the Ninth International Conference on Information and Knowledge Management (CIKM 2000), ACM Press, 2000, pp. 504–511.
[67] T. Li, Web-document prediction and presenting using association rule sequential classifiers, Master's thesis, Simon Fraser University, 2001.
[68] Y.-H. Wu, A.L.P. Chen, Prediction of web page accesses by proxy server log, World Wide Web 5 (1) (2002) 67–88.
[69] B. Berendt, Using site semantics to analyze, visualize, and support navigation, Data Mining and Knowledge Discovery 6 (1) (2002) 37–59.
[70] Y. Fu, M. Creado, C. Ju, Reorganizing web sites based on user access patterns, in: Proceedings of the Tenth International Conference on Information and Knowledge Management, ACM Press, 2001, pp. 583–585.
[71] T. Kamdar, Creating adaptive web servers using incremental web log mining, Master's thesis, Computer Science Department, University of Maryland, Baltimore County, 2001.
[72] O.R. Zaïane, Web usage mining for a better web-based learning environment, in: Proceedings of Conference on Advanced Technology for Education, 2001, pp. 450–455.
[73] C. Bounsaythip, E. Rinta-Runsala, Overview of data mining for customer behavior modeling, Technical Report TTE1-2001-18, VTT Information Technology (2001).
[74] M. Eirinaki, M. Vazirgiannis, Web Mining for web personalization, ACM Transactions on Internet Technology (TOIT) 3 (1) (2003) 1–27.
[75] WebSideStory HitBox, http://www.websidestory.com (2003).
[76] Accrue, http://www.accrue.com (2003).
[77] NetIQ WebTrends Log Analyzer, http://www.netiq.com (2003).
[78] Sane NetTracker, http://www.sane.com/products/NetTracker (2003).
[79] Funnel Web Analyzer, http://www.quest.com (2003).
[80] WUM: A Web Utilization Miner, http://wum.wiwi.huberlin.de (2003).
[81] Accrue G2, http://www.accrue.com/products/g2 (2003).
[82] Accrue Insight 5, http://www.accrue.com/products/insight (2003).
[83] Pilot Hitlist, http://www.accrue.com/products/hitlist (2003).
[84] Lumio Recognition, http://www.lumio.com (2003).
[85] Elytics Analysis Suite, http://www.elytics.com (2003).
[86] E.piphany E.6, http://www.epiphany.com (2003).
[87] NetGenesis, http://www.spss.com/netgenesis (2003).
[88] SPSS, http://www.spss.com (2003).
[89] SAS WebHound, http://www.sas.com/products/webhound (2003).
[90] SAS IntelliVisor, http://www.sas.com/solutions/intellivisor (2003).
[91] Megaputer WebAnalyst, http://www.megaputer.com (2003).
[92] Prudsys ECOMMINER, http://www.prudsys.com (2003).
[93] InterShop, http://www.intershop.com (2003).
[94] Logisma Business Webstore, http://www.logisma.de (2003).
[95] H.R. Kim, P.K. Chan, Learning implicit user interest hierarchy for context in personalization, in: Proceedings of the 2003 International Conference on Intelligent User Interfaces, ACM Press, 2003, pp. 101–108.
[96] W. Lin, S.A. Alvarez, C. Ruiz, Efficient adaptive–support association rule mining for recommender systems, Data Mining and Knowledge Discovery 6 (1) (2002) 83–105.
[97] M. Eirinaki, M. Vazirgiannis, I. Varlamis, Sewep: using site semantics and a taxonomy to enhance the web personalization process, in: Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM Press, 2003, pp. 99–108.
[98] F. Bonchi, F. Giannotti, C. Gozzi, G. Manco, M. Nanni, D. Pedreschi, C. Renso, S. Ruggieri, Web log data warehousing and mining for intelligent web caching, Data Knowledge Engineering 39 (2) (2001) 165–189.
[99] A. Nanopoulos, D. Katsaros, Y. Manolopoulos, A data mining algorithm for generalized web prefetching, IEEE Transactions on Knowledge and Data Engineering 15 (5) (2003) 1155–1169.
[100] Q. Yang, H.H. Zhang, Web-log mining for predictive web caching, IEEE Transactions on Knowledge and Data Engineering 15 (4) (2003) 1050–1054.

[101] C.R. Anderson, P. Domingos, D.S. Weld, Relational markov models and their application to adaptive web navigation, in: Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-2002), 2002.
[102] M. Spiliopoulou, C. Pohle, Data mining for measuring and improving the success of web sites (1–2) (2001) 85–114.
[103] R. Srikant, Y. Yang, Mining web logs to improve website organization, World Wide Web (2001) 430–437.
[104] J. Zhu, J. Hong, J.G. Hughes, Using markov chains for link prediction in adaptive web sites, in: D.W. Bustard, W. Liu, R. Sterritt (Eds.), Soft-Ware 2002: Computing in an Imperfect World, First International Conference, Soft-Ware 2002, Belfast, Northern Ireland, 8–10 April 2002, Proceedings, vol. 2311 of Lecture Notes in Computer Science, Springer, 2002, pp. 60–73.
[105] W.-L. Chang, S.-T. Yuan, A synthesized learning approach for web-based crm, in: WEBKDD 2000—Web Mining for E-Commerce—Challenges and Opportunities, Second International Workshop, 2000.
[106] Directive 94/46/ec of the european parliament and of the council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data, Official Journal of the European Community (281) (1995) 31–50, http://europa.eu.int/comm/internalmarket/privacy/index.htm.
[107] Platform for Privacy Preferences (P3P) Project, http://www.w3.org/TR/P3P/ (2003).
[108] R. Agrawal, R. Srikant, Privacy-preserving data mining, in: Proceedings of the ACM SIGMOD Conference on Management of Data, ACM Press, 2000, pp. 439–450.
[109] Y. Lindell, B. Pinkas, Privacy preserving data mining, Lecture Notes in Computer Science 1880 (2000) 36.
[110] A. Evfimievski, R. Srikant, R. Agrawal, J. Gehrke, Privacy preserving mining of association rules, in: Proceedings of 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2002.
[111] S.J. Rizvi, J.R. Haritsa, Maintaining data privacy in association rule mining, in: Proceedings of 28th International Conference on Very Large Data Bases (VLDB), 2002.
[112] R. Meo, P.L. Lanzi, M. Matera, Integrating web conceptual modeling and web usage mining, Technical Report, Dipartimento di Elettronica e Informazione—Politecnico di Milano, accepted at WEBKDD2004 (2004).

Federico Michele Facca obtained his Master degree in Computer Engineering from the Politecnico di Milano in 2004. His master thesis regards XML mining techniques with a specific focus on the application to the mining of Web logs.

Pier Luca Lanzi received the Laurea degree in computer science in 1994 from the Università degli Studi di Udine and the Ph.D. degree in Computer and Automation Engineering from the Politecnico di Milano in 1999. Since 2001 he is a research professor at the Politecnico di Milano, Department of Electronics and Information. His research areas include evolutionary computation, reinforcement learning, and machine learning. He is interested in applications to data mining and autonomous agents. He is member of the editorial board of the *Evolutionary Computation Journal*.