A Survey on web usage mining in the field of E-commerce

K.Rajeshwari[#1], C.Vinothini[*2]

[#]*Computer Science Department, Dr..NGP Institute of Technology*
Coimbatore, India.

[1]kmraji2011@gmail.com

[2]vinucrazy56@gmail.com

***Abstract*— Today World Wide Web becomes an interactive tool for sharing the information. As the growth of web users, huge amount of information is stored in web environment. Extracting that heterogeneous information is a tiresome process. Web mining exploits the data mining techniques to extract interesting patterns from web data. Three different types of web mining are usage mining, content mining, structure mining. Web usage mining is a salient research area in the field of web mining. Web usage mining consists of four phases namely collection of data, preliminary processing, discovery of access pattern and analysing the pattern. This paper purveys a detailed discussion about all the phases and related work in various fields.**

***Keyword*—Web Usage Mining (WUM), Pre-processing, Pattern discovery, Pattern analysis, E-commerce.**

I. INTRODUCTION

Today, internet is playing a vital role in everybody life, it is very difficult to survive without it. World Wide Web becomes a significant and it acts as a tool to collect, share and disseminate information at any place at any time. World Wide Web has influenced a lot to both users as well as the website owners. The website owners are able to reach to all the targeted audience nationally and internationally [1].Web data are complex in nature. Most of the web data is unstructured or semi structured. So we cannot apply data mining techniques directly to extract the information from weblog. Rather another discipline is evolved called web mining that can be applied to such web data.

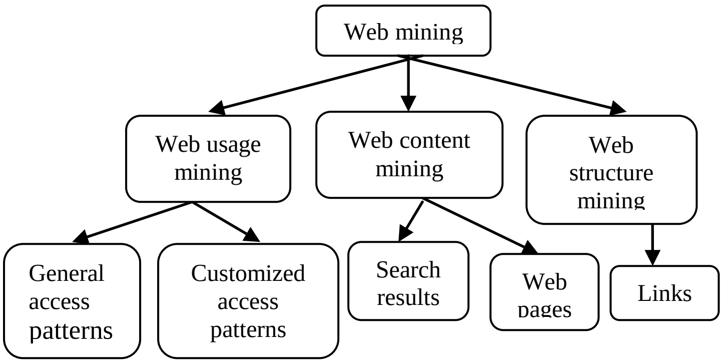

Fig.1 [4] Classification of Web Mining

Web mining makes use of various data mining techniques to automatically discover web and retrieve information from the web documents .But it is not sufficient to mine the web data, so we are using other techniques .There are three kinds of information presented in web: web content, web structure, web usage data, each one is exposed to a separate research area. Web usage mining is refers to the discovery of user access patterns from web usage logs. Web structure mining finds some useful knowledge from the structure of hyperlinks .Main goal of structure mining is to extract or mine useful knowledge from web structure. Web content mining extracts text, pictures and graphs of a web page to find out the relevance of the content to the search query. First is usage processing used to accomplish pattern discovery. It is most difficult to achieve the complete information because only bits of information such as IP address, time. Second is content processing consists of the alteration of web information like text, scripts, and images into some useful forms. Third is structure processing consists of analysis of the structure of each page hold in a website. With the explosion of e-commerce website the way companies are doing marketing has been changed.

There are millions of visitors interact daily with websites, massive amounts of data are being generated .Web information is very precious to the company in the field of e-commerce. By predicting access pattern of the visitor usage mining enhances the quality of e-commerce services and performances of web structure and web server. Section 2 describes about the overview of web usage mining .Section 3 describes the related work done on various fields.

