Web Usage Mining: A Review of Recent Works

Rosli Omar, Abu Osman Md Tap, Zainatul Shima Abdullah

Department of Computer Science, Kulliyah of Information and Communication Technology
International Islamic University Malaysia (IIUM), Kuala Lumpur 50728, Malaysia
E-mail: rasau68@gmail.com, abuosman@iium.edu.my, shima@iium.edu.my

***Abstract*— Web mining is the application of data mining on web data and web usage mining is an important component of web mining. The goal of web usage mining is to understand the behavior of web site users through the process of data mining of web access data. Knowledge obtained from web usage mining can be used to enhance web design, introduce personalization service and facilitate more effective browsing. This paper presents a review of literature containing latest works done in this field. Our objective is to provide an overview of web usage mining concepts relevant to pattern mining phase of web usage mining process. We provide review of pattern discovery algorithms which utilize association rules, classification and sequential patterns, and since sequential pattern mining is gaining much interest from WUM research community extra emphasis is given to related papers.**

***Keywords*— data mining; web usage mining; sequential patterns, association rules**

I. Introduction

The number of Internet applications has grown and continue to grow significantly, affecting the lives of people in various aspects of their life including education, health, business and etc. The convenience and flexibility of services offered by web applications are the contributing factors why web applications are fast gaining popularity. In the process, web applications almost invariably churn out huge data containing user transactions and activity logs of user operations. Within the broad conceptual framework of Knowledge Discovery from Databases (KDD)[1], many studies have been conducted to explore ways of extracting potentially useful information embedded from large databases which can enhance decision making process. The core process of KDD, referred to as data mining, constitutes a number of different tasks aimed at extracting frequent patterns including association rules and sequential patterns mining. The application of data mining on web data is termed as web mining [2]. Two different approaches were initially taken in defining Web mining. First was a process-centric view, which takes the view of Web mining as a sequence of tasks [3]. Second was a data-centric view, which defined Web mining in terms of the types of Web data that was being used in the mining process [2]. The second definition has become more acceptable, as is evident from the many approaches and this is the definition adopted by this paper.

Cooley, Mobasher and Srivastava [2] have further categorized web mining into three main components: web usage mining(WUM), web structure mining(WSM) and web content mining(WCM). WCM is the task of discovering useful information based on the content of web pages. Web contents include multimedia data, structured content such as XML documents, semi-structured such as HTML documents and unstructured data such plain text. Web content mining applications include the task of organizing and clustering the web pages based on content and as well as ranking of web pages based on contents. WSM focuses on the structure of web sites using source data in the form of the structural information present in Web pages; typical applications are link-based categorization of Web pages, ranking of Web pages through a combination of content and structure and reverse engineering of Web site models. Taxanomy of WUM literature is presented in Figure 1.

Due to enormous interest in this field, there are plenty of studies being done since the last decade. In addition, there are a number of existing works on reviewing web mining and WUM approaches by focusing issues at the different levels of WUM namely pre-processing, pattern discovery and pattern analysis as in [4][5][6][7][8][9]and [10]. Chitraa [4] and Hussain, Asghar and Masood [5] review previous works which deals with the issues at the pre-processing stage and discuss proposed techniques to overcome those issues. Citing [11] that since the 80% of time spent on WUM is spent on pre-processing raw data, the authors argued that considerable attention should be given to address problems at this stage in order to ensure accuracy of later phases of WUM. Both papers present studies on issues specific to steps involved in pre-processing phase which are data cleaning, data filtering, user and session identification and path completion. The papers analyze and evaluate proposed techniques.

