International Conference on Recent Innovations in Science, Agriculture, Engineering and Management
University College of Computer Applications Guru Kashi University, Bathinda, Punjab (India)
20th November 2017, www.conferenceworld.in ISBN: 978-93-86171-80-1

Web Usage Mining: A Review

Sunny Sharma[1], Sunita Mahajan[2], Vijay Rana[3]

[1,2,3] *Department of CSE, Arni University, (India)*

ABSTRACT

The manuscript discusses about web usage mining involves the automatic discovery and analysis of web user access patterns from one or more Web servers. Web usage mining is the uses of prominent data mining techniques to automatic discover the patterns of web user from Web data, in order to know and better serve the requirements of Web-based applications. Web usage mining consists of four phases, namely data collection, preprocessing, pattern discovery, and pattern analysis. Given its application potential this paper illustrates each of these phases in detail.

Keywords: Data Mining, Web Personalization, Web Usage Mining, World Wide Web.

I. INTRODUCTION

With the large amount of information available on the Web, it is much tedious to retrieve the information from the web. The ease and pace with which business communication can be conceded over the Web has been a key driving force in the quick escalation of electronic businesses. Specially, e-commerce activity so as to involve the end user is undergoing a major revolution. The capacity to track clients' surfing behavior over the web has brought the seller and customer nearer than ever before. It is now feasible for a seller to personalize his product message for individual clients at a huge scale, a phenomenon that is being referred to as mass customization. In this circumstance, the area of Web usage mining is an important source of ideas and methods for the achievement of personalization functionality. Web Usage mining [1], which is the practice of applying data mining techniques to the discovery of various usage patterns from Web data, targeted towards various applications. Data mining techniques [2] applied with the Web, called Web mining, can be generally classified into three programs, i.e. web content mining, web usage mining, and web structure mining.

Recently, a figure of approaches has been developed dealing with explicit aspects of Web usage mining [3] for the principle of automatically discovering user profiles. For instancee, Perkowitz and Etzioni [4] proposed the idea of optimizing the structure of Web sites based co-occurrence patterns of pages within usage data for the site. Schechter et al [5] have developed techniques for using path profiles of users to predict future HTTP requests, which can be used for network and proxy caching. Spiliopoulou et al [6], Cooley et al [7], and Buchner and Mulvenna [7] have applied data mining techniques to extract usage patterns from Web logs, for the purpose of deriving marketing intelligence. Shahabi et al [8], Yan et al [9], and Nasraoui et al [10] have proposed clustering of user sessions to predict future user behavior. Section 2 presents the various types of Web data that can be helpful for Web Usage mining. Section 3 describes various data sources involved in discovering usage

patterns from Web data. Section 4 provides a detailed taxonomy of Web Usage mining, and Section 5 concludes the paper.

II. WEB DATA

One of the mean steps in Knowledge Discovery in Databases is to build an appropriate target data set for the data mining tasks. In Web Mining, data can be collected at the client-side, proxy servers, server side, or obtained from an organization's database [11] (which contains business data or consolidated Web data). Each type of data collection contains diverse kinds of data. There are diverse of data that can be used in Web Mining. This paper classified such data into the following types:

I. CONTENT: The real data in the Web pages, i.e. the data the Web page was designed to convey to the users. This usually consists of, but is not limited to, text and graphics.

II. STRUCTURE: 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.

III. USAGE: Data that describes the pattern of usage of Web pages, such as IP addresses, page references, and the date and time of accesses.

IV. USER PROFILE: Data that provides demographic information about users of the Web site. This includes registration data and customer profile information.

III. DATA SOURCES

The web usage data together at the different sources will represent the navigation patterns of different segments of the overall Web sessions, ranging from single-user, single-site browsing behavior to multi-user, multi-site accesses patterns.

I. SERVER LEVEL COLLECTION: A Web server log is an important source for performing Web Usage Mining because it explicitly records [12] the browsing behavior of site visitors. The data recorded in server logs re- e ts the (possibly on current) access of a Web site by multiple users. These log les an be stored in various formats such as Common log or Extended log formats. However, the site usage data recorded by server logs may not be entirely reliable due to the presence of various levels within the Web environment. Cached page views are not recorded in a server log. In addition, any important information passed through the POST method will not be available in a server log. Packet sniffing technology [13] is an alternative method to collecting usage data through server logs. Packet sniffers monitor network coming to a Web server and extra t usage data directly from TCP/IP packets. The Web server an also store other kinds of usage information such as cookies and query data in separate logs. Cookies are tokens generated by the Web server for individual client browsers in order to automatically track the site visitors. Tracking of individual users is not an easy task due to the stateless connection model of the HTTP protocol. Query data is also typically generated by online visitors while searching for pages relevant to their information needs. Besides usage data, the server side also provides content

data, structure information and Web page meta-information (such as the size of a le and its last modified time). The Web server [14] also relies on other utilities such as CGI s scripts to handle data sent back from client browsers. Web servers implementing the CGI standard parse the URI1 of the requested le to determine if it is an application program. The URI for CGI programs may contain additional parameter values to be passed to the CGI application. On e the CGI program has completed its execution, the Web server send the output of the CGI application back to the browser.

