Recent Developments in Web Usage Mining Research

Federico Michele Facca and Pier Luca Lanzi*

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

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.

1 Introduction

Web Mining [29] is that area of Data Mining which deals with the extraction of interesting knowledge from the World Wide Web. More precisely [40], *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 (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 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. [20]), 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; 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 than 150 papers, of the overall 400, have been before 2001; around the 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 of interest for this area. This paper is a survey of the recent developments in the area of Web Usage Mining. It is based on the more than 150 papers published since 2000 on the topic of Web Usage Mining; see the on-line bibliography on the web site of the cInQ project [1].

* Contact Author: Pier Luca Lanzi, `pierluca.lanzi@polimi.it`.

Y. Kambayashi, M. Mohania, W. Wöß (Eds.): DaWaK 2003, LNCS 2737, pp. 140–150, 2003.

2 Data Sources

Web Usage Mining applications are based on data collected from three main sources [58]: (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 [2], Extended Log Format [3], LogML [53]. When exploiting log information from web servers, the major issue is the identification of users' sessions (see Section 3).

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 [52]. 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. Packet sniffers are rarely used in practice because of rise scalability issue on web servers with high traffic [52], and the impossibility to access encrypted packets like those used in secure commercial transactions a quite severe limitation when applying web usage mining to e-businesses [13]. Probably, the best approach for tracking web usage consists of directly accessing the server application layer, as proposed in [14]. Unfortunately, this is not always possible.

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 collect data of *groups of users* accessing *huge groups* of web servers.

The Client Side. Usage data can be tracked also on the client side by using JavaScript, java applets [56], or even modified browsers [22]. 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 [30]. However, these approaches rely heavily on the users' cooperation and rise many issues concerning the privacy laws, which are quite strict.

3 Preprocessing

Data preprocessing has a fundamental role in Web Usage Mining applications.

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 [27, 12] 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, a heuristic based on navigational behavior can be used to separates robot sessions from actual users' sessions (see [60, 61]).

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. 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 identifies. This problem can be partially solved by the use of cookies [25], by URL rewriting, or by requiring the user to log in when entering the web site [12]. Web browser caching is a more complex issue. Log 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. 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*. [17] 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. [24] proposed an adaptive time out heuristic. [26] 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.

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. [26] has been the first to employ content based information to enrich the web log data. 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, [59] proposes to use Semantic Web for Web Usage Mining: web pages are mapped into ontologies to add meaning to the observed frequent paths. [15] 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.

Data Formatting. This is the final step of preprocessing. Once the previous phases have been completed, data are properly formatted before applying mining techniques. [11] 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. [47] introduces a method based on signature tree to index log stored in databases for efficient pattern queries. A tree structure, WAP-tree, is also introduced in [51] to register access sequence to web pages; this structure is optimized to exploit the sequence mining algorithm developed by the same authors [51].

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 [32] for a detailed description of these techniques).

Association Rules. are probably the most elementary data mining technique and, at the same time, the most used technique in Web Usage Mining. 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 [38] and [46] by using a modification of the Apriori algorithm [32]. [37] proposes and evaluates some interestingness measures to evaluate the association rules mined from web usage data. [21] 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 form [45]: the 70% of users who *first* visited *A.html* and *then* visited *B.html* afterwards, in the same session, have also accessed page *C.html*. Sequential patterns might appear *syntactically* similar to association rules; in fact algorithms to extract association rules can also be used for sequential pattern mining. There are essentially two classes of algorithms that are used to extract sequential patterns: one includes methods based on association rule mining; one includes methods based on the use of tree-structures, data projection techniques, and Markov chains to mine navigation patterns. Some well-known algorithms for mining association rules have been modified to extract sequential patterns. [44] presents a comparison of different sequential pattern algorithms applied to Web Usage Mining. The comparison includes PSP+, FreeSpan, and PrefixSpan. While PSP+ is an evolution of GSP, based on candidate generation and test heuristic, FreeSpan and the newly proposed PrefixSpan use a data projection based approach. According to [44] PrefixSpan outperforms the other two algorithms and offers very good performance even on long sequences. [54] proposes an hybrid method: data are store 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 resembles 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 [18], and later improved in [54] 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 [56, 34, 36, 15]. [65] was the first to suggest that the focus of web usage mining should be shifted from single user sessions to group of user sessions; [65] was also the first to apply clustering for identifying such cluster of similar sessions. [15] proposes similarity graph in conjunction with the time spent on web pages to estimate group similarity in concept-based clustering. [33] uses *sequence alignment* to measure similarity, while [65] exploits belief functions. [57] uses Genetic Algorithms [35] to improve the results of clustering through user feedback. [48] couples Fuzzy Artificial Immune System and clustering techniques to improve the users' profiles obtained through clustering. [34] applies multimodal clustering, a technique which build clusters by using multiple information data features. [49] presents an application of matrix clustering to web usage data.

