International Journal of Research in Advent Technology, Vol.2, No.3, March 2014
E-ISSN: 2321-9637

Survey on Latest Trends in Web Mining

D.Sridevi[1], Dr.A.Pandurangan[2], Dr.S.Gunasekaran[3]
[1]*Department of Computer Applications,,Valliammai Engineering College,*
[2]*Former Prof.Valliammai Engineering College,*
[3]*Prof&Head,Applied Research,GRU*
Email:ndsdevi@gmail.com

Abstract-The main purpose of this paper is to study the process of web mining techniques, features, and mainly on applications on e-commerce and e-business. Web mining widely used in e-businesses and the application runs more efficiently with the application of mining techniques such as data mining and usage mining. In this web mining is considered best.

Index Terms- Data Mining; Web Mining; Web Usage Mining;

1.INTRODUCTION

Web mining used to extract interesting and potentially useful patterns and hidden information from web documents and web activities .With the rapid increase of web information, search engine has gradually developed since 1994. In general, search engine is a web site that provides public information search services. It discovers network information by certain techniques and strategies on the Internet. It processes network information. And it provides users with search services in order of information navigation. Web Mining automatically extracts information from the web document and full fills all the needs of web user. Now a day's very useful for business people and by this e-commerce, the profit increased more than their targets.

2.CLASSIFICATION OF WEB MINING

Web mining is the application of data mining technology, which is to extract interesting and potentially useful patterns and hidden information from web documents and web activities [1, 2]. Web Mining is broadly categorized into Web content mining (WCM), Web structure mining (WSM), and Web usage mining (WUM) [2].

2.1 Web Content Mining: Web content mining is related to the uncovering of useful information from web contents, including text, image, audio, video, etc. Research in web content mining encompasses resource discovery from the web, document categorization and clustering, and information extraction from web pages.

2.2 Web structure mining: Web structure mining studies the web's hyperlink structure. It usually involves analysis of the in-links and out-links of a web page, and it has been used for search engine result ranking [3, 4].

2.3 Web usage mining: Web usage mining focuses on analyzing search logs or other activity logs to find interesting patterns.The Web mining classifications [5], as shown in figure1.

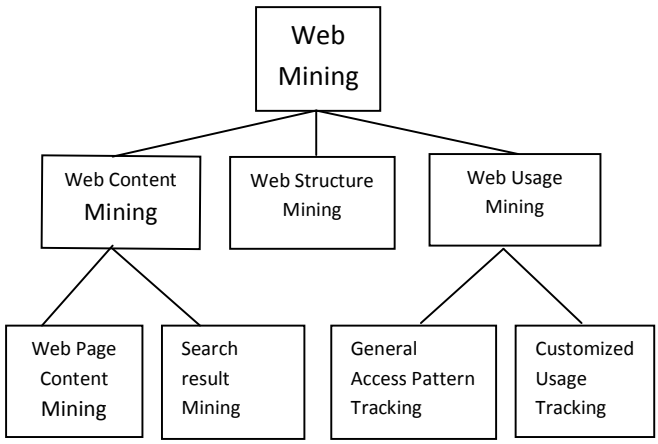

Fig1. Categories of Web Ming

3. PROCESS OF WEB MINING

Processes of Web mining are divided into four Stages: 1. Resources discovery 2.Information Choice and Preprocess, 3.Pattern Discovery, 4. Pattern Analysis [14]

3.1Resources Discovery

The task of retrieving intended web documents. The source of data in Web mining is the web log files and it records all the behavior of the user.

3.2 Information Choice and preprocess
Automatically selecting and pre-processing specific information from retrieved web resources. Here the data collected from the web may be incomplete, redundant and ambiguous. Preprocessing is done to accurate and concise data for mining. This process includes data cleaning, user identification, user session certifications, access path supplements and transaction identification.

3.3 Pattern Discovery
Automatically discovers general patterns at individual web sites as well as across multiple sites. Some of them are path analysis, association rule discovery, sequential pattern discovery, clustering analysis and classification.

3.4 Pattern Analysis
Validation and/ or interpretation of the mined patterns.. Its Purpose is to find out a valuable model. Few techniques used for analysis are visualization tool, OLAP techniques, data and knowledge querying and usability analysis.