Pabarskaite & Raudys [11] provide a more extensive review of WUM papers prior to 2005 which covers literature related to both the pre-processing and pattern discovery phases. Most of the rest of the reviews investigate studies of WUM on specific applications. For example, Lappas [7] surveys studies on WUM application in the areas which have direct impact to society such as e-government, e-education and e-politics. These are areas which were given less attention by WUM research community compared to business and computing. Review on the application of WUM to facilitate the prediction of future user request is done by Patil[6]. Accurate and efficient prediction of users future request will be able to overcome the propagation delay in heavily visited web sites. The author analyzes a

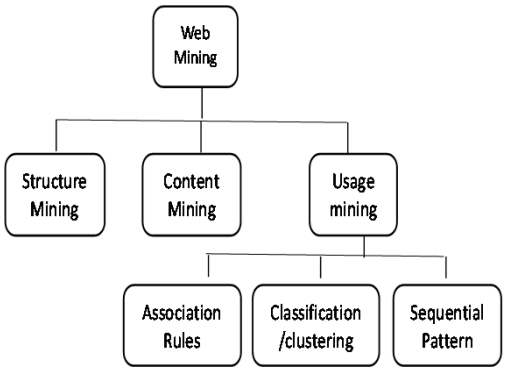

Figure 1. Taxonomy Of WUM

number of proposed techniques and highlights related future research issues.. Ngai reviews studies of WUM in customer relationship management [12] and fraud detection[13].

Our review of WUM literature focuses on the learning algorithms applied for pattern discovery, giving additional emphasis on sequence discovery methods. Since pattern discovery is an essential phase of WUM, we believe that it is pertinent to give due attention to pattern discovery algorithms as the algorithms have a direct impact to the accuracy and quality of resulting patterns. Furthermore, since user web access transactions are temporal in nature and sequential pattern techniques yield more accurate results [14], in this review extra attention is given to studies involving learning algorithms which utilize sequential pattern methods.

II. WEB USAGE MINING

Information used form mining the usage patterns often comes from Web log file generated by the web server which contains the user traversal data based on user interactions on the web. The web log data are usually presented in the some standard formats such as Common Log Format and Extended Log Format specified by the World Wide Web Community (W3C). Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site.

WUM itself can be classified further depending on the kind of usage data considered. The first kind of web log data is the one generated by the Web Server. Web server log data usually contain the IP address of clients, web site page requested and the access time. The second type of web log data is generated by the Application Server. Information contained in this log tends to be more specific, representing various kinds of business events related to the applications. The third kind of web log is called Application Level Log which is designated to individual applications. The content of the log is even more specific than the previous type. Due to the difficulty to collecting and accessing web log from various locations, most WUM methods rely on the Web Server Log file.

Web log information can be used for business intelligence in order to improve sales and advertisement by supplying product recommendation. It can identify frequent access behavior to improve the overall performance of future access. To improve latency time, caching and pre-fetching policies can be made based on frequent accessed pages. In addition, web log data can also be used to improve website design if we know access behavior of users. Finally, personalization for a user can be obtained through WUM.

III. THE MAIN PROCESSES OF WUM

WUM, from the data mining perspective, is the application of data mining techniques to discover useful knowledge of user behavior from Web data in order to understand and facilitate browsing experience for the user. Like most other data mining task, the process of WUM comprises of three main steps namely pre-processing, pattern discovery and pattern analysis. In the pre-processing step data are collected, then cleaned to remove unrelated objects such as graphic and multimedia entries. After that, the process of categorizing sessions according to different users is performed. A session is represented by a set of transactions of a user over a period of time as he traverses a web site[15]. The output of pre-processing phase serves as the input for the pattern discovery phase whereby learning algorithms are applied to mine for potential interesting patterns which may be embedded in the log data. Finally, in the pattern analysis phase, uninteresting rules or patterns found from the discovery phase are identified to be omitted. Pattern analysis technique is very much dependent upon the specific application that used them and one of the more popular pattern analysis application is SQL.

Sessionization, the process to identify the sessions from the raw data, is a major challenge, because the server logs do not always contain all the information needed. Moreover, the data have to be transformed into a suitable format before they can be used as the input for the mining algorithms. Once the data are prepared for mining, a data mining technique that suits the intended goal will be applied to the data. Finally, the results of the mining algorithms will be analyzed and interpreted into useful knowledge which can be used to facilitate decision making.

