Journal of Computer Science 1 (3): 413-418, 2005
ISSN 1549-3636

A Classification of Techniques for Web Usage Analysis

Haider Ramadhan, Muna Hatem, Zuhoor Al-Khanjri and Swamy Kutti
Department of Computer Science, Sultan Qaboos University, P.O. Box 36, Muscat 123, Oman

Abstract: Techniques which assist in recognizing various access patterns and interests of the Web users are normally referred to as an application of Web usage mining. The application provides useful insights to address crucial points such as user interests into a particular page, server load balancing, Web site reorganization, clustering of similar browsing patterns, classification of user profiles, content caching or data distribution and replication. In this study, we provide an updated focused survey of major aspects and problems related to the task of modeling the user behavior. We also point out some recent advancement in areas related to automatic Web navigation and implicit capturing of user interests in regard to a particular page. Main future directions are also outlined in the conclusion.

Keywords: Web Usage Analysis, User Access Patterns, Automatic Web Navigation, Web Mining

INTRODUCTION

Explosive increase in the use of the Internet has made automatic knowledge extraction from various web log files a necessity. This can be used to improve the effectiveness of the web sites by adapting the information structure of the sites to the user behavior. The ease and speed with which business transactions can be carried out over the Web has been a key driving force in the rapid growth of e-commerce. Specifically, e-commerce activity that involves the end user is undergoing a significant revolution. The ability to track user browsing behavior down to individual mouse clicks has brought the vendor and end customer closer than ever before.

Service providers can now clearly recognize user visiting patterns to their sites and pages and hence can reorganize their site structure as per the interests exhibited by their users. For example, by automatically analyzing the interests of the users in various links on the their main index page, vendors can reorganize the structure of that index page by placing popular links towards the top portions of the page, hence making their services more easily accessible.

Through the ability of automatically capturing users interests in particular pages, site owners can easily get implicit ratings about their pages and use such information in page reorganization or relocation. Moreover, with the advancement in Automatic Web Navigation [36], now it is possible to discover user access patterns from the log files and construct an overall model about the user. This model can be used to perform Web surfing on the behalf of the user by pre-fetching pages included in the user model (see more examples in section 4). To sum it up, it is now possible for a vendor to personalize his product message for individual customers at a massive scale, a phenomenon that is being referred to as *mass customization.* The scenarios described above are some of the many possible applications of Web Usage mining, which is the process of applying data mining techniques to the discovery of usage patterns from Web data [8, 31], targeted towards various applications.

WEB MINING

Web servers use the log files to record an entry for every single access they get. As the complexity of the web site or application increases, simple statistics give no meaningful hints on how the web site is being used. Moreover, the log files of popular web sites may grow of several hundreds of megabytes per day, making analysis tasks awkward. See [8, 31] on description of techniques for mining information and knowledge from large databases. Web mining refers to the application of such techniques to web data repositories, to enhance the analytical capabilities of the known statistical tools. An early taxonomy of web mining has been proposed in [3].

Web mining involves three tasks: (1) structure mining, (2) content mining and (3) usage mining. Web structure mining refers to the process of information extraction from the topology of the Web and aims at extracting data which describes the organization of the content. *Intra-page* structure information includes the arrangement of various HTML or XML tags within a given page. This can be represented as a tree structure, where the *html* tag becomes the root of the tree. The principal kind of *inter-page* structure information is hyper-links connecting one page to another. Alternatively, web structure mining has application in categorizing web pages. In [32], a method is described to discover authority sites (authorities) for the subjects and overview sites (hubs) pointing to the authorities.

Web content mining is the process of extracting useful information from the web sites and the pages they are composed of. One of the main challenges is the definition of what the web content is. Web content is

composed of multiple data types: text, images, audio, video, metadata and hyperlinks and multimedia data mining has become a specific instance of web content mining. As the greatest percentage of web content is unstructured text, great relevance is given to knowledge discovery in text [33, 15, 2]. It is worth noting that Web structure mining projects such as [39, 40] and Web content mining projects such as [41, 42] are beyond the scope of this study.

