ISSN No. 0976-5697

Volume 1, No. 4, Nov-Dec 2010

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

A Comprehensive Survey on Frequent Pattern Mining from Web Logs

Mrs. Poonam Mishra
Lecturer, Shri Vaishnav Institute of Management,
Indore, India
poonam_dwvd@rediffmail.com

Abstract: Web usage mining is the type of Web mining activity that involves the automatic discovery of user access patterns from one or more Web servers. As more organization rely on the Internet and the World Wide Web to conduct business, the traditional strategies and techniques for market analysis need to be revisited in this context. Organizations often generate and collect large volumes of data in their daily operations. Most of this information is usually generate automatically by Web servers and collected in server access logs. Other sources of user information include referrer logs which contains information about the referring pages for each page reference, and user registration or survey data gathered via tools such as CGI scripts. In this paper we have surveyed various applications of web usage mining and analyzed their productivity.

Keywords: Web usage mining, World Wide Web, Data mining, Web mining

I. INTRODUCTION

Web usage Mining is the application of data mining techniques to Web click stream data in order to extract usage patterns .As Web sites continue to grow in size and complexity, the results of Web usage mining have become critical for a number of applications such as Web site design, business and marketing decision support, personalization, usability studies, and network traffic analysis. The two major challenges involved in Web usage mining are preprocessing the raw data to provide an accurate picture of how a site is being used, and filtering the results of the various data mining algorithms in order to present only the rules and patterns that are potentially interesting [10].Web usage mining consists of three phases, namely preprocessing, pattern discovery, and pattern analysis. Given its application potential, Web usage mining has seen a rapid increase in interest, from both the research and practice communities.

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

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

Web mining is the use of data mining techniques to automatically discover and extract information from Web documents and services [14].This area of research is so huge today partly due to the interests of various research communities, the tremendous growth of information sources available on the web and the recent interest in e-commerce Web mining decomposing into these subtasks, namely:

(a) Resource finding: the task of retrieving intended be documents.
(b) Information selection and preprocessing: automatically selecting and pre-processing specific information from retrieved Web resources.
(c) Generalization: automatically discovers general patterns at individual Web sites as well as across multiple sites.
(d) Analysis: validation and /or interpretation of the mined patterns.

The Web is an excellent tool to deliver on-line courses in the context of distance education. However, counting only on web traffic statistical analysis does not take advantage in the potential of hidden patterns inside the web logs. Web usage mining is a non-trivial process of extracting useful implicit and previously unknown patterns from the usage of the Web. Significant research is invested to discover these useful patterns to increase profitability of e-commerce sites. However, the goals of these applications and methods," turning visitors into purchasers", are different the goals in e-learning: "turning learners into effective better learners".

While some tools using data mining techniques to help educators and learners are being developed, the research is still in its infancy. In addition, with the awareness of the potential advantages of integrated web usage mining and the sufficient data recorded by web servers, there is need for more specialized logs from the application side to enrich the information already logged by the web server. This added value by specific event recording on the e-learning side will give click streams and the patterns discovered a better meaning and interpretation [13].

Web personalization is the process of customizing a Web site to the needs of specific users, taking advantage of the knowledge acquired from the analysis of the user's navigational behavior (usage data) in correlation with other information collected in the Web context, namely, structure, content, and user profile data. [3] Due to the explosive growth of the Web, the domain of Web personalization has gained great momentum both in the research and commercial areas. We present a survey of the use of Web mining for Web personalization. More specifically, we introduce the modules that comprise a Web personalization system, emphasizing the Web usage mining module. A review of the most common methods that are used as well as technical issues that occur is given, along with a brief overview of the most popular tools and applications available from software vendors. Moreover, the most important research initiatives in the Web usage mining and personalization areas are presented.