IV. RELATED WORKS ON WUM

This section describes recent development in the area of WUM, from the perspective of the different types of patterns being mined.

A. Basic Mining Algorithms

At the heart of WUM are the generic mining algorithms which perform the task of extracting frequent patterns from data files. Some of these algorithms include AprioriAll [16], Generalized Sequential Pattern (GSP) [17], Sequential Pattern Discovery using Equivalence classes (SPADE) [18], Frequent Pattern-Projected Sequential Pattern mining (FreeSpan) and Prefix-Span [19]. Although these algorithms

have the same objective of mining for frequent patterns, they employ different methods to achieve the goal. For WUM specific purposes, the algorithms are required to process only single-element sequences which are suitable for web navigational sequences. Some of the more popular WUM algorithms include: WAP-tree [20], and PLWAP-tree [21].

The Apriori-based algorithms require multiple scans of database. For each iteration i, the algorithms generate candidate itemsets of size i by joining frequent itemset of length (i 1) with itself and subsequently pruning out candidates which contain infrequent subsequences. By Apriori property, these candidates cannot be frequent since all subsets a frequent itemset are frequent. To compute the support values for all candidate itemsets being generated, the database is scanned again. Candidate itemsets which satisfy support threshold are considered frequent itemsets. The process continues until all frequent itemsets are discovered. This means that the database is access at least k number of times, where k is the maximum number of iterations it takes to mine all frequent itemsets. This is one of the main drawbacks of Apriori method. The non-Apriori techniques avoid the time consuming level wise candidate generation process by using pattern-growth trees [19] and database projections. Here, the original database is scanned not more than twice where initially efficient and compact conditional trees are constructed and populated with frequent itemsets. Then the algorithms mine the trees for frequent itemsets using divide-and-conquer strategy. Besides this, there are also some less popular approaches which integrate Apriori style techniques with other non-Apriori techniques like pattern growth, as in [22].

B. Mining association rules from WUM

Mining association rules from web data is well studied due to its popularity. Association rules in WUM describe the relationships between two or more web pages. For example, the association rule pattern page1 page3, where page1 and page3 are pages within a set of pages contain in user access session, states that sessions that contain page1 will most likely also contains page3. In other words, association rules in WUM describe web pages which are frequently visited together. Having this knowledge may help a web designer restructure the website by placing frequently visited pages closed to each other so as to enhance the speed of browsing. Online business can identify which items are frequently bought together, and design the promotional activities accordingly.

Recent studies involving mining association rules from WUM include [23], [24], and [25]. Reference [23] proposed a method for web path traversal pattern from web logs. The authors presented an efficient algorithm for web page prediction from large web logs visited by a user. A significant weight is assigned to each page based on time spent by user on each page, visiting frequency and click event done on each page. The objective is to mine weighted association rules to extract knowledge of user behavior from web logs.

A methodology for extracting useful information from users history databases associated to an e-commerce website is presented in [24]. The authors described the main phases of WUM including data collection, data preprocessing, extraction and analysis of knowledge. Both unsupervised and supervised learning algorithms are applied for knowledge extraction. Results obtained shown to be useful for website designers, providing some guidelines for improving its usability and user satisfaction.

Langhnoja et al., [25] study user behavior from access patterns captured in the web log. Web usage mining includes three phases namely pre-processing, pattern discovery and pattern analysis. Their method combines the technique of clustering and association rule mining to extract significant user behaviors.

C. Mining for Classification and Clustering

An application of WUM for predicting user traversal pattern for a colleges web site can be found in [26]. This paper describes web usage mining for the college log files to analyze the behavioral patterns and profiles of users interacting with a Web site. The discovered patterns are represented as clusters that are frequently accessed by groups of visitors with common interests. In this paper, the visitors and hits were forecasted to predict the further access statistics.