Web usage mining is devoted to the investigation on how people use web pages they access and on recognizing their navigational behavior. It involves automatic discovery of user access patterns from one or more Web servers or clients. Web usage mining is based on the analysis of secondary data describing interactions between users and the Web. Web usage data include data recorded in Web server access logs, proxy server logs, browser logs, user profiles, registration data, user sessions and transactions, cookies, user queries, bookmark data, mouse clicks and scrolls and any other data derived from the above interactions [4]. As it can be easily understood, boundaries between these three categories are blurred [4, 34]. As our main concern is with user access behavior, in this study we mainly concentrate on the usage mining as a mean to track the behavioral patterns of users surfing either a web site or a page.

WEB USAGE ANALYSIS

Web usage analysis relates to the development of techniques for discovering and/or predicting the behavior of a user interacting with the web. The data to be analyzed are (see Fig. 1) data logged by the user client (e.g. web browser) or server side (web server, proxy server, application server). Data logged at different locations represent different navigation patterns: client side data describe, in general, single-user multi-site navigation, server side logs describe multi-user single-site interaction and proxy server logs describe multi-user multi-site interaction. Widely adopted log file formats are described in [29, 30].

Server side logged data include client IP address (machine originating the request, maybe a proxy), user ID (if authentication is needed), time/date, request (URI, method and protocol), status (action performed by the server), bytes transferred, referrer and user agent (operative system and browser used by the client). Since this information is incomplete (e.g., not showing hidden request parameters automatically downloaded using the POST command) and not entirely reliable, the information should be integrated using packet sniffers and, where available, application server log files. Client side data collection has been implemented using remote agents (Java applets, HTML embedding Java script code) or ad-hoc browsers. Recently, especially tailored browsers have been used to implicitly and

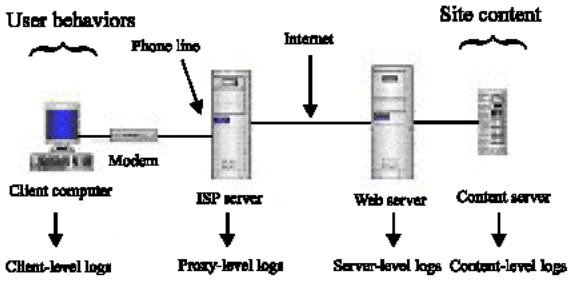

Fig. 1: Various Data Sources

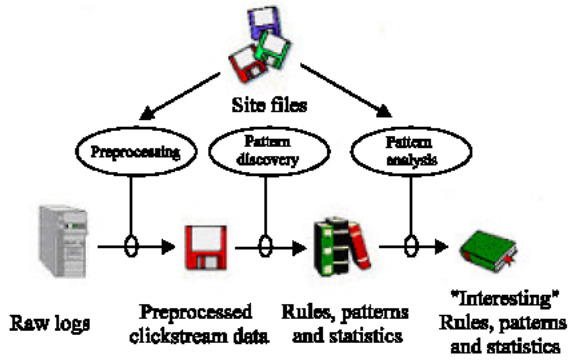

Fig. 2: Phases of Web Usage Mining

automatically capture users interests in a particular page and provide an overall rating for that page [45, 46]. Proxy server data describe, on one side, access to cached pages and on the other, access to sites from actual clients seen as a single anonymous entity from the web server. In [5] web usage mining techniques are analyzed. Three different phases, as shown in Fig. 2, are identified: pre-processing, pattern discovery and pattern analysis.

Pre-Processing: In this phase, abstraction data may be built representing, as examples, users (single actor using a browser to access files served by a web server), page views (the set of files served to the browser in response to a user action such as a mouse click), click-streams (a sequential series of page views requests), user sessions (click stream for a single user across the Web) and server sessions (set of page views for a user session on a single web site). See [9] for more definitions of the terms relevant for web usage analysis.

A particular stress is given to data preparation and preprocessing. As stated earlier, data may be incomplete (especially when client side logs are unavailable) leading to difficulties in user identification and difficulty to detect the user session termination (a 15 and 30 minute default time is normally assumed [10, 43] for session termination).

Depending on the data actually available for the analysis, typically known problems in user behavior reconstruction include: multiple server sessions

associated to a single IP client address (as in the case of users accessing a site through a proxy); multiple IP addresses associated to multiple server sessions (the so called mega-proxy problem); user accessing the web from multiple machines; and a user using multiple agents to access the Web. Assuming that the user has been identified, the associated click-stream has to be divided into user sessions. As expected, the first relevant problem is the identification of the session termination. Other relevant issues are the need to access application and content server information and the integration with proxy server logged information relative to cached resources access. Related to the usage preprocessing is the content preprocessing. Page views might be classified or clustered depending on their intended use and the results of this process be used to limit discovered usage patterns.