II.EVOLUTION

Web usage mining is explored by various researchers and focuses on major research area. In 1996 O. Etzioni explored the question of whether effective Web mining is feasible in practice. He believed that the Web is too unstructured for Web mining to succeed. Indeed, data mining has been applied to database traditionally, yet much of the information on the Web lies buried in documents designed for human consumption such as home pages or product catalogs. Furthermore, much of the information on the Web is presented in natural language text with no machine-readable semantics; HTML annotations structure the display of Web pages, but provide little insight into their content [1]. In 1998 Alex G. Biichner, Maurice D. Mulveena proposed which combined existing online analytical mining as well as web usage mining approaches, and incorporates marketing expertise. The data that is considered not only covers various types of server and web Meta information, but also marketing data and knowledge. Furthermore, heterogeneity resolution thereof and Internet- and electronic commerce-specific pre-processing activities are embedded. A generic web log data hypercube is formally defined and schematic designs for analytical and predictive activities are given. From these materialized views, various online analytical web usage data mining techniques are shown, which include marketing expertise as domain knowledge and are specifically designed for electronic commerce purposes [2].

The behavior of Web site's users may change so quickly that attempting to make predictions, according to frequent patterns coming from the analysis of an access log file, becomes challenging [13]. In order for the obsolescence of the behavioral patterns to become as null as possible, the ideal method would provide frequent patterns in real time, allowing the result to be available immediately. In 2001 They proposed a method allowing to find frequent behavioral patterns in real time, whatever the number of connected users is. Considering how fast the frequent behavior patterns can change since the last analysis of the access log file, this result thus provide completely adapted navigation schemas for user behavior predictions. Based on distributed heuristic, their method also tackled problems with the data mining framework: Discovering "interesting zones" (a great number of frequent patterns concentrated over a period of time or discovering of "super-frequent" patterns), discovering very long sequential patterns and interactive data mining ("on the fly" modification of minimum support)[11].In 2002 Bamshad Mobasher, Honghua Dai, Tao Luo Miki Nakagawa described an efficient framework for Web Personalization based on sequential and non-sequential pattern discovery from usage data. Their experimental results performed on real usage data indicate that more restrictive patterns, such as continuous sequential patterns (e.g., frequent navigational paths) are more suitable for predictive tasks, such as Web perfecting, which involve predicting which item is accessed next by a user, while less constrained patterns, such as frequent item sets or general sequential patterns are more effective alternatives in the context of Web personalization and recommender systems [7].

In Ref [4] Web Usage Mining (WUM) focused on the interaction behavior between web users and requested Web pages in order to identify navigation patterns. They approached the evaluation educational site design as a WUM application. Focus was set on site structure, as well as usage and combination technological resources for accomplishing a learning activity. The WUM process adopted considers characteristics that are specific to the learning context. Even considering that the pattern evaluation phase was not completed, the results obtained demonstrated that the abstractions adopted and the yielded pattern types are suitable and useful for site usage evaluation. The domain expert has shown particular interest on sequential patterns, for they show most frequent students' trails on the site. These were considered a powerful tool for identifying possible site restructuring points. It is hoped that site restructuring becomes a less hazardous activity if comprehensive WUM infrastructure becomes available.

In 2003 Juan Vel'asquez, Hiroshi Yasuda and Terumasa Aoki proposed a way to study the visitor behavior in a Web site, based in web content and usage mining. A web site is a semi structured collection of different kinds of data, whose motivation is show relevant information to visitor and by this way capture her/his attention. Understand the specifics preferences that define the visitor behavior in a web site, is a complex task. An approximation is suppose that it depend the content, navigation sequence and time spent in each page visited. These variables can be extracted from the web log files and the web site itself, using web usage and content mining respectively. Combining the describe variables, a similarity measure among visitor sessions is introduced and used in a clustering algorithm, which identifies group of similar sessions, allowing the analysis of visitors behavior. In order to prove the methodology's effectiveness, it was applied in a certain web site, showing the benefits of the described approach [5]. In July 2004 They proposed the impact of developing the WUM preprocessing phase according to concepts, problems and goals specific to Web-based learning environments. They present a tool prototype that automates typical tasks performed in the pre-processing phase, by offering operators that implement these tasks. The tool seeks for the active involvement of domain-related people(e.g. instructors).Functionality makes easier the alignment of mining goals with required pre-processing tasks, by addressing the configuration of operators visually and by enabling the reuse of existing configurations[8].