Jagli & Oswal [26] provide an overview of the different stages involved in a general WUM process. As an example, a WUM approach is presented which is based on the use of fuzzy clustering to discover user categories starting based on usage patterns. Improving the quality of service by web site administrator is the goal in [27]. They proposed a method for predicting potential next page to be visited by the user using classification of user navigational behavior from historical records. They implement a supervised learning to train the classifier.

D. Mining Sequential Patterns in WUM

The application of sequential mining techniques on WUM from web log accesses, involves mining for sequential patterns of web pages visited by web users in order to understand user browsing behavior. Web access pattern tree (WAP-tree) mining [20] is a sequential pattern mining technique for web log access sequences. In the first stage, it scans the sequence database while constructing a compact prefix tree to store the web access sequence database. In the second stage, the mining algorithm uses the WAP-tree to mine frequent sequences from the WAP-tree. The mining process involves recursively re-constructing intermediate trees, starting with suffix sequences and ending with prefix sequences. The strength of WAP-tree method includes avoiding the time consuming candidate generating process. However, WAP-tree algorithm requires recursive reconstruction of intermediate WAP-trees during mining process. This process is very time-consuming.

To overcome above the problem, [21] proposed a technique for assigning a position code to each node of the tree. The code consists of single binary position codes of all of this nodes ancestors in the binary tree equivalent. Using the position code to label each node of the WAP-tree, the WAP-tree head links is built in a pre-order fashion rather than in the order the nodes arrive as done by the WAP-tree algorithm. The proposed algorithm, called Pre-Order Linked Web Access Pattern Tree (PLWAP), eliminates the need to reconstruct WAP-tree repeatedly.

Meanwhile, another similar approach to PLWAP was proposed by Vijayalakshmi et al, [28], called Adaptive Web Access Pattern Tree (AWAPT). The AWAPT approach is based on WAP-tree but AWAPT is more efficient since it does not require multiple reconstructions of intermediate trees during mining process. This is enabled by maintaining binary position codes which let the algorithm determine the suffix of any given frequent pattern without having to construct a corresponding WAP-trees.

PLWAP is extended to enable incremental mining of sequential patterns of web usage data [21]. Incremental mining allows mining process to be performed even when a database is being updated. This is achieved by using recent additional data sequences in an incremental manner in conjunction with already mined patterns. With this method, online prediction of next user request can be achieved. The authors propose two algorithms, RePL4UP (Revised PLWAP For UPdate), and PL4UP (PLWAP For UPdate). A tree structure is used to store incrementally updated web sequential patterns efficiently. As the database is updated with new user transactions, the algorithm does not require rescanning of the entire database to compute a new set of sequential patterns. The algorithms also handle the case where previous infrequent items becoming frequent. In the tree constructing stage, the RePL4UP is responsible for maintaining the position codes of infrequent items. During mining stage, only the recently added database sequences are scanned by RePL4UP, and subsequently it updates the existing PLWAP tree to revise the status of both previous infrequent and frequent items. The newly added sequences may result in previously infrequent items to become frequent and vice versa. Updating the tree is achieved with the information contained in the infrequent item position codes without having to recursively construct any additional PLWAP-tree. However, at the beginning PL4UP generally constructs a larger tree to store both frequent sequences as well as infrequent sequences which may become frequent later as the mining process progresses. The position code enables PLWAP to efficiently mine these trees to extract current frequent patterns as the database is being updated.