II.WEB USAGE MINING OVERVIEW

While user interacts with the web, usage mining concentrates on the techniques that could predict the navigational pattern of user. Main goal of web usage mining is guessing the user interests for improving the usability of web. Web usage mining is the accumulation of web access information for web pages, this usage data purveys the paths leading to accessed web pages. This information can be collected automatically into access logs through web server. Companies make use of web usage mining to produce information pertaining to the future of their business function ability. From business perspective analysing web usage data has the ability to produce results more effective to their businesses and increasing of sales, use of this log data helps to gather the important information from customers visiting the site. Web usage mining is one of the salient research area due to the following reasons.

1. Web log information can be used for business intelligence in order to improve sales and advertisement by supplying product recommendation.
2. It can identify frequent access behaviour to improve the overall performance of future access.
3. To improve latency time caching and pre-fetching policies can be made based on frequent accessed pages.
4. To improve website design access behaviour of users can be used.
5. Personalization for a user can be obtained through web usage mining.

By keep tracking of accessed pages, we can predict the typical behaviour and desired pages.

There are four basic steps followed in web usage mining [1]: data collection, data pre-processing, and pattern discovery and pattern analysis.

A. Data collection

Data collection is the very first and significant step of web usage mining. Quality of data can be directly affecting the final recommendation of characteristic services. Data can be collected from various logs such as server logs, browser logs, and proxy logs or from organizational database.

B. Pre-processing

Because of enormous quantity of irrelevant information in web log, raw data cannot be used for further processing. Data pre-processing describes any kind of processing executed on raw data to organize it for another processing process .Data pre-processing converts the data which is convenient to user for further processing. Main purpose of this is to enhance the quality and competency of pre-processing steps of web usage mining. Various steps followed in pre-processing are: Data cleaning [10], Data filtering, and Data integration.

1) *Data Cleaning*: For any kind of web log analysis, data cleaning techniques can be used. Main goal of data cleaning is to remove irrelevant items in web log. This step provides removing useless requests from the log files. In web usage mining, this method eliminates requests regarding images and multimedia files .It also recognizes web robots and displaces their requests. By eliminating the useless data, we can downsize the log files to use less storage space and alleviate to further tasks.

2) Data filtering: Usually web data sources contains huge amount of data, but we need specific subset of data that meets certain conditions. For such kind of application data filtering technique can be used.

3) *Data integration:* The basic idea behind data integration is to organize the data from multiple sources.

C. Pattern Discovery

Web usage mining is the process of utilizing data mining techniques to the discovery of usage patterns from web data. WUM is used to expose patterns in server logs but executed only on samples of data. Various pattern discovery methods are [8]:

1. Statistical analysis
2. Clustering
3. Classification
4. Sequential patterns
5. Association rules

D. Pattern Analysis

This is the final step in web usage mining. To filter trifling information and obtain valuable information, the collected usage patterns are analysed. This process can be done after pre-processing step. But filtering the insignificant information and interpreting the interesting patterns is a major challenge in pattern analysis. For that we are using various techniques [6] like Structured Query Language (SQL) and Online Analytical Processing (OLAP).

II.RELATED WORK

This section provides a detailed discussion about WUM techniques in various fields. In first method C.J. Carmona, et. al., [3] proposed the methodology used in an e-commerce website of olive oil sale called www.OrOliveSur.com. They describe the set of phases including data collection, data pre-processing, extraction and analysis. Knowledge is extracted using supervised and unsupervised learning data mining algorithms through various tasks include clustering, classification, association and subgroup discovery. Results obtained will be discussed especially for the interests of the website, providing some guidelines for improving its usability and user satisfaction. Since users visit a very low number of pages and second the majority of visits use Internet Explorer, web master team must pay attention in visits obtained by reference websites.

In this second method Xuejun Zhang, et.al.,[11] described a toolset that make use of web usage mining techniques to identify customer Internet accessing patterns. These patterns are then used to prop up a personalized product recommendation system for online sales. To discover user group files and click stream data, they have used Self Organizing Maps (SOM) within the architecture. Based on that make a match to a specific user group and recommend a unique set of product browsing options appropriate to an individual user.SOM model execute better on this dataset: it leads to a small misclassification error based on mean absolute error (MAE) and a larger correlation coefficient. They have also shown that a personalized product recommendation system which enable SOM predictive model is able to produce useful recommendations.