In Aug 2004, Xin Jin, Yanzan Zhou, Bamshad Mobasher have developed a unified framework for the discovery and analysis of Web navigational patterns based on PLSA. Probabilistic Latent Semantic Analysis (PLSA) is particularly useful in this context, since it can uncover latent semantic associations among users and pages based on the co-occurrence patterns of these pages in user sessions [6]. They show the flexibility of this framework in characterizing various relationships among users, user tasks and Web objects. Since these relationships are measured in terms of probabilities, They are able to use probabilistic inference to perform a variety of analysis tasks such as task identification and user segmentation, as well as predictive tasks such as collaborative recommendations. They have demonstrated the effectiveness of their approach through experiments performed on two real-world data sets.

In Sep 2004 Daby M.Sow, David P. Olshefski, Madis Beigi and Guruduth Banavar introduced a new technique for perfecting web content by learning the access patterns of individual users. The prediction scheme for perfecting is based on a learning algorithm, called Fuzzy-LZ, which mines the history of user access and identifies patterns of recurring accesses. This algorithm is evaluated analytically via a metric called *learn ability* and validated experimentally by correlating learn ability with prediction accuracy. A web perfecting system that incorporates Fuzzy-LZ is described and evaluated. Their experiments demonstrated that Fuzzy-LZ perfecting provides a gain of 41.5 % in cache hit rate over pure caching. This gain is highest for those users who are neither highly predictable nor highly random, which turns out to be the vast majority of users in our workload. The overhead of our perfecting technique for a typical user is 2.4 perfected pages per user request [9].

In 2005, they described a framework for a recommender system that predicts the user's next requests based on their behavior discovered from Web Logs data. They have compared results from three usage mining approaches: association rules, sequential rules and generalized sequential rules. They have used two selection rules criteria: highest confidence and last subsequence. Experiments are performed on three collections of real usage data: one from an Intranet Web site and two from an Internet Web site.

In Ref [15] They proposed an approach for discovering the profiles of visitor groups. To this end, They begin by mapping user interests into symbolic objects, which is the basis of the Symbolic Data Analysis and represents here a successful interaction of the user with the web site. They identified groups of users with similar behavior by means of a dynamic clustering approach applying a context dependent dissimilarity measure. The method was applied to identify visitor groups of a web site in the educational domain and also to analyze the traces of different user behavior.

In 2006, Natheer Khasawneh and Chien-Chung Chan presented new techniques for preprocessing web log data including identifying unique users and sessions. They introduced a fast active user-based user identification algorithm with time complexity of *O (n)*. For session identification, we used an ontology based session identification algorithm that uses the website structure to identify users' sessions. Models were introduced for determining three website parameters: number of records per user, inactive user time and number of recorded records per second in the web log to support the active-user-based approach. The output of this preprocessing step can be used as an input for different data mining techniques.

In July 2007, I-Hsien Ting, Chris Kimble and Daniel Kudenko proposed a users' browsing behavior analysis approach which is based on applying web usage mining techniques. Two web usage mining techniques in the approach are introduced, including Automatic Pattern Discovery (APD) and Co-occurrence Pattern Mining with Distance Measurement (CPMDM). A combination method is also discussed to show how potential browsing problems can be identified [16].

In Dec 2007, they focused on extraction of Sequential Patterns (SPs) with very low support from a large preprocessed Web usage data, to discover the behaviors of minority users of a Web site. Due to the sequential nature of the Web user's activity, Sequential Pattern Mining (SPM) is particularly well adapted for the study of Web usage data. Traditional SPM techniques with very low support produce large number of SPs. They are unsuitable for extraction of knowledge about the minority users because of large diversified user's behaviors and difficult to locate. Here, They proposed a novel approach called Cluster and Extract Sequential Patterns (CESP) that works based on divisive principle, where initial large Web log data split into smaller clusters (sub-logs) through ART1 neural network based clustering, and then Apriori like SPM technique is applied on each Cluster to extract SPs which reveal the behaviors of minority users. Several experiments were conducted on diversified Web log files, enabled us to discover interesting SPs having very low support (0.06 %). The study reveals that discovery of such SPs by a traditional SPM algorithms were impractical.