However, the above mining methods are limited to binary attributes only. In other words, to compute the support of frequent sequences, these algorithms only check for the presence or absence of these sequences in the sequence database. In real life, many applications, including WUM, involve non-binary attributes. For example in WUM, it may be useful to have a method for mining sequence of web pages that also factors in the duration of time a user spends on each page. This would give more insight into the browsing behavior of the user. Mining for high utility patterns involves the mining of pattern with non-binary attributes and considers non-binary values called utilities of web pages in a sequence. Zhou et al.[29] introduced the concept of utility in web path traversal mining model to express the significance of web pages in terms of browsing time spent by the user. Thilagu & Nadarajan [30] apply the notion of high utility patterns in conjunction with the transitional pattern mining in order to detect changes in user behavior from web usage log. The authors argue that in reality, human behavior is not static over time and subject to changes. Therefore it is useful to find a method which identifies milestones when changes occur. They constructed a two-phase algorithm which first finds all frequent patterns of high utility and then transitional patterns and their significant milestones are found based on these frequent patterns.

The utilities of subsequences in sequences in [30] approach are difficult to be calculated due to the three kinds of utility calculations. To simplify the utility calculation, [31] then presents a maximum utility measure, which is derived from the principle of traditional sequential pattern mining that the count of a subsequence in the sequence is only regarded as one. Hence, the maximum measure is properly used to simplify the utility calculation for subsequences in mining.

V. CONCLUSION

With the growth of web-based applications, there has been increasing research interest in the discovery and the analysis of web usage patterns. Understanding the browsing behavior of users and applying the discovered knowledge may provide potential increase to the quality of browsing experience. In this work we discuss recent WUM approaches for mining usage patterns. We described the overview of a general WUM process and we highlight recent works done in mining different types of frequent patterns although we laid extra emphasis on the mining of sequential patterns from web data and explain the mining process associated this type of patterns.

WUM is fast gaining attention from research community, as confirmed by the large number of works published on this topic and the variety of contexts where WUM processes can be conveniently applied. However, much work can still be done to enhance the efficacy of the overall process of discovery and analysis of usage patterns.

REFERENCES

[1] U. Fayyad, G. Piatetsky-shapiro, and P. Smyth, From Data Mining to Knowledge Discovery in, pp. 3754, 1996.

[2] R. Cooley, B. Mobasher, and J. Srivastava, Web Mining: Information and Pattern Discovery on the World Wide Web * 1 Introduction, pp. 558567, 1997.

[3] O. Etzioni, The World-Wide Web: quagmire or gold mine?, Commun. ACM, pp. 16, 1996.