4. WHY WEB USAGE MINING?
In this paper, we will emphasize on Web usage mining. Reasons are very simple: With the explosion of E-commerce, the way companies are doing businesses has been changed. E-commerce, mainly characterized by electronic transactions through Internet, has provided us a cost-efficient and effective way of doing business.

While web structure mining shows that page A has a link to Page B, web usage mining shows who or how many people took that link, which site they came from and where they went when they left page B. The important factors considered here are hyperlinks, dynamic content generation as per user references, quality of the content in web pages, huge size of the data.

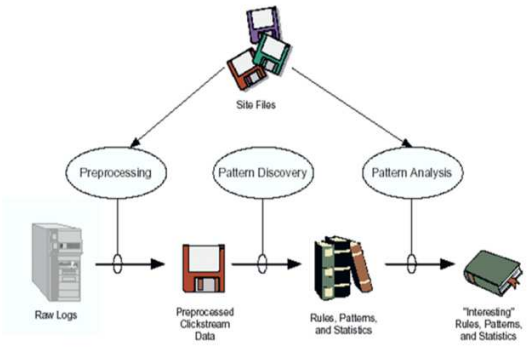

Fig 2: High Level Web Usage Mining

In preprocessing state user sessions are inferred from log data. In Pattern discovery, it searches for patterns in the data by making use of standard data mining techniques, such as association rules or mining for sequential patterns. In the pattern analysis stage information filter bases on domain knowledge and the web site structures is applied to the mining patterns in search for the interesting patterns. Links between pages and the similarity between contents of pages provide evidence that pages are related. The preprocessing phase allows the option of converting the server sessions into episodes prior to performing knowledge discovery.

4.1 Data Sources: web usage mining applications are based on data collected from three main sources [06]: (i) web servers, (ii) proxy servers, and (iii) web clients.

4.2 Data Preprocessing: Data preprocessing has a fundamental role in Web Usage Mining applications. The preprocessing of web logs is usually complex and time demanding. It comprises four different tasks: (i) the data cleaning, (ii) the identification and the reconstruction of users' sessions, (iii) the retrieving of information about page content and structure, and (iv) the data formatting.

4.3 Data Cleaning: This step consists of removing all the data tracked in web logs that are useless for mining purposes [7, 8].

4.4 Session Identification and Reconstruction: This step consists of (i) identifying the different users' sessions from the usually very poor information available in log files and (ii) reconstructing the users' navigation path within the identified sessions. Most of the problems encountered in this phase are caused by the caching performed either by proxy servers either by browsers. Proxy caching causes a single IP address (the one belonging to the proxy Server) to be associated with different users' sessions, so that it becomes impossible to use IP addresses as users identifies. This problem can be partially solved by the use of COOKIES [09], by URL rewriting, or by requiring the user to log in when entering the web site [10].
Content and Structure Retrieving: The vast majority of Web Usage Mining applications use the visited URLs as the main source of information for mining purposes. URLs are however a poor source of information since; for instance, they do not convey any information about the actual page content. [11] has been the first to employ content based information to enrich the web log data.

4.5 Data Formatting: This is the final step of preprocessing. Once the previous phases have been completed, data are properly formatted before applying mining techniques. [12] Stores data extracted from web logs into a relational database using a click fact schema, so as to provide better support to log querying finalized to frequent pattern mining.

5. WEB MINING IN E-COMMERCE

The tools of Web mining in e-commerce: The available tools of web mining in e-commerce are: 1) Statistic analysis. 2) Knowledge discovery 3) Prediction model.

The roles of Web mining: The roles of web mining are: 1) Optimize Web site, customized web site design. 2) Retain existing customers, and tap potential customers 3) Enhance e-commerce security. 4) Reduce operating costs and improve competitiveness of enterprises. [13].

6. BENEFITS OF WEB MINING

In the field of Financial Analyses that includes reviewing of costs and revenues, calculation and comparative analysis of corporate income statements, analysis of corporate balance sheet and profitability, cash flow statement, analysis of financial markets and sophisticated controlling. Web mining can be an effective tool.

In the field of Marketing Analyses that includes analysis of sales receipts, sales profitability, profit margins, meeting sales targets, time of orders, actions undertaken by competitors, stock exchange quotations, and market identification and segmentation. Web mining can be used here as a key tools that helps in building effective marketing strategy.