In Feb 2008, They proposed a complete framework and findings in mining Web usage patterns from Web log files of a real Web site that has all the challenging aspects of real-life Web usage mining, including evolving user profiles and external data describing ontology of the Web content. Even though the Web site under study is part of a nonprofit organization that does not "sell" any products, it was crucial to understand "who" the users were, "what" they looked at, and "how their interests changed with time," all of which are important questions in Customer Relationship Management (CRM). Hence, They presented an approach for discovering and tracking evolving user profiles. They also described how the discovered user profiles can be enriched with explicit information need that is inferred from search queries extracted from Web log data. Profiles are also enriched with other domain-specific information facets that give a panoramic view of the discovered mass usage modes. An objective validation strategy is also used to assess the quality of the mined profiles, in particular their adaptability in the face of evolving user behavior.

III. RECENT SCENARIO

Sebastian A. Rios and Juan D. Velasquez [19] developed a Semantic WUM process, which uses a concept-based approach to add semantics into the mining process. The solution proposed, was applied to a real web site to

produce offline enhancements of contents and structure. The method was compared with four different WUM methods.

Yan LI, Boqin FENG and Qinjiao MAO proposed algorithms avoid the complicated procedure of mining site topology and don't produce the user privacy issues. The modification of the reference length of the pages after path completion has also been implemented; it is very helpful for more accurate investigation on the user access pattern [20].

Saud R.Aghabozorgi and The Ying Wah proposed an off-line model based web usage mining that is generated by clustering algorithm. Then, they will use users' transactions periodically to change the off-line model to a dynamic-model. This proposed approach will solve the problem of the decrease of accuracy in the off-line models over time resulted of new users joining or changes of behavior for existing users in model-based approaches. Finally, They discussed the on-line model for user behavior prediction in the web personalization system [21].

IV. APPLICATIONS

The general goal of Web Usage Mining is to gather interesting information about users navigation patterns .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 be used for various purposes [10].

A. Personalization of web content: Personalizing the Web Experience for a user is the holy grain of many Web-based applications. 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 [22, 23, and 24].The Web Watcher [25], Site Helper [26], Letizia [27], and clustering have all concentrated on providing Web Site Personalization based on usage information.

B. System Improvement: Web usage mining provides the key to understanding Web traffic behavior, which can in turn be used for developing policies for Web caching, network transmission [28], load balancing, or data distribution.

C. Web Site Design: Web usage mining provides detailed feedback on user behavior, providing the Website designer information on which to base redesign decisions.

D. Business Intelligence: 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.

V. CHALLENGES

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 user's privacy that is currently highly relevant for the whole data mining area.

A. The main challenge is to come up with guidelines and rules such that site administrators can perform various analyses on the usage data without compromising the identity of an individual user.

B. The W3C has an ongoing initiative called Platform for Privacy Preferences (P3P) [30].P3P provides a protocol which allows the site administrators to publish the privacy policies followed by a site in a machine readable format.

C. The European Union ,the United States, and all other countries are publishing very strict laws about privacy.[29]

VI. CONCLUSION AND FUTURE DIRECTIONS

With the growth of Web based application, specifically electronic commerce, there is significant interest in analyzing Web usage data to better understand Web usage, and apply the knowledge to better serve users. This has lead to a number of open issues in Web Usage Mining area. In many practical applications, due to the introduction of stricter laws, privacy respect represents big challenge. In this survey paper, we briefly explored various applications of web usage mining suggested by authors. We also analyzed some problems and challenges of Web usage mining. Anyway we believe that the most interesting research area deals with the integration of semantics within Web site design so to improve the results of Web Usage Mining applications. 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 Semantic Web.

VII. REFERENCES

[1] O. Etzioni, The world -wide Web: Quagmire or gold mine ? Communications of the ACM 39 (11) (1996) 65-68.

[2] Alex G. Biichner, Maurice D. Mulveena" Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining", (1998).

[3] M. Eirinaki , M. vazirgiannis, 'Web Mining for web personalization",ACM Transactions on Internet Technology (TOIT) 3 (1) (2003) 1-27.

[4] Leticia dos Santos Machado, Karin Becker, " Distance Education: a Web Usage Mining Case Study for the Evaluation of Learning Sites".2003.

[5] Juan Vel'asquez Hiroshi Yasuda and Terumasa Aoki "Combining the web content and usage mining to understand the visitor behavior in a web site", 2003.