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 [58]: (i) to personalize the delivery of web content; (ii) to improve user navigation through prefetching 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, in real time, the user behavior 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 [10, 21, 63, 43]. Personalized Site Maps [62] are an example of recommendation system for links (see also [45]). [50] proposed an adaptive technique to reorganize the product catalog of the products according to the forecasted user profile. A survey on existing commercial recommendation systems, implemented in e-commerce web sites, is presented in [55].

Prefetching 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 prefetching and caching strategies so as to reduce the server response time, as done in [23, 41, 42, 46, 64].

Support to the Design. Usability is one of the major issues 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. [16] uses stratograms to evaluate the organization and the efficiency of web sites from the users' viewpoint. [31] 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 [39, 66].

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 [19, 14, 28].

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 exploit Data Mining techniques. [28] provides an up to date review of available commercial tools for web usage mining. With respect to Web Mining commercial tools, it is worth noting that since the review made in [58], the number of existing products almost doubled. Companies which sold Web Usage Mining products in the past have been disappeared (e.g., Andromeda's Aria); others have been bought by other companies. In most cases, Web Usage Mining tools are part of integrated Customer Relation Management (CRM) solutions for e-commerce (e.g., [8] and [4]). Sometimes, these tools are simple web log analyzers (e.g., [6, 7, 5]). One software developed in a research environment, WUM [9], appears to have reached an interesting maturity level; WUM has currently reached the version 7.0.

We presented a survey of the recent developments in the area of Web Usage Mining, based on the more than 150 papers published since 2000 on this topic. Because, it was not possible to cite all the papers here we refer the interested reader to provide an on-line bibliography on the web site of the cInQ project [1].

Acknowledgements

This work has been supported by the *consortium on discovering knowledge with Inductive Queries* (*cInQ*) [1], a project funded by the Future and Emerging Technologies arm of the IST Programme (Contr. no. IST-2000-26469). The authors wish to thank Maristella Matera for discussions.

References

[1] **c**onsortium on discovering knowledge with **In**ductive **Q**ueries (*cInQ*). 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/`. 140, 146

[2] Configuration File of W3C httpd, 1995. `http://www.w3.org/Daemon/User/Config/`. 141

[3] W3C Extended Log File Format, 1996. `http://www.w3.org/TR/WD-logfile.html`. 141

[4] Accrue, 2003. `http://www.accrue.com`. 146

[5] Funnel Web Analyzer, 2003. `http://www.quest.com`. 146

[6] NetIQ WebTrends Log Analyzer, 2003. `http://www.netiq.com`. 146

[7] Sane NetTracker, 2003. `http://www.sane.com/products/NetTracker`. 146

[8] WebSideStory HitBox, 2003. `http://www.websidestory.com`. 146

[9] WUM: A Web Utilization Miner, 2003. `http://wum.wiwi.hu-berlin.de`. 146

[10] Gediminas Adomavicius and Alexander Tuzhilin. Extending recommender systems: A multidimensional approach. 145
[11] Jesper Andersen, Anders Giversen, Allan H. Jensen, Rune S. Larsen, Torben Bach Pedersen, and Janne Skyt. Analyzing clickstreams using subsessions. In *International Workshop on Data Warehousing and OLAP (DOLAP 2000)*, 2000. 143
[12] Corin R. Anderson. *A Machine Learning Approach to Web Personalization*. PhD thesis, University of Washington, 2002. 142
[13] Suhail Ansari, Ron Kohavi, Llew Mason, and Zijian Zheng. Integrating e-commerce and data mining: Architecture and challenges. In *WEBKDD 2000 - Web Mining for E-Commerce – Challenges and Opportunities, Second International Workshop*, August 2000. 141
[14] Suhail Ansari, Ron Kohavi, Llew Mason, and Zijian Zheng. Integrating e-commerce and data mining: Architecture and challenges. In Nick Cercone, Tsau Young Lin, and Xindong Wu, editors, *Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM 2001)*. IEEE Computer Society, 2001. 141, 145
[15] A. Banerjee and 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. 143, 144
[16] Bettina Berendt. Using site semantics to analyze, visualize, and support navigation. *Data Mining and Knowledge Discovery*, 6(1):37–59, 2002. 145
[17] Bettina Berendt, Bamshad Mobasher, Miki Nakagawa, and Myra 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. 142
[18] Jose Borges. *A Data Mining Model to Capture UserWeb Navigation Patterns*. PhD thesis, Department of Computer Science University College London, 2000. 144
[19] Catherine Bounsaythip and Esa Rinta-Runsala. Overview of data mining for customer behavior modeling. Technical Report TTE1-2001-18, VTT Information Technology, 2001. 145
[20] Sergey Brin and Lawrence Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1–7):107–117, 1998. 140
[21] S. Shiu C. Wong and S. Pal. Mining fuzzy association rules for web access case adaptation. In *Case-Based Reasoning Research and Development: Proceedings of the Fourth International Conference on Case-Based Reasoning*, pages ?–?, 2001. 143, 145
[22] Lara D. Catledge and James E. Pitkow. Characterizing browsing strategies in the World-Wide Web. *Computer Networks and ISDN Systems*, 27(6):1065–1073, 1995. 141
[23] Cheng-Yue Chang and Ming-Syan Chen. A new cache replacement algorithm for the integration of web caching and prefectching. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 632–634. ACM Press, 2002. 145
[24] Mao Chen, Andrea S. LaPaugh, and Jaswinder Pal 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*, pages 65–72, 2002. 142