In the field of Customer Analysis that mainly concern time maintaining contacts with customers, customer profitability, modeling customers' behavior and reactions, customer satisfaction, churn analysis etc. web mining tells us what strategy should be used to get number of customers with quality.

In the field of Production Management Analysis where work is mainly to identify production 'bottlenecks' and delayed orders and enabling organizations to examine production dynamics and to compare production results obtained by departments or plants, etc.

In the field of Logistic Analysis, web mining can be effective to identify partners of supply chain quickly, reverse logistics analysis and handling.

In the field of Wage analysis, analysis of wage related data including wage component reports made with reference to the type required, reports made from the perspective of a given enterprise, wage report, distinguishing employment types, payroll surcharges, personal contribution reports, analyze of average wages, etc.

In the field of Personal data analyses that includes examination of employment turnover, employment types, presentation of information on individual employee's personal data, etc. [15][16] [17].

7. CONCLUSION

The advantages of using web mining in search engines and e-commerce, CRM, customer behavior analysis, cross selling; web site service quality improvement is noticeable. The recommendation of using web mining techniques can be applied successfully with a keen analysis of clearly understood business needs and requirements. Also one more governing factor is the amount of data, as the data is voluminous the results can be more towards the correct trends and patterns to be predicted from the given set of data. Web mining enhances users' ability to access information so that they feel very easy and comfortable to surf and thus the more applications have to be developed.

REFERENCES

[1] Dunham., Margaret H., Data Mining Introductory and Advanced Topics. Beijing: Tsinghua University Press, 2003, pI95-220.

[2] Han Jiawei and KamberMicheline Data Mining Concepts and Techniques [M].Beijing: China Machine Press, 2001, p290-297.

3] Wang Bin; Liu Zhijing, "Web mining research", Proceedings of fifth International Conference on Computational Intelligence and Multimedia Applications(ICCIMA 2003) , 2003 , pp: 84 - 89.

[4] Mei Li and Cheng Feng, Overview of WEB Mining Technology and Its Application in E-commerce, 2010 2nd International Conference on Computer Engineering and Technology ,Volume 7.DOI 978-1-4244-6349-7/10 .2010 IEEE pp V7-277-V7-280

[5] R. Cooley, B. Mobasher, J. Srivastava, "Web Mining:Information and Pattern Discovery on the World Wide Web," ictai, 9th International Conference on Tools with Artificial Intelligence (ICTAI '97), 1997, pp: 2-3.

[06] JaideepSrivastava, Robert Cooley, MukundDeshpande, and Pang-Ning Tan, Web usage mining: Discovery and applications of usage patterns from web data.SIGKDD Explorations, 1(2):12–23, 2000.

[7] Boris Diebold and Michael Kaufmann. Usage-based visualization of web localities.In Australian

symposium on Information visualisation, pages 159–164, 2001.
[8] Corin R. Anderson. A Machine Learning Approach to Web Personalization.PhD thesis, University of Washington, 2002.
[9] R. Cooley. Web Usage Mining: Discovery and Application of Interesting Patterns from Web Data. PhD thesis, University of Minnesota, 2000.
[10] Corin R. Anderson. A Machine Learning Approach to Web Personalization.PhD thesis, University of Washington, 2002.
[11] Robert Cooley, BamshadMobasher, and JaideepSrivastava.Data preparation for mining World Wide Web browsing patterns. Knowledge and Information Systems, 1(1):5–32, 1999.
[12] Jesper Andersen, Anders Giversen, Allan H. Jensen, Rune S. Larsen, Torben Bach Pedersen, and JanneSkyt.Analyzing clickstreams using subsessions. In Interna
tional Workshop on Data Warehousing and OLAP (DOLAP 2000), 2000.
[13] Open Market Inc. Open market web reporter. http://www.openmarket.com,1996.
[14]O.Etzioni. The World Wide Web: Quagmire or gold mine. Communications of the ACM, 30(11):65-68, 1996.
[15] Surat khan, Bin Zhang, Faizullah Khan, Siqi Chen," Business Intelligence in the Cloud: A Case of Pakistan", 2011, IEEE.
[16] Deepak Pareek,"Business Intelligence for Telecommunication", 2007.
[17] Celina M. Olszak and EwaZiemba ," Approach to Building and Implementing Business Intelligence Systems".