[6] Xin Jin, Yanzan Zhou, Bamshad Mobasher"Web Usage Mining Based on Probabilistic LATENT Semantic Analysis",2004.

[7] Bamshad Mobasher, Honghua Dai, Tao Luo Miki Nakagawa,"Using sequential and Non-Sequential Patterns in Predictive Web Usage Mining Tasks",2002.

[8] Carlos G. Marquardt, Karin Becker Duncan D. Ruiz," A Pre-processing Tool for Web Usage Mining in the Distance Education Domain", 2004.

[9] Daby M.Sow,David P. Olshefski, Mandis Beigi and Guruduth Banavar,"Prefecting Based on Web Usage Mining",2004.

[10] J.Srivastava, R.Cooley, M.Deshpande, P-N,Web usage mining:discovery and applications of usage patterns from web data,SIGKDD Explorations 1 (2) (2000) 12-23.

[11] Florent Masseglia, Maguelonne Teisseire, Pascal Poncelet"Real time Web usage mining: a Heuristic based Distributed miner", 2001.

[12] Mathias Gery ,Hatem Haddad,"Evaluation of Web Usage Mining Approaches for User's Next Request Prediction",2005.

[13] O.R. Zaiane,Web usage mining for a better web-based learning environment,in: Proceeding of Conference on Advanced Technology for Education, 2001,pp.450-455.

[14] Kosala, Blockeel, Web Mining research : a survey , SIGKDD explorations: newsletter of the special interest group (SIG) on knowledge discovery & data mining , ACM 2 (1) (2000) .

[15] Alzennyr da Silva, Yves Lechevallier, Francisco de Carvalho, Brigite Trousse“ Mining Web Usage Data for Discovering Navigation Clusters”,2006.

[16] Natheer Khasawneh , Chien-Chung Chan “Active User-Based and Ontogy-Based Web Log Data Preprocessing for Web Usage Mining” ,Dec 2006.

[17] G T Raju, Kunal and P S Satyanarayana,”Knowledge Discovery from Web Usage Data: Extraction of Sequential Patterns through ART1 Neural Network based Clustering Algorithm”,Dec 2007.

[18] Olfa Nasraoui, Member, IEEE, Maha Soliman, Member IEEE, Esin Saka, Member, IEEE,Antonio Badia, Member, IEEE, and Richard Gerain,”A Web Usage Mning Framework for Mining Evolving User Profioes in Dynamic Web Sites”,IEEE Transactions on Knowledge and Data Engineering, Vol. 20, No.2,February 2008.

[19] Sebastian A. Rios and Juan D.Velasquez, “Semantic Web Usage Mining by a Concept-based Approach for Off-line Website Enhancement”, 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology.

[20] “Research on Path Completion Techniques in Web Usage Mining”Monis Akhlaq, M Noman Jafri, Muzammil A Khan, and Baber Aslam”Addressing Security Concerns of Data Exchange in AODV Protocol”.

[21] Saud R. Aghabozorgi and The Ying Wah,“Dynamic Modelling by Usage Data for Personalization” 2009 13th International Conference Information Visualization.

[22] G.Adomavicius, A.Tuzhilin,Extending recommender systems: A multidimensional Approach.

[23] 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.

[24] 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).

[25] T.Joachims,D.Freitag ,and T.Mitchell,Webwatcher:A tour guide for the world wide web.In The 15th International Conference on Artifical Intelligence,Nagoya,Japan .1997.

[26] D.S W.Ngu and X.Wu Sitehelper: A localized agent that helps incremental exploration of the world wide web .In 6th International World Wide Web Conference, Santa Clara ,CA,1997.

[27] H. Lieberman. Letizia: An agent that assists web browsing .In Proc.of the 1995 International Joint Conference on ArtificalIntelligence,Nagoya,Japan ,1997.

[28] E.Cohen , B. Krishnamurthy, and J. Rexford .Improving end-to-end performance of the web using server volumes and proxy filters.In Proc.ACM SIGCOMM,pages 241-253,1998.

[29] Roger Clarke. Internet privacy concerns conf the case for intervention.42(2):60-67,1999.

[30] Reagle Joseph and Cranor Lorrie Faith.The platform for privacy preferences. 42(2):48-55,1999.

