Global Journals LaTeX JournalKaleidoscope™
Artificial Intelligence formulated this projection for compatibility purposes from the original article published at Global Journals. However, this technology is currently in beta. *Therefore, kindly ignore odd layouts, missed formulae, text, tables, or figures.*

A Survey on Web Usage Mining

Dr J. Vellingiri[1] and S. Chenthur Pandian[2]

[1] Kongunadu College of Engineering and Technology

Received: 9 January 2011 Accepted: 4 February 2011 Published: 17 February 2011

Abstract

Now a day World Wide Web become very popular and interactive for transferring of information. The web is huge, diverse and active and thus increases the scalability, multimedia data and temporal matters. The growth of the web has outcome in a huge amount of information that is now freely offered for user access. The several kinds of data have to be handled and organized in a manner that they can be accessed by several users effectively and efficiently. So the usage of data mining methods and knowledge discovery on the web is now on the spotlight of a boosting number of researchers. Web usage mining is a kind of data mining method that can be useful in recommending the web usage patterns with the help of users? session and behavior. Web usage mining includes three process, namely, preprocessing, pattern discovery and pattern analysis. There are different techniques already exists for web usage mining. Those existing techniques have their own advantages and disadvantages. This paper presents a survey on some of the existing web usage mining techniques.

Index terms— Web Usage Mining, Preprocessing, Web Log, User Navigation.

1 Introduction

ecently, millions of electronic data are included on hundreds of millions data that are previously on-line today. With this significant increase of existing data on the Internet and because of its fast and disordered growth, the World Wide Web has evolved into a network of data with no proper organizational structure. In addition, survival of plentiful data in the network and the varying and heterogeneous nature of the web, web searching has become a tricky procedure for the majority of the users. This makes the users feel confused and at times lost in overloaded data that persist to enlarge. Moreover, e-business and web marketing are quickly developing and significance of anticipate the requirement of their customers is obvious particularly. As a result, guessing the users' interests for improving the usability of web or so called personalization has turn out to be very essential. Web personalization can be depicted as some action that builds the web experience of a user personalized according to the user's interest.

Generally, three kinds of information have to be handled in a web site: content, structure and log data. Content data contains of anything is in a web page, structure data is nothing but the organization of the content and usage data is nothing but the usage patterns of web sites.The usage of the data mining process to these dissimilar data sets is based on the three different research directions in the area of web mining: web content mining, web structure mining and web usage mining.

Web usage mining is the type of data mining process for discovering the usage patterns from web information for the purpose of understanding and better provide the requirements of web-based applications. Web usage mining imitates the actions of humans as they interact with the Internet. Examination of user actions in communication with web site can offer insights causing to customization and personalization of a user's web practice. As a result of this, web usage mining is of extreme attention for e-marketing and ecommerce professionals. Web usage mining involves of three phases, namely, preprocessing, pattern discovery and pattern analysis. There are different techniques available for web usage mining with its own advantages and disadvantages. This paper provides some discussion about some of the techniques available for web usage mining.

2 II.

3 Literature Survey

This section provides some discussion about several web usage mining techniques available today.

The discovery of the users' navigational patterns using SOM is proposed by Etminani et al., [1]. Huge amount of information are collected repeatedly by Web servers and gathered in access log files. Analysis of server access data can offer important and helpful data. Web usage mining is the technique of using data mining procedure for discovering the usage patterns from Web data and is targeted towards applications. It extracts the secondary information resulting from the interactions of the users through some period of Web sessions. Web usage mining involves three processes, namely preprocessing, pattern discovery, and pattern analysis. Provided its application possibility, Web usage mining has seen a quick raise in interest, from the research and practice area. The author used the Kohonen's SOM (Self Organizing Map) to preprocessed Web logs for extracting the common patterns.

Jianxi et al., [2] presented a Web usage mining technique based on fuzzy clustering in Identifying Target Group. Data mining is a process of non-trivial mining of inherent, previously unidentified, and highly R helpful data from very large quantity of information. Web mining can be defined basically as the usage of data mining procedures to Web data. Web usage mining is a significant and rapidly developing field of Web mining where many research has been performed previously. The author uses the fuzzy clustering technique for discovering groups that share similar interests and behaviors by examining the data gathered in Web servers.

Nina et al., [3] suggests a complete idea for the pattern discovery of Web usage mining. Web site creators must have clear knowledge of user's profile and site intentions and also emphasized information of the approach users will browse Web site. The creators can examine the visitor's behavior by means of Web analysis and identify patterns of the visitor's activities. This Web analysis includes the transformation and interpretation of the Web log records to identify the hidden data or predictive pattern by the data mining and knowledge discovery process. This result provides a great view coupled with the Web warehousing.