[25] R. Cooley. *Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data*. PhD thesis, University of Minnesota, 2000. 142
[26] Robert Cooley, Bamshad Mobasher, and Jaideep Srivastava. Data preparation for mining world wide web browsing patterns. *Knowledge and Information Systems*, 1(1):5–32, 1999. 142, 143
[27] Boris Diebold and Michael Kaufmann. Usage-based visualization of web localities. In *Australian symposium on Information visualisation*, pages 159–164, 2001. 142
[28] Magdalini Eirinaki and Michalis Vazirgiannis. Web mining for web personalization. *ACM Transactions on Internet Technology (TOIT)*, 3(1):1–27, 2003. 145, 146
[29] Oren Etzioni. The world-wide web: Quagmire or gold mine? *Communications of the ACM*, 39(11):65–68, 1996. 140
[30] Kurt D. Fenstermacher and Mark 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)*, pages 205–212, 2002. 141
[31] Yongjian Fu, Mario Creado, and Chunhua Ju. Reorganizing web sites based on user access patterns. In *Proceedings of the tenth international conference on Information and knowledge management*, pages 583–585. ACM Press, 2001. 145
[32] Jiawei Han and Micheline Kamber. *Data Mining Concepts and Techniques*. Morgan Kaufmann, 2001. 143
[33] Birgit Hay, Geert Wets, and Koen Vanhoof. Clustering navigation patterns on a website using a sequence alignment method. 144
[34] Jeffrey Heer and Ed 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. 144
[35] John H. Holland. *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, 1975. Republished by the MIT press, 1992. 144
[36] Joshua Zhexue Huang, Michael Ng, Wai-Ki Ching, Joe Ng, and David Cheung. A cube model and cluster analysis for web access sessions. In R. Kohavi, B. Masand, M. Spiliopoulou, and J. Srivastava, editors, *WEBKDD 2001 - Mining Web Log Data Across All Customers Touch Points, Third International Workshop, San Francisco, CA, USA, August 26, 2001. Revised Papers*, volume 2356 of *Lecture Notes in Computer Science*, pages 48–67. Springer, 2002. 144
[37] Xiangji Huang, Nick Cercone, and Aijun An. Comparison of interestingness functions for learning web usage patterns. In *Proceedings of the eleventh international conference on Information and knowledge management*, pages 617–620. ACM Press, 2002. 143
[38] Karuna P. Joshi, Anupam Joshi, and Yelena Yesha. On using a warehouse to analyze web logs. *Distributed and Parallel Databases*, 13(2):161–180, 2003. 143
[39] Tapan Kamdar. Creating adaptive web servers using incremental web log mining. Master's thesis, Computer Science Department, University of Maryland, Baltimore County, 2001. 145
[40] Kosala and Blockeel. Web mining research: A survey. *SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery & Data Mining, ACM*, 2(1):1–15, 2000. 140
[41] Bin Lan, Stephane Bressan, Beng Chin Ooi, and Kian-Lee Tan. Rule-assisted prefetching in web-server caching. In *Proceedings of the ninth international conference on Information and knowledge management (CIKM 2000)*, pages 504–511. ACM Press, 2000. 145

[42] Tianyi Li. Web-document prediction and presending using association rule sequential classifiers. Master's thesis, Simon Fraser University, 2001. 145

[43] Bamshad Mobasher, Honghua Dai, Tao Luo, and Miki Nakagawa. Effective personalization based on association rule discovery from web usage data. In *Web Information and Data Management*, pages 9–15, 2001. 145

[44] Behzad Mortazavi-Asl. Discovering and mining user web-page traversal patterns. Master's thesis, Simon Fraser University, 2001. 144

[45] Eleni Stroulia Nan Niu and Mohammad 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)*, pages 53–64. IEEE, 2002. 144, 145