Pattern Discovery: Techniques adopted for this phase, strictly depend on the aim of the analysis. Methods available draw upon several fields such as statistics, data mining, machine learning and pattern recognition. Statistical analysis is in general applied to discover information such as most accessed pages or average length of a navigation path through a web site [35]. Use of association rules to find correlation between pages most often referenced together in a server session with a support value exceeding a given threshold is discussed in [35]. The results may find application in developing marketing strategies for e-business sites as well as for providing hints for restructuring a web site. Clustering is used to group items having similar characteristics.

In this case clustering may be used to group users exhibiting similar navigation behavior (usage clusters) or groups of pages having a related content (page clusters). In the first case, the information is again relevant to marketing scopes while, in the second one, it might be used by search engines. Classification techniques are often used to associate navigation behaviors to groups of users (or profiles). See [36, 43] for discussion on the application of sequential pattern discovery techniques to identify set of items followed by further items in a time ordered sequence. This is relevant for marketing purposes, i.e. for placing advertisements along the navigation path of certain users. Dependency modeling is also used with the goal of developing a model to represent significant dependencies among various variables in the web (for instance, modeling the stages a user undergoes during a visit to an on-line store). This is useful not only in predicting the user behavior but also in predicting web resource consumption. With Semantic Web [44], pattern discovery benefits from semantics included into web pages. This enables mapping Http requests to meaningful units of application events. Ontology, Resource Description Framework (RDF) repository and user profile can be updated with new information. In addition to the relationships between concepts, ontology also contains logical axioms that enable inferring new knowledge. The meaning that is constituted by the set of web pages accessed can be captured and taken into account. Hence, usage pattern can reveal semantic relationships that can help in learning the ontology itself or ontology instances.

Pattern Analysis: The last phase deals with pattern analysis. The aim is to filter out irrelevant information and extract only interesting information from the output of pattern discovery phase. One of the approaches used in analyzing patterns is the use of visualization techniques, such as graphing patterns, charts, diagrams, coloring schemes and coordinated views. The goal is to highlight overall patterns and trends in the data [8]. Relational databases have also been used for this phase. See [18] for description on the modeling of the data in a data cube to perform OLAP (On-Line Analytical Processing) operations.

MAIN APPLICATIONS

According to [4], applications of web usage mining may be classified into two categories: those dedicated to the discovery of user access patterns (behaviors) and hence customizing the Web the site accordingly [11, 12, 13] and those dedicated to an architectural (related to site topology) improvement of the web site effectiveness [14]. These two applications are discussed below.

Web Site Improvement: Making dynamic recommendations to a user based on his profile as well as navigation patterns has great relevance to e-business applications. Most of the sub-fields in this area are based on applying some sort of intelligent techniques, which include machine learning, knowledge representation and automated reasoning. See [37] for knowledge discovery process to recognize marketing intelligence from web data and use the information for implementing automatic Web navigation on behalf of the user, i.e. using information compiled from the user access patterns, to automatically pre-fetch updated pages without any manual interference from the user. As stated earlier, ad-hoc browsers can also provide local logs which can help in the discovery of the user interests in a particular page. The logs are then used to infer user intention and hence provide implicit page ratings. This is done through recording the user actions in terms of mouse clicks and movements, scrolling and elapsed time [45]. Other browsers provide more sophisticated indicators such as add to bookmark, print, forward, number of visits (visit frequency), save and more effective total visit time calculations [46]. Popularity of the links in relation to their functional locations in the web page is another application which is attracting more attention within the e-Business

community. Systems incorporating such feature not only analyze links as per their popularity but also suggest an optimal page design based on the frequency of the visits and the time spent on the links, hence adding a novel improvement to adaptive Web systems [46].

Moreover, the analysis of the web site usage may give important hints for site redesign, so as to enhance accessibility and attractiveness. Web usage mining may also provide insight into web traffic behavior, allowing to develop adequate policies for content caching and distribution, load balancing, network transmission and security management.