Wu et al., [4] given a Web Usage Mining technique based on the sequences of clicking patterns in a grid computing environment. Examining user's browsing pattern is an significant process of web usage mining. It can assist the web supervisors or creators enhance the web structure or increase the performance of the web servers. Mining on the sequences of such clicking patterns (MSCP) can be regarded as a data mining task. MSCP is generally an expensive procedure because of its significant quantity of time for computation and storage for archiving a large quantity of information. Running MSCP becomes ineffective or even not practical on a computer with restricted resources. The author discovers the usage of MSCP in a distributed grid computing surroundings and expresses its effectiveness by empirical cases.

Aghabozorgi et al., [5] proposed the usage of incremental fuzzy clustering to Web Usage Mining. Currently wide increase of information on the Web has produces a huge quantity of log records on Web server databases. By using the Web usage mining procedure on this huge quantity of historical information can identify potentially helpful patterns and expose user access patterns on the Web page. Cluster analysis has broadly been used to produce user behavior pattern on server Web logs. The majority of these off-line procedure have the drawback of reduction of accuracy over time resulted of new users joining or modifications of pattern for present users in model-based techniques. The author presented a new technique to produce dynamic model from off-line model produced by fuzzy clustering. In this technique, users' transactions are used periodically for modifying the off-line model. To this intend, an enhanced technique of leader clustering along with a static technique is used to regenerate clusters in an incremental fashion.

[6] Personalized Web page recommendation is strictly restricted by the nature of web logs, the intrinsic complexity of the problem and the higher efficiency needs. When handled by existing Web usage mining methods, because of the existence of an large number of meaningful clusters and profiles for visitors of a usually highly rated Website, the model-based or distance-based techniques are likely to create very strong and simple assumptions or, on the other hand, to turn out to be highly complex and slow. The author designed a heuristic majority intelligence trchnique, which effortlessly adjusts to changing navigational patterns; with the low cost explicitly individuate them ahead of navigation. The proposed technique imitates human behavior in an unidentified environment in occurrence of several individuals working in parallel and it has the ability to predict with better accuracy and in real time the next page group visited by a user. This technique has been checked on real data from users who browse a popular Website of common content. Average accuracy on test sets is better on a 17 class problem and, most importantly, it continues to be steady as the Web navigation goes on.

Rough set based feature selection for web usage mining is proposed by Inbarani et al., [7].

Web usage mining utilizes data mining methods to find out precious data from navigation pattern of World Wide Web (WWW) users. The necessary data is collected by Web servers and stored in Web usage data logs. The initial stage of Web usage mining is the pre processing stage. In the preprocessing stage, initially the related data is filtered from the logs. Data preprocessing is a important process in Web usage mining. The outcome of data preprocessing is related to the next stages like transaction identification, path analysis, association rule mining, sequential pattern mining, etc. Feature selection is a preprocessing stage in data mining, and it is highly helpful in decreasing dimensions, minimizing the unrelated information, enhancing the learning accuracy and enhancing comprehensiveness. The author presents a new technique for feature selection based on rough set theory for Web usage mining.

Jalali et al., [8] put forth a web usage mining technique based on LCS algorithm for online predicting recommendation systems. The Internet is one of the quickly developing fields of intelligence gathering. Through their navigation Web users provide several records of their action. This vast quantity of information can be a helpful resource of knowledge. Advanced mining techniques are required for this information to be extracted, understood and used. Web Usage Mining (WUM) scheme is particularly proposed to perform this process by examining the data indicating usage data concerning a particular Web site. Web usage mining can model user behavior and, hence, to predict their upcoming movements. Online prediction is a kind of Web Usage Mining application. But, the accuracy of the prediction and classification in the present architecture of predicting users' future requests systems can not still assure users particularly in vast Web sites. For offering online prediction effectively, the author provides advance design for online predicting in Web Usage Mining system and presents a new technique based on LCS algorithm for categorizing user navigation behavior for forecasting users' future requests. The simulation result indicates that the technique can enhance accuracy of classification in the system.

For providing the online prediction effectively, Shinde et al., [9] provides a architecture for online recommendation for predicting in Web Usage Mining System .The author presents the architecture of on line recommendation in Web usage mining (OLRWMS) for improving the accuracy of classification by interaction between classifications, evaluation, and present user activates and user profile in online phase of this architecture.