Olatz Arbelaitz et.al, [9] proposed a web usage mining method for Bidasoa Turismo website. The tourism industry has experienced a switch from offline to online travellers and this has made the tourism web service essential. This service system must provide consumers and service providers with the most appropriate information, greater mobility and the most enjoyable travel experiences. They have presented the design of a general and non-protruding web mining system, using the information stored in a web server. The proposed system integrate both web usage and content mining techniques with the following main objectives: Generating user navigation profiles, enriching the profiles with semantic information to diversify them and attaining global and language-dependent user interest profiles for web designs, and allows them to create future marketing groups for particular targets. In this work they have used a crisp-based approach to allot interest profiles to clusters, selected a single topic as representative. They have also developed a general system which is also can applied to other websites.

Yu-Shiang Hung et.al, [12] provide a method for analyzing elder self-care behavior patterns. This study mainly focuses on analysing the daily self-care activities and health status of elders who live at home alone. They have collected log data from actual cases of elder self-care service. first they have analysed the self-care behaviour use cycle ,time, number of participants ,quality of services and then association rules were used to identify the relationship among those functions. Second, k-means algorithm is used to mine cluster patterns. Finally sequential profiles were captured for sequence behaviour mining cluster patterns for the elders. The results can be used in various research fields such as medicine, public health, nursing and psychology.This methodology uses association analysis, and sequence-based representation schemes in association with Markov models combined withART2-enhance K-mean algorithm.

Liao et.al, [7] proposed a system to implement online shopping and home delivery for hypermarkets. Internet population has increased globally due to the advancement in modern technology. Online shopping has become a new type of consumption. In addition, business-to-customer (B2C) home delivery markets have taken shape gradually, because virtual stores have risen and developed, e.g. mail-order, TV marketing E-commerce. This study combines online shopping and home delivery, and attempts to use association rules to keep track unknown bundling of fresh products and non-fresh products in a hypermarket. By using clustering analysis, the customers are divided up in clusters .Based on each of the cluster's consumption preferences catalogue is designed. By this method, to increase the catalogue's attraction to customers, hypermarkets are offered an online shopping and home delivery business model for sales services. With such a model, we can expect to attract more customers open up more broad markets, and earn the higher profits for hypermarkets.

Jae Kim et.al, [5] proposed a system for a personalized recommendation procedure for internet shopping support. To conquer the product overload of Internet shoppers, many recommender systems have been developed. Recommendation systems trail past actions of a cluster of customers to make a recommendation to individual members of the group. They introduce a personalized recommendation procedure by which they can get further recommendation effectiveness when they applied to Internet shopping malls. The suggested system is based on Web usage mining, product classification and decision tree induction. They implement the procedure to a leading Internet shopping mall in Korea for performance evaluation, and provide some experimental results. Finally the experimental results show that selecting the right level of product taxonomy and the right customers enhances the quality of recommendations. They have also compare the procedure with other methods such as collaborative filtering and rule-based.

C. J. Carmona et.al, [2] proposed system for subgroup discovery in a psychiatric emergency department by using evolutionary fuzzy rule extraction. They describe the application of evolutionary fuzzy systems for subgroup discovery to a medical problem. This study is based on the type of patients who incline to visit the psychiatric emergency department in a given period of time of the day. The main objective is to distinguish subgroups of patients according to their time of arrival at the emergency department. To resolve this problem, various subgroup discovery algorithms have been used to determine the better results among those algorithms. For this purpose, several subgroup discovery algorithms have been applied to determine which of them obtains better results: the classical CN2-SD algorithm, the evolutionary algorithm SDIGA and the multi-objective evolutionary algorithm MESDIF. Compared to various evolutionary algorithm, multi-objective evolutionary algorithm MESDIF provide a better results, so it has been used to mine interesting information respective to the rate of admission to the psychiatric emergency department.