In [6] an approach is described to the problem of the user navigation pattern reconstruction from the analysis of the logged data. This is relevant both for customizing the content presented to the user and for improving the site structure. Two techniques are identified: mapping the data to relational tables and then applying data mining algorithms or directly applying analysis algorithms to logged data. See [38] for an interesting method about modeling disk I/O as well as network or web traffic (in general, everything that can be described by Bursty and Selfsimilar sequences) starting from data mining techniques.

Performance Improvement: Another important application of Web usage mining deals with improving the Web Performance. Here too we need intelligent systems techniques to identify, characterize and understand user behaviors. In fact, users navigation activity determines network traffic and, as a consequence, influences web server performance. Resources of the server are concurrently accessed and consumed and performance metrics must be continuously tuned in order to make services available and reliable. In [16, 17] pre-fetching is proposed in order to improve Web latency. Of course, this technique is useful if pre-fetched object or page is the target of the next request of the user.

So, pre-fetching must be subordinated to user browsing activity prediction. Browsing strategies need to be classified and in [24] client-side traces are analyzed. Here, according to common sequences size, users are categorized. Pre-fetching from the server side is considered in [16]. In [17] time-series analysis and digital signal processing are used to model web users. During a session, when a resulting threshold is reached, pre-fetching starts. Anyway, as pointed out in [25], the number of clicks per session exhibits strong regularities. The observed distribution showed to be inverse Gaussian. The authors discuss how this limits pre-fetching strategies. The goal of improving Web performance can be reached if the Web workload is well understood (see [23] for a survey of proposed Web characterizations). The nature of Web traffic has been deeply studied and analyzed. The first set of invariants has been processed in [21] and in [20] a discussion about the self-similar nature of Web traffic is given. A workload characterization could be also used to create synthetic workload [22], in order to benchmark a site, a monitoring agent checks the resource usage metrics during the test to search for system and network bottlenecks. In [26] two important models characterizing user sessions are introduced: Customer Behavior Model Graph (CBMG) and Customer Visit Models (CVM). Unlike traditional workload characterizations, these models are designed specifically for e-Commerce sites, where user sessions are represented in terms of navigational patterns. Moreover, these models can be obtained from the HTTP logs available at the server side. The CBMG is a state transition graph, in which the nodes corresponds to states in the session. A state can be viewed as a collection of web pages semantically related (e.g., browsing, searching, paying, etc.). The arcs in the CBMG correspond to transitions between states. As in the Markov Chains, probabilities are associated with transitions. In [19] K-means algorithms are used to obtain clusters of CBMG with similar patterns. A CVM is a more compact model than the CBMG. It represents sessions as a collection of session vectors, one per sessions. A value in a session vector is the number of times that each of the different functions (i.e., state in a CBMG) were invoked during the session. In [27] fractal clustering is used to find similar patterns in a collection of CVMs. Such workload characterizations fit well the nature of web traffic, allowing performance analysis and observations about user common behaviors.

CONCLUSION

The area of Web usage mining is still in its preliminary stages, as a consequence, related issues need further exploration and research. Some important points that need special attention include: (1) as the content served by a web server becomes more and more dynamic, the need of integration in the analysis process of the data coming from sources different from Web servers become mandatory. In particular, as multiple applicative services are able (and have to) to identify single user sessions, the user session reconstruction might become relatively straightforward, (2) development of advanced tools to deal with highly structured content such as XML, which requires more than just text mining, (3) development of techniques which would improve automatic Web navigation and page pre-fetching on the behalf of the users, (4) use of overall user interests discovered to design future theme based search engines as opposed to current key word based engines and (5) establishment of more robust procedures for reconstructing user behavior patterns while visiting Web sites and pages. In the Web performance application field, an interesting research direction is represented by the generation of representative synthetic workload that cannot rely on

traditional tools. For instance, to benchmark an e-commerce site, different user behaviors must be taken into account. In [28], synthetic workload is created from CBMGs. More versatile user session models are needed to characterize security parameters; many web services provide server and/or client authentication as well as ciphered communications (e.g., HTTP over SSL). This important aspect of the Web strongly influences performance and it is widely used in e-commerce. In such context, site topology must be carefully structured and analyzed to let different users to easily surf towards functions which contribute towards placing purchase orders.