[46] Alexandros Nanopoulos, Dimitrios Katsaros, and Yannis Manolopoulos. Exploiting web log mining for web cache enhancement. In R. Kohavi, B. Masand, M. Spiliopoulou, and J. Srivastava, editors, *WEBKDD 2001 - Mining Web Log Data Across All Customers Touch Points, Third International Workshop, San Francisco, CA, USA, August 26, 2001. Revised Papers*, volume 2356 of *Lecture Notes in Computer Science*, pages 68–87. Springer, 2002. 143, 145

[47] Alexandros Nanopoulos, Maciej Zakrzewicz, Tadeusz Morzy, and Yannis Manolopoulos. Indexing web access-logs for pattern queries. In *Fourth ACM CIKM International Workshop on Web Information and Data Management (WIDM'02)*, 2002. 143

[48] O. Nasraoui, F. Gonzalez, and 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*, pages 711–716, 2002. 144

[49] Shigeru Oyanagi, Kazuto Kubota, and Akihiko 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. 144

[50] Hye-Young Paik, Boualem Benatallah, and Rachid Hamadi. Dynamic restructuring of e-catalog communities based on user interaction patterns. *World Wide Web*, 5(4):325–366, 2002. 145

[51] Jian Pei, Jiawei Han, Behzad Mortazavi-asl, and Hua Zhu. Mining access patterns efficiently from web logs. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 396–407, 2000. 143

[52] Pilot Software. Web Site Analysis, Going Beyond Traffic Analysis `http://www.marketwave.com/products_solutions/hitlist.html`, 2002. 141

[53] John R. Punin, Mukkai S. Krishnamoorthy, and Mohammed J. Zaki. Logml: Log markup language for web usage mining. In R. Kohavi, B. Masand, M. Spiliopoulou, and J. Srivastava, editors, *WEBKDD 2001 - Mining Web Log Data Across All Customers Touch Points, Third International Workshop, San Francisco, CA, USA, August 26, 2001. Revised Papers*, volume 2356 of *Lecture Notes in Computer Science*, pages 88–112. Springer, 2002. 141

[54] T. B. Pedersen S. Jespersen and J. Thorhauge. A hybrid approach to web usage mining. Technical Report R02-5002, Department of Computer Science Aalborg University, 2002. 144

[55] J. Ben Schafer, Joseph A. Konstan, and John Riedl. E-commerce recommendation applications. *Data Mining and Knowledge Discovery*, 5(1-2):115–153, 2001. 145

[56] Cyrus Shahabi and Farnoush Banaei-Kashani. A framework for efficient and anonymous web usage mining based on client-side tracking. In R. Kohavi, B. Masand, M. Spiliopoulou, and J. Srivastava, editors, *WEBKDD 2001 - Mining Web Log Data Across All Customers Touch Points, Third International Workshop, San Francisco, CA, USA, August 26, 2001. Revised Papers*, volume 2356 of *Lecture Notes in Computer Science*, pages 113–144. Springer, 2002. 141, 144

[57] Cyrus Shahabi and Yi-Shin Chen. Improving user profiles for e-commerce by genetic algorithms. *E-Commerce and Intelligent Methods Studies in Fuzziness and Soft Computing*, 105(8), 2002. 144

[58] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, and Pang-Ning Tan. Web usage mining: Discovery and applications of usage patterns from web data. *SIGKDD Explorations*, 1(2):12–23, 2000. 141, 145, 146

[59] G. Stumme, A. Hotho, and B. Berendt. Usage mining for and on the semantic web. In *National Science Foundation Workshop on Next Generation Data Mining*, 2002. 143

[60] Pang-Ning Tan and Vipin Kumar. Modeling of web robot navigational patterns. In *WEBKDD 2000 - Web Mining for E-Commerce – Challenges and Opportunities, Second International Workshop*, August 2000. 142

[61] Pang-Ning Tan and Vipin Kumar. Discovery of web robot sessions based on their navigational patterns. *Data Mining and Knowledge Discovery*, 6(1):9–35, 2002. 142

[62] Fergus Toolan and Nicholas Kushmerick. Mining web logs for personalized site maps. 145

[63] Debra VanderMeer, Kaushik Dutta, and Anindya Datta. Enabling scalable online personalization on the web. In *Proceedings of the 2nd ACM E-Commerce Conference (EC'00)*, pages 185–196. ACM Press, 2000. 145

[64] Yi-Hung Wu and Arbee L. P. Chen. Prediction of web page accesses by proxy server log. *World Wide Web*, 5(1):67–88, 2002. 145

[65] Yunjuan Xie and Vir V. Phoha. Web user clustering from access log using belief function. In *Proceedings of the First International Conference on Knowledge Capture (K-CAP 2001)*, pages 202–208. ACM Press, 2001. 144

[66] Osmar R. Zaïane. Web usage mining for a better web-based learning environment. In *Proceedings of Conference on Advanced Technology for Education*, pages 450–455, 2001. 145