II. CLIENT LEVEL COLLECTION: Client-side data collection can be implemented by using a remote agent by using JavaScript or Java applets or by modifying the sour e ode of an existing to enhance its data collection capabilities.

III. PROXY LEVEL COLLECTION: A Web proxy acts as an intermediate level between client browsers and Web servers. Proxy servers can be used to reduce the loading time of a Web page experienced by users as well as the network load at the server and client sides. The performance of proxy depends on their ability to predict future page requests correctly. Proxy traces may reveal the actual HTTP requests from multiple clients to multiple Web servers. This may serve as a data sour e for characterizing the browsing behavior of a group of anonymous users sharing a common proxy server.

IV. WEB USAGE MINING

The web usage mining generally includes the following several steps: data collection, data pre-preprocessing, and knowledge discovery and pattern analysis.

4.1 DATA COLLECTION: Data collection is the first step of web usage mining, the data authenticity and integrity will directly affect the following works smoothly carrying on and the final recommendation of characteristic service's quality. Therefore it must use scientific, reasonable and advanced technology to gather various data, towards web usage mining technology, the main data origin has three kinds: server data, client data and middle data (agent server data and package detecting).

4.2 DATA PREPROCESSING: Some databases are insufficient, inconsistent and including noise. The data pre-treatment is to carry on a unification transformation to those databases [15]. The result is that the database will to become integrate and consistent, thus establish the database which may mine. In the data pre-treatment work, mainly include data cleaning, user identification, session identification and path completion. Data Cleaning: The purpose of data cleaning is to eliminate irrelevant items, and these kinds of techniques are of importance for any type of web log analysis not only data mining. According to the purposes of different mining applications, irrelevant records in web access log will be eliminated during data cleaning. Since the target of Web Usage Mining is to get the user's travel patterns, following two kinds of records are unnecessary and should be removed: 1) The records of graphics, videos and the format information. The records have filename

suffixes of GIF, JPEG, CSS, and so on, which can found in the URI field of the every record; 2) The records with the failed HTTP status code. By examining the Status field of every record in the web access log, the records with status codes over 299 or under 200 are removed. It should be pointed out that different from most other researches, records having value of POST or HEAD in the Method field are reserved in present study for acquiring more accurate referrer information. User and Session Identification: The task of user and session identification is to find out the different user sessions from the original web access log. User's identification is, to identify who access web site and which pages are accessed. The goal of session identification is to divide the page accesses of each user at a time into individual sessions. A session is a series of web pages user browse in a single access. The difficulties to accomplish this step are introduced by using proxy servers, e.g. different users may have same IP address in the log. A referrer-based method is proposed to solve these problems in this study [16]. The rules adopted to distinguish user sessions can be described as follows:

a) The different IP addresses distinguish different users;

b) If the IP addresses are same, the different browsers and operation systems indicate different users;

c) If all of the IP address, browsers and operating systems are same, the referrer information should be taken into account. The Refer URI field is checked, and a new user session is identified if the URL in the Refer URI field hasn't been accessed previously, or there is a large interval (usually more than 10 seconds) between the accessing time of this record and the previous one if the Refer URI field is empty.

d. The session identified by rule 3 may contains more than one visit by the same user at different time the time-oriented heuristics is then used to divide the different visits into different user sessions. After grouping the records in web logs into user sessions, the path completion algorithm should be used for acquiring the complete user access path.

Path completion: Another critical step in data preprocessing is path completion. There are some reasons that result in path's incompletion, for instance, local cache, agent cache, "post" technique and browser's "back" button can result in some important accesses not recorded in the access log file, and the number of Uniform Resource Locators (URL) recorded in log may be less than the real one. Using the local caching and proxy servers also produces the difficulties for path completion because users can access the pages in the local caching or the proxy servers caching without leaving any record in server's access log. As a result, the user access paths are incompletely preserved in the web access log. To discover user's travel pattern, the missing pages in the user access path should be appended. The purpose of the path completion is to accomplish this task. The better results of data preprocessing, we will improve the mined patterns' quality and save algorithm's running time. It is especially important to web log files, in respect that the structure of web log files are not the same as the data in database or data warehouse. They are not structured and complete due to various causations. So it is especially necessary to pre-process web log files in web usage mining. Through data pre-processing, web log can be transformed into another data structure, which is easy to be mined.