Zhang et al., [10] given an intelligent algorithm of data pre-processing in Web usage mining. Web usage mining is the kind of data mining methods to Web usage logs of huge Web data stores for producing results used in some parts like Web site design, Web server design, users categorization, creating adaptive Web sites and Web site personalization. Data preprocessing is a important process in Web usage mining. The outcome of data preprocessing are appropriate to the next process like transaction identification, path analysis, association rules mining, sequential patterns mining, etc. An algorithm called "USIA" was proposed and its merits and demerits were examined. The experimental evaluation of USIA indicates the better efficiency and also it determines the exact user and session.

Nasraoui et al., [11] provides a whole framework and findings in mining Web usage navigation from Web log files of a genuine Web site which has every challenging characteristics of real-life Web usage mining, together with evolving user profiles and external data describing an ontology of the Web content. Although the Web site considered is element of a nonprofit organization that does not sell any products, it was essential to recognize who the users were, what they looked at, and how their attentions modified with time, every one of which are significant questions in Customer Relationship Management (CRM). Therefore the author provides a technique for identifying and tracing growing user profiles. The author also illustrates how the discovered user profiles can be enriched with explicit data need which is gathered from search queries extracted from Web log data. Profiles are also enhanced with other domain-specific data features that provide a panoramic view of the discovered mass usage modes. An objective validation approach is also applied to evaluate the excellence of the mined profiles, in specific their adaptability in the face of evolving user pattern.

Temporal Web usage mining includes application of data mining process on temporal Web usage information for discovering temporal navigation that illustrates the temporal activities of Web users. Clusters and associations in Web usage mining do not essentially have crispy boundaries. Hogo et al., [12] proposed the temporal Web usage mining of Web users on single educational Web site with the help of the adapted Kohonen SOM based on rough set properties.

Along with the raped development of ecommerce, it is very essential to recognize the user access mode. By means of Web usage mining, the server log, registration data and other related data obtained from user access can be mined with the user access mode that will afford basis for decision-making of organizations. DeMin et al., [13] presented a SQL Server2000 based Web usage mining system and demonstrates with genuine instances on how to make use of DTS, T-SQL and other tools under SQL Server2000 to appreciate data transfer, data cleansing, user recognition, session recognition and further data pre-processing purpose, and demonstrates on how to utilize Online Analysis and Process (OLAP) and also the Data Mining (DM) under SQL Server2000 as to recognize mode discovery and mode analysis.

A development of data preprocessing technique for Web usage mining and the information of algorithm for path completion are provided by Yan et al., [14]. After user session detection, the missing pages in user access paths are added with the help of referrer-based technique that is an effective solution to the problems occurred when proxy servers and local caching is used. The reference length of pages in whole path is changed by considering the average reference length of auxiliary pages that is predicted in advance with the help of maximal forward references and the reference length techniques. As confirmed by practical Web access log, the presented path completion technique efficiently adds the missing data and enhances the reliability of access data for more Web usage mining evaluations.

Sophisticated data mining techniques are required for the knowledge to be extracted, understood and used. Baraglia et al., [15] proposed a Web usage mining (WUM) system, called SUGGEST, which continuously creates the suggested connections to Web pages of probable importance for a user. SUGGEST is proposed to effectively combines the WUM process with the regular Web server functionalities. It can afford valuable data to make the user's Web navigation simpler and to enhance the Web server's performance.

Web usage mining is the application of data mining methods to huge Web log databases for the purpose of extracting the usage navigation. Conversely, the majority of the earlier studies on usage patterns discovery only

focus on extracting intra-transaction associations, i.e., the associations between items inside the identical user transaction. A cross-transaction association rule illustrates the association link between various user transactions. Jian et al., [16] uses the closure property of frequent item sets for the purpose of extracting the cross-transaction association rules from Web log databases. The approach and technical framework based on this technique is proposed and examined.

A World Wide Web usage mining and examination tool called SpeedTracer, was created by Wu et al., [17] in order to realize user browsing pattern by investigating the Web server log files with data mining procedures. As the attractiveness of the Web has exploded, there is a powerful need to recognize user browsing pattern. Conversely, it is complex to carry out user-oriented data mining and analysis straightforwardly on the server log files since they inclined to be vague and deficient. With innovative technique, SpeedTracer initially recognize user sessions by rebuilding the user traversal paths. It does not need cookies or user registration for the purpose of session identification. User privacy is protected. Once user sessions are recognized, data mining techniques are then used to determine the very common traversal paths and groups of pages regularly visited simultaneously. The essential user navigation patterns are identified from the frequent traversal paths and page groups, assisting the understanding of user browsing pattern. Three kinds of reports are organized: user-based reports, path-based reports and groupbased reports. The author illustrates the design of SpeedTracer and shows some of its features with a little sample reports.