REFERENCES

1. Réka Albert, Hawoong, Jeong, Albert-Làszl and Barabàsi, 1999. Diameter of the World-Wide Web, in Nature, 401: 130- 131.
2. Vernon Leighton, H. and J. Srivastava, 1997. Precision Among WWW Search Services (Search Engines): Alta Vista, Excite, Hotbot, Infoseek, Lycos. http://www.winona.msus.edu/is-f/libraryf /webind2/webind2.htm
3. Robert Cooley, Bamshad Mobasher and Jaideep Srivastava, 1997. Web mining: Information and pattern discovery on the world wide web. In Intl. Conference on Tools with Artificial Intelligence, Newport Beach, IEEE, pp: 558-567.
4. Kosala, R. and H. Blockeel, 2000. Web Mining Research: A Survey. SIGKDD Explorations, 2: 1-15.
5. Srivastava, J., R. Cooley, M. Deshpande and P.-N. Tan, 2000. Web usage mining: Discovery and applications of usage patterns from web data. SIGKDD Explorations, Vol. 1.
6. Borges, J. and M. Levene, 2000. Data Mining of User Navigation Patterns, in Web Usage Analysis and User Profiling, Published by Springer-Verlag as Lecture Notes in Computer Science, 1836: 92-111.
7. Dai, H., T. Luo and M. Nakagawa, 2002. Discovery and Evaluation of Aggregate Usage Profiles for Web Personalization. In Data Mining and Knowledge Discovery, Kluwer Publishing, 6: 61-82.
8. Fayyad, U., G. Piatetsky-Shapiro and P. Smyth, 1994. From Data Mining to Knowledge Discovery: An overview. In Proc. ACM KDD.
9. Web Characterization Activity http://www.w3.org/WCA/
10. Catledge, L. and J. Pitkow, 1995. Characterizing browsing strategies in the World Wide Web, in Proceedings of the 3rd Intl. WWW Conference, Darmstadt, Elsevier, Germany, pp: 1065-1073.
11. Langley, P., 1999. User modelling in adaptive interfaces. In Proceedings of the 7th Intl. Conference on User Modelling, pp: 357-370.
12. Perkowitz, M. and O. Etzioni, 1997. Adaptive web sites: an AI challenge. In Proc. 15th Intl. Joint Conf. AI, pp: 16-23.
13. Perkowitz, M. and O. Etzioni, 1997. Adaptive Sites: Automatically Learning From User Access Patterns. In Proceedings of the 6th Intl. World Wide Web Conference, poster no. 722.
14. Spiliopoulou, M., C. Pohle and L.C. Faulstich, 1999. Improving the effectiveness of a Web site with Web usage mining. In Proceedings of the Workshop on Web Usage Analysis and User Profiling (WebKKD99), San Diego.
15. Vernon Leighton, H. and J. Srivastava, 1997. Precision Among WWW Search Services (Search Engines): Alta Vista, Excite, Hotbot, Infoseek, Lycos., http://www.winona.msus.edu/is-f/libraryf /webind2/webind2.htm
16. Pedmanabhan, V.N., J.C. Mogul, 1996. Using Predictive Pre-fetching to Improve World Wide Web Latency, Computer Communication Review, Vol. 26.
17. Cunha, C.R. and C.F.B. Jaccoud, 1997. Determining WWW User's Next Access and its Application to Pre-fetching, Proc. of the Intern. Symp. on Computers and Communication'97, Alexandria, Egypt, pp: 1-3.
18. Zaïane, O.R., M. Xin and J. Han, 1998. Discovering Web Access Patterns and Trends by Applying OLAP and Data Mining Technology on Web Logs, in Proc. Advances in Digital Libraries Conf. (ADL'98), Santa Barbara, CA.
19. Menascé, D.A., V.A.F. Almeida, R. Fonseca and M.A. Mendes, 1999. A Methodology for Workload Characterization of Ecommerce Sites, in Proc. of ACM Conf. on E-Commerce, Denver, CO.
20. Crovella, M.E. and A. Bestavros, 1996. Self-Similarity in World Wide Web Traffic: Evidence and Possible Causes. In Proc. of ACM SIGMETRICS.
21. Arlitt, M.F. and C.L. Williamson, 1996. Web Server Workload Characterization: The Search for Invariants. In Proc. of ACM SIGMETRICS.
22. Barford, P. and M. Crovella, 1998. Generating Representative Web Workloads for Network and Server Performance Evaluation. In Proc. of ACM SIGMETRICS.
23. Pitkow, J.E., 2000. Summary of WWW Characterization, The Web Journal.
24. Catledge, L.D. and J.E. Pitkow, 1995. Characterizing browsing strategies in the World Wide Web, Computer Networks and ISDN Systems 26: 1065-1073.
25. Huberman, B., P. Pirolli, J. Pitkow and R. Lukose, 2000. Strong regularities in WWW surfing. Science, Vol. 280.
26. Menascé, D.A., V.A.F. Almeida, R. Fonseca, M.A. Mendes, 2000. Scaling for E-business: Technologies, Models and Performance and Capacity Planning, Prentice Hall, NJ.