[4] V. Chitraa, A Survey on Pre-processing Methods for Web Usage Data, vol. 7, no. 3, pp. 7883, 2010.
[5] T. Hussain, S. Asghar, and N. Masood, Web usage mining: A survey on pre-processing of web log file, 2010 Int. Conf. Inf. Emerg. Technol., pp. 16, Jun. 2010.
[6] U. Patil and S. Pardeshi, A Survey on User Future Request Prediction: Web Usage Mining, vol. 2, no. 3, pp. 14, 2012.
[7] G. Lappas, An Overview of Web Mining in Societal Benefit Areas Technological Educational Institution of Western Macedonia, 2007.
[8] R. Kosala and H. Blockeel, Web mining research: A survey, ACM Sigkdd Explor. Newsl., vol. 2, no. 1, 2000.
[9] F. M. Facca and P. L. Lanzi, Mining interesting knowledge from weblogs: a survey, Data Knowl. Eng., vol. 53, no. 3, pp. 225241, Jun. 2005.
[10] P. I. Hofgesang, Online Mining of Web Usage Data: An Overview, pp. 123, 2009.
[11] Z. Pabarskaite and A. Raudys, A process of knowledge discovery from web log data: Systematization and critical review, J. Intell. Inf. Syst., vol. 28, no. 1, pp. 79104, Dec. 2006.
[12] E. W. T. Ngai, L. Xiu, and D. C. K. Chau, Application of data mining techniques in customer relationship management: A literature review and classification, Expert Syst. Appl., vol. 36, no. 2, pp. 25922602, Mar. 2009.
[13] E. W. T. Ngai, Y. Hu, Y. H. Wong, Y. Chen, and X. Sun, The application of data mining techniques in financial fraud detection: A classification framework and an academic review of literature, Decis. Support Syst., vol. 50, no. 3, pp. 559569, Feb. 2011.
[14] M. Gry and H. Haddad, Evaluation of web usage mining approaches for users next request prediction, Proc. fifth ACM Int. Work. Web Inf. data Manag. - WIDM 03, p. 74, 2003.
[15] K. Sudheer Reddy, M. Kantha Reddy, and V. Sitaramulu, An effective data pre-processing method for Web Usage Mining, 2013 Int. Conf. Inf. Commun. Embed. Syst., pp. 710, Feb. 2013.
[16] R. Srikant and R. Agrawal, Mining sequential patterns: Generalizations and performance improvements, Adv. Database Technol., 1996.
[17] R. Srikant and R. Agrawal, Mining sequential patterns: Generalizations and performance improvements, Adv. Database Technol., 1996.
[18] M. Zaki, SPADE: An efficient algorithm for mining frequent sequences, Mach. Learn., pp. 3160, 2001.
[19] J. Han, J. Pei, Y. Yin, and R. Mao, Mining frequent patterns without candidate generation: A frequent-pattern tree approach, Data Min. Knowl. Discov., pp. 5387, 2004.
[20] J. Pei, J. Han, B. Mortazavi-Asl, and H. Zhu, Mining access patterns efficiently from web logs, Discov. Data Mining. Curr. , 2000.
[21] Y. I. Ezeife, Lu, Mining Web Log Sequential Patterns with Position, pp. 538, 2005.
[22] P. Tang, L. Rock, and M. P. Turkia, Mining Frequent Web Access Patterns with Partial Enumeration, pp. 226231, 2007.
[23] R. Agarwal, K. V. Arya, S. Shekhar, and R. Kumar, An Efficient Weighted Algorithm for Web Information Retrieval System, 2011 Int. Conf. Comput. Intell. Commun. Networks, pp. 126131, Oct. 2011.
[24] C. J. Carmona, S. Ramrez-Gallego, F. Torres, E. Bernal, M. J. del Jesus, and S. Garca, Web usage mining to improve the design of an e-commerce website: OrOliveSur.com, Expert Syst. Appl., vol. 39, no. 12, pp. 1124311249, 2012.
[25] S. G. Langhnoja, M. P. Barot, and D. B. Mehta, Web Usage Mining Using Association Rule Mining on Clustered Data for, vol. 02, no. 01, 2013.
[26] D. Jagli and S. Oswal, Web Usage Mining: Pattern Discovery and Forecasting, vol. 2, no. November, pp. 187190, 2012.
[27] S. C Kamoji and P. Naik, Mining the web data using data mining techniques for identifying and classifying the user access behavioral patterns, IOSR J. Comput. Eng., vol. 16, no. 2, pp. 6371, 2014.
[28] S. Vijayalakshmi, V. Mohan, and R. Suresh, Mining of Users Access Behaviour for Frequent Sequential Pattern From Web Logs, Int. J. Database Manag. Syst., vol. 2, no. 3, pp. 3145, Aug. 2010.
[29] L. Zhou, Y. Liu, J. Wang, and Y. Shi, Utility-Based Web Path Traversal Pattern Mining, Seventh IEEE Int. Conf. Data Min. Work. (ICDMW 2007), pp. 373380, Oct. 2007.
[30] M. Thilagu and R. Nadarajan, Investigating Significant Changes in Users Interest on Web Traversal Patterns, Int. J. Cybern. Informatics, vol. 2, no. 4, pp. 3955, Aug. 2013.
[31] G.-C. Lan, T.-P. Hong, V. S. Tseng, and S.-L. Wang, Applying the maximum utility measure in high utility sequential pattern mining, Expert Syst. Appl., vol. 41, no. 11, pp. 50715081, Sep. 2014.