Jalali et al., [18] proposed a novel technique for web usage mining and visualization that is based on the bio-mimetic relational clustering technique Leader Ant and the description of prototypes based on typicality computation for producing an efficient visualization of the activity of users on a website. The simulation result illustrates that it can effortlessly construct meaningful visualizations of typical user browsing patterns.

Lee et al., [19] put forth a Web Usage Mining technique based on clustering of browsing characteristics. Guessing of user's navigation pattern is a significant technique in E-commerce application. The forecasted result can be helpful for personalization, creating appropriate Web site, enhancing marketing strategy, promotion, product supply, getting marketing data, predicting market trends, and enhancing the competitive power of enterprises etc. The author makes usage of the hierarchical agglomerative clustering to cluster users' navigation patterns. The prediction results by two levels of prediction model framework work better in general aspects. Conversely, two levels of prediction model experience from the heterogeneity user's navigation pattern. The author enhances the two levels of prediction technique to attain higher hit ratio.

Tzekou et al., [20] given an effective site customization technique based on Web Semantics and Usage Mining. The increased expansion of online information and the variety of goals that may be practiced over the Web have considerably enlarged the monetary value of the Web traffic. To knock into this faster growing market, Web site operators attempt to amplify their traffic by modifying their sites to the requirements of particular users. Web site personalization includes two big challenges: the efficient recognition of the user importance and the encapsulation of those importances into the sites' design and content. The author analyzed the techniques for efficiently identifying the user likings that are secreted behind user browsing behavior and new recommendation technique is proposed that utilizes Web mining techniques for correlating the recognized importance to the sites semantic content, for the purpose of modifying them to certain users. The simulation result indicates that the user interests can be exactly identified from their browsing pattern and that the recommendation technique that uses the identified interests and provides better enhancement in the sites' usability.

Xidong et al., [21] provides a technique that can discover users' frequent browsing patterns underlying users' browsing Web behaviors. Initially, the author proposes the technique of access pattern based on the user's access path. Next, the author presents a revised algorithm (FAP-Mining) based on the FP-tree technique for mining the common access patterns. The novel technique initially builds a common navigation pattern tree and then extracts the users' common access patterns on the tree. This technique is accurate and scalable for extracting the common browsing patterns with various lengths.

Content adaptation on the Web decreases the existing data to a subset that contests a user's anticipated requirements. Recommender techniques based on relevance scores for individual content items; particularly, pattern-based recommendation uses cooccurrences of items in user sessions to view any prediction about relevancy. To improve the discovered patterns' quality, Adda et al., [22] presents a technique with the help of metadata about the content that they imagine is stored in domain ontology. This technique includes a dedicated pattern space constructed on top of the ontology, navigation primitives, mining procedure and recommendation methods.

4 III.

5 Conclusion

The web is a very essential means to carry out business and commerce. So the design of web pages is highly essential for the system managers and web creators. These characteristics have huge impact on the number of users who access the page. Therefore the web analyzer has to examine with the data of server log file for identifying the navigation pattern. So it is essential for good understanding of the data preparation technique and pattern discovery method. Web usage mining systems will offer those techniques stated. This lead to support many researchers to provide their ideas in this field. There are several techniques proposed by different researchers for

the web usage mining with its own merits and demerits. This paper discussed about various techniques available for web usage mining.

The evaluation of clustering performance can be carried out in an integrated manner for problems like link prediction and customer relationship management. It is obvious that enhanced cluster recovery provides highly accurate guessing of a Web user's future visit if the user's cluster can be exactly determined. [1] [2] [3]

[1] March 2011 2 ©2011 Global Journals Inc. (US)
[2] March 2011©2011 Global Journals Inc. (US)
[3] March 2011

[Shinde and Kulkarni ()] 'A New Approach for on Line Recommender System in Web Usage Mining'. S K Shinde , U V Kulkarni . *International Conference on Advanced Computer Theory and Engineering*, 2008. p. .

[Labroche et al. ()] 'A New Web Usage Mining and Visualization Tool'. N Labroche , M J Lesot , L Yaffi . *19th IEEE International Conference on Tools with Artificial Intelligence*, 2007. 1 p. .

[Jalali et al. ()] 'A Web Usage Mining Approach Based on LCS Algorithm in Online Predicting Recommendation Systems'. M Jalali , N Mustapha , N B Sulaiman , A Mamat . *12th International Conference Information Visualisation*, 2008. p. .