4.3. KNOWLEDGE DISCOVERY Use statistical method [17] to carry on the analysis and mine the pre-treated data. We may discover the user or the user community's interests then construct interest model. At

present the usually used machine learning methods mainly have clustering, classifying, the relation discovery and the order model discovery. Each method has its own excellence and shortcomings, but the quite effective method mainly is classifying and clustering at the present.

4.4. PATTERN ANALYSIS Challenges of Pattern Analysis is to filter uninteresting information and to visualize and interpret the interesting patterns to the user. First delete the less significance rules or models from the interested model storehouse; Next use technology of OLAP and so on to carry on the comprehensive mining and analysis; Once more, let discovered data or knowledge be visible; Finally, provide the characteristic service to the electronic commerce

V. CONCLUSION

This paper has attempted to provide an up-to-date survey of the rapidly growing area of Web Usage mining. With the rapid growth of Web-based applications, especially electronic commerce, there is significant interest in analyzing Web usage data to better understand Web usage, and apply the facts to better serve the need of users. This has led to a number of commercial offerings for doing such analysis. However, Web Usage mining raises some hard questions that must be answered before robust tools can be developed. This article has aimed at describing such challenges, and the hope is that the research community will take up the challenge of addressing them.

REFERENCES

[1.] Mobasher, B. (2005). Web usage mining. In Encyclopedia of data warehousing and mining (pp. 1216-1220). IGI Global.

[2.] Hegland, M. (2001). Data mining techniques. *Acta numerica*, *10*, 313-355.

[3.] Mobasher, B., Cooley, R., & Srivastava, J. (2000). Automatic personalization based on web usage mining. *Communications of the ACM*, *43*(8), 142-151.

[4.] Mobasher, B., Cooley, R., & Srivastava, J. (1999). Creating adaptive web sites through usage-based clustering of URLs. In *Knowledge and Data Engineering Exchange, 1999.(KDEX'99) Proceedings. 1999 Workshop on* (pp. 19-25). IEEE.

[5.] Schechter, S., Krishnan, M., & Smith, M. D. (1998). Using path profiles to predict HTTP requests. *Computer Networks and ISDN Systems*, *30*(1-7), 457-467.

[6.] Spiliopoulou, M. (2000). Web usage mining for web site evaluation. *Communications of the ACM*, *43*(8), 127-134.

[7.] Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. N. (2000). Web usage mining: Discovery and applications of usage patterns from web data. *Acm Sigkdd Explorations Newsletter*, *1*(2), 12-23.

[8.] Shahabi, C., Zarkesh, A. M., Adibi, J., & Shah, V. (1997, April). Knowledge discovery from users web-page navigation. In *Research Issues in Data Engineering, 1997. Proceedings. Seventh International Workshop on* (pp. 20-29). IEEE.

[9.] Gibbons, P. B., Karp, B., Ke, Y., Nath, S., & Seshan, S. (2003). Irisnet: An architecture for a worldwide sensor web. *IEEE pervasive computing*, *2*(4), 22-33.

[10.] Nasraoui, O., Frigui, H., Joshi, A., & Krishnapuram, R. (1999, August). Mining web access logs using relational competitive fuzzy clustering. In *Proceedings of the Eight International Fuzzy Systems Association World Congress* (Vol. 1, pp. 195-204).

[11.] Srivastava, J., Cooley, R., Deshpande, M., & Tan, P. N. (2000). Web usage mining: Discovery and applications of usage patterns from web data. *Acm Sigkdd Explorations Newsletter*, *1*(2), 12-23.

[12.] Mobasher, B., Cooley, R., & Srivastava, J. (2000). Automatic personalization based on web usage mining. *Communications of the ACM*, *43*(8), 142-151.

[13.] Ansari, S., Rajeev, S. G., & Chandrashekar, H. S. (2002). Packet sniffing: a brief introduction. *IEEE potentials*, *21*(5), 17-19.

[14.] Dosztányi, Z., Csizmok, V., Tompa, P., & Simon, I. (2005). IUPred: web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, *21*(16), 3433-3434.

[15.] Sharma, S., & Rana, V. (2017). Web Personalization through Semantic Annotation System. *Advances in Computational Sciences and Technology*, *10*(6), 1683-1690.

[16.] Mahajan, S., Sharma, S., & Rana, V. (2017). Design a Perception Based Semantics Model for Knowledge Extraction. *International Journal of Computational Intelligence Research*, *13*(6), 1547-1556.

[17.] Devlin, B., & Roeder, K. (1999). Genomic control for association studies. *Biometrics*, *55*(4), 997-1004.