III. CONCLUSION

This paper has endeavored to provide an up-to-date survey of the rapidly growing research area of Web Usage Mining. Since web-based applications specifically e-commerce has increased, analyzing web usage data is essential to better understand the web users. Doing such business analysis is eventually important in the field of business intelligence. It is very substantial for good understanding of the data preparation technique and pattern discovery method. Web usage mining system provides those techniques as stated. This guide to support many researchers to provide their ideas in this field.Web usage mining mainly focuses on the discovery of access patterns of Web users. For better understanding of user behaviour, web usage mining techniques has been used. Researchers proposed several techniques for the web usage mining in various fields. This paper mainly concentrates on techniques available for web usage mining in various fields and discusses about four different phases as stated .In this paper web usage mining technique has been used in various fields such as medicine, e-commerce and health care industry.

IV. References

[1] Arun Singh, Avinav Pathak, Dheeraj Sharma. Web Usage Mining: Discovery of Mined Data Patterns and their Applications. International Journal of Computer Science and Management Research Volume 2 Issue 5 May 2013 ISSN 2278-733X.

[2]Carmona C, González, P Del Jesus, M. J., & Herrera, F. (2010). NMEEF-SD: Non-dominated multi-objective evolutionary algorithm for extracting fuzzy rules in subgroup discovery. IEEE Transactions on Fuzzy Systems, 18, 958–970.

[3] C.J. Carmona , S.Ramírez -Gallego , F. Torres , E. Bernal, M.J. del Jesus , S. García(2012). Web usage mining to improve the design of an e-commerce website: OrOliveSur.com. Expert Systems with Applications 39 (2012) 11243–11249.

[4]http://www.infovis.net/imagenes/T1_N172_A1001_WebMiningEng.gif

[5]Jae Kyeong Kim, Yoon Ho Cho, Woo Ju Kim, Je Ran Kim, Ji Hae Suh. A personalized recommendation procedure for internet shopping support,Electronic Commerce Research and Applications 1 (2002) 301–313.

[6]Kamika Chaudhary, Santosh Kumar Gupta. Web Usage Mining Tools & Techniques: A Survey. International Journal of Scientific & Engineering Research, Volume 4, Issue 6, June-2013 1762 ISSN 2229-5518

[7]Liao S. H., Chen, Y, Lin, Y. T (2011).Mining customer knowledge to implement online shopping and home delivery for hypermarkets. Expert Systems with Applications, 38, 3982–3991.

[8]P.Nithya, Dr. P.Sumathi. A Survey on Web Usage Mining: Theory and Applications. International Journal, Computer Technology & Applications, Volume 3 (4), 1625-1629

[9]Olatz Arbelaitz , Ibai Gurrutxaga, Aizea Lojo, Javier Muguerza, Jesús Maria Pérez, Iñigo Perona(2013).Web usage and content mining to extract knowledge for modeling the users of the Bidasoa Turismo website and to adapt it. Expert Systems with Applications xxx (2013) xxx–xxx.

[10]S.K.Pani,,L.Panigrahy, V.H.Sankar,Bikram Keshari Ratha, A.K.Mandal,S.K.Padhi. Web Usage Mining: A Survey on Pattern Extraction from Web Logs. International Journal of Instrumentation, Control & Automation (IJICA), Volume 1, Issue 1, 2011.

[11]Xuejun Zhang, John Edwards , Jenny Harding(2007).Personalized online sales using web usage data mining.Computers in Industry 58 (2007) 772–782.

[12]Yu-Shiang Hung , Kuei-Ling B. Chen , Chi-Ta Yang, Guang-Feng Deng(2012).Web usage mining for analyzing elder self-care behavior patterns .Expert Systems with Applications 40 (2013) 775–783.