[Nasraoui et al. ()] 'A Web Usage Mining Framework for Mining Evolving User Profiles in Dynamic Web Sites'. O Nasraoui , M Soliman , E Saka , A Badia , R Germain . *IEEE Transactions on Knowledge and Data Engineering* 2008. 20 (2) p. .

[Maratea and Petrosino ()] 'An Heuristic Approach to Page Recommendation in Web Usage Mining'. A Maratea , A Petrosino . *Ninth International Conference on Intelligent Systems Design and Applications*, 2009. p. .

[Huiying and Wei ()] 'An intelligent algorithm of data pre-processing in Web usage mining'. Zhang Huiying , Liang Wei . *Fifth World Congress on Intelligent Control and Automation*, 2004. 4 p. .

[Chen et al. ()] 'Discovering Web usage patterns by mining cross-transaction association rules'. Jian Chen , Jian Yin , A K H Tung , Bin Liu . *International Conference on Machine Learning and Cybernetics*, 2004. 5 p. .

[Wang et al. ()] 'Discovery of user frequent access patterns on Web usage mining'. Xidong Wang , Yiming Ouyang , Xuegang Hu , Yan Zhang . *The 8th International Conference on Computer Supported Cooperative Work in Design*, 2004. 1 p. .

[Tzekou et al. ()] 'Effective Site Customization Based on Web Semantics and Usage Mining'. P Tzekou , S Stamou , L Kozanidis , N Zotos . *Third International IEEE Conference on Signal-Image Technologies and Internet-Based System*, 2007. p. .

[Dong ()] 'Exploration on Web Usage Mining and its Application'. Demin Dong . *International Workshop on Intelligent Systems and Applications*, 2009. p. .

[Nina et al. ()] 'Pattern Discovery of Web Usage Mining'. S P Nina , M Rahman , K I Bhuiyan , K Ahmed . *International Conference on Computer Technology and Development*, 2009. 1 p. .

[Li et al. ()] 'Research on Path Completion Technique in Web Usage Mining'. Yan Li , Boqin Feng , Qinjiao Mao . *International Symposium on Computer Science and Computational Technology* 2008. 1 p. .

[Inbarani et al. ()] 'Rough Set Based Feature Selection for Web Usage Mining'. H H Inbarani , K Thangavel , A Pethalakshmi . *International Conference on Conference on Computational Intelligence and Multimedia Applications*, 2007. 1 p. .

[Wu et al. ()] 'SpeedTracer: A Web usage mining and analysis tool'. K L Wu , P S Yu , A Ballman . *IBM Systems Journal* 1998. 37 (1) p. .

[Baraglia and Palmerini ()] 'SUGGEST: a Web usage mining system'. R Baraglia , P Palmerini . *International Conference on Information Technology: Coding and Computing*, 2002. p. .

[Hogo et al. ()] 'Temporal Web usage mining'. M Hogo , M Snorek , P Lingras . *International Conference on Web Intelligence*, 2003. p. .

[Adda et al. ()] 'Toward Recommendation Based on Ontology-Powered Web-Usage Mining'. M Adda , P Valtchev , R Missaoui , C Djeraba . *IEEE Internet Computing* 2007. 11 (4) p. .

[Aghabozorgi and Wah ()] 'Using Incremental Fuzzy Clustering to Web Usage Mining'. S R Aghabozorgi , T Y Wah . *International Conference of Soft Computing and Pattern Recognition*, 2009. p. .

[Lee and Fu ()] 'Web Usage Mining Based on Clustering of Browsing Features'. Chu-Hui Lee , Yu-Hsiang Fu . *Eighth International Conference on Intelligent Systems Design and Applications*, 2008. 1 p. .

[Zhang et al. ()] 'Web Usage Mining Based On Fuzzy Clustering in Identifying Target Group'. Jianxi Zhang , Peiying Zhao , Lin Shang , Lunsheng Wang . *International Colloquium on Computing, Communication, Control, and Management* 2009. 4 p. .

[Wu et al. ()] 'Web Usage Mining on the Sequences of Clicking Patterns in a Grid Computing Environment'. Chih-Hung Wu , Yen-Liang Wu , Yuan-Ming Chang , Ming-Hung Hung . *International Conference on Machine Learning and Cybernetics (ICMLC)*, 2010. 6 p. .

[Etminani et al. ()] 'Web Usage Mining: Discovery of the Users' Navigational Patterns Using SOM'. K Etminani , A R Delui , N R Yanehsari , M Rouhani . *First International Conference on Networked Digital Technologies*, 2009. p. .