27. Menascé, D., B. Abrah_o, D. Barbar , V. Almeida and F. Ribeiro, 2002. Fractal Characterization of Web Workloads, World Wide Web Conference.
28. Ballocca, G., R. Politi, G. Ruffo and V. Russo, 2002. Benchmarking a Site with Realistic Workload, Tech. Report, -CSP/Università di Torino.
29. http://www.w3.org/Daemon/User/Config/Logging.html
30. Extended Log File Format W3C Working Draft WD-logfile-960323,http://www.w3.org/pub/WW/TR/ WDlogfile- 960323.html
31. Chen, M.S., J. Han, P.S. Yu, 1996. Data Mining: An Overview from a Database Perspective. IEEE Transactions on Knowledge and Data Engineering, 8: 866-883.
32. David Gibson, Jon M. Kleinberg and Prabhakar Raghavan, 1998. Inferring web communities from link topology. In HyperText.
33. Hearst, M., 1999. Untangling text data mining. In Proceedings of ACL'99: the 37th Annual Meeting of the Association for Computational Linguistics.
34. Cooley, R.W., 2000. Web usage mining: discovery and application of interesting patterns from web data. PhD thesis, University of Minnesota, USA.
35. Agrawal, R. and R. Srikant, 1994. Fast algorithms for mining association rules. Proc. VLDB-94.
36. Rakesh Agrawal and Ramakrishnan Srikant, 1995. Mining Sequential Patterns. IEEE 11th Intl. Conference on Data Eng., IEEE Computer Society Press.
37. Buechner, A.G., S.S. Anand, M.D. Mulvenna and J.G. Hughes, 1999. Discovering Internet Marketing Intelligence through Web Log Mining, ACM SIGMOD Record, Vol. 27.
38. Mengzhi Wang, Tara M. Madhyastha, Ngai Hang Chan, Spiros Papadimitriou and Christos Faloutsos, 2002. Data Mining Meets Performance Evaluation: Fast Algorithms for Modeling Bursty Traffic, ICDE.
39. Fayyad, U., G. Piatetsky-Shapiro and P. Smyth, 1994. From data mining to knowledge discovery: An overview. In Proc. ACM KDD.
40. Spertus, E., 1997. Parasite: Mining structural information on the web. Computer Networks and ISDN Systems: The Intl. J. Com. and Telecommunication Networking, 29:1205-1215.
41. Pazzani, M., L. Nguyen and S. Mantik, 1995. Learning from hotlists and coldlists: Towards a www information filtering and seeking agent. In IEEE 1995 Intl. Conf. Tools with Artificial Intelligence.
42. Balabanovic, M. and Y. Shoham, 1995. Learning information retrieval agents: Experiments with automated web browsing. In On-line Working Notes of the AAAI Spring Symposium Series on Information Gathering from Distributed, heterogeneous Environments.
43. Ramadhan, H., Z. Al-Khanjari, A. Al-Hamadani and S. Kutti, 2004. Automatic Construction of the User Web Access Profiles. Transactions on Systems, 3: 1497-1506.
44. www.SemanticWeb.org/knowmarkup.html
45. Claypool, M., P. Le, M. Wased and D. Brown, 2001. Implicit interest indicators. In Proc. 6th intl. Conf. Intelligent User Interfaces, pp: 33-40.
46. Chan, P.K., 1999. A non-invasive learning approach to building web user profiles, In KDD-99 Workshop on Web Usage Analysis and User Profiling, pp: 7-12.