IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 5, Issue 2, February 2016

Web Usage Mining - A Review

Mr. Ankit Rathi[1], Prof. Abhijeet Raipurkar[2]

M. Tech, Computer Science and Engineering, RCOEM, Nagpur, India[1]

Assistant Professor, Computer Science and Engineering, RCOEM Nagpur, India[2]

Abstract: The web has recently become a powerful platform for retrieval of Information and discovering knowledge from web data. The beginning of discovering useful patterns in data has been given a variety of names like data mining, knowledge extraction, discovery of information and data pattern processing. Web mining is the application of data mining techniques for knowledge drawing out from web data. The data is collected from web server when the web accessed by user. First perform the preprocessing for finding access pattern because, raw data which is collected from the web server is incomplete. The aim of this paper is to understand the preprocessing of usage data and also the discovery of Patterns and their analysis. The discovery and analysis of patterns focuses on data accessed by the user. Web browsing behavior of users is captured by Web usage data through web site. User activities are stored in web logs. Due to more usage, the files in log are increasing at higher rate in size. The Preprocessing plays an important role in efficient mining process as Log data is normally noisy and not distinct.

Keywords: Web Usage Mining, Data preprocessing, Pattern discovery, Pattern analysis.

I. INTRODUCTION

The term Data mining is defined as the automatic extraction of unidentified, useful and understandable patterns from large databases. In order to increase the performance of Website, the essential thing is good web site design. The interests of the users help in designing better Websites. Web mining is used to retrieve, extract and evaluate data for information discovery from documents on Web. Web mining consists of Web content mining, Web structure mining and Web usage mining [3]. Web Content Mining deals with the discovery of information which is useful from the web data or documents. Web Structure Mining mines the hyperlinks structure within the web itself. The Structure represents the graph of the link in a website. Web Usage Mining mines data at log file stored in the web server.

II. WEB USAGE MINING

Web usage mining is the application of data mining techniques on large web log repositories to discover knowledge which is useful about behavioural pattern of user and also website usage statistics that can be used for various website design tasks. The four stages under web usage mining are:

Data Collection: The data in log is collected from sources like server side, client side and proxy servers and so on. Data Preprocessing: This is done on raw data which present in log file wrapping up of data cleaning, user identification and session identification. Pattern discovery: The patterns are discovered in this phase. Also the statistical analysis, association rules, clustering, pattern matching and so on perform in this. Pattern analysis: once patterns were discovered from web logs, the rules or patterns which are not interesting are filtered out. All the four stages are shown through the following figure –

1. Data Collection:
The data collection step includes various data sources.

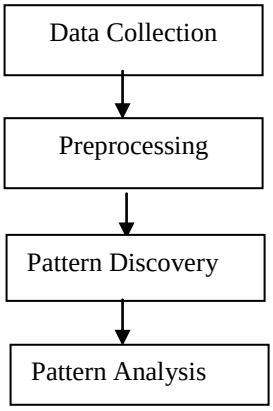

Figure 1: Web usage Mining Process

The Primary source of data in web usage mining is the log at server. There are some additional data source are also use for some user and some application which includes log on client side and Proxy side log [3]. In Log at client side, usage data can be tracked also on the client side. In many respects, collecting navigation data at the proxy level and at server level is same. The main difference is only that proxy servers collect data of user groups accessing big groups of web servers.

2. Data Preprocessing:
The information available in the web is Varied and unstructured. Therefore, the preprocessing phase is a required for discovering patterns. The purpose of this is to transform the raw data into a group of user profiles. Data preprocessing is important and this led to various algorithms and heuristic techniques for it such as Data Cleaning, User and Session Identification etc.

A. Data Cleaning: Data Cleaning is a process of removing items which are irrelevant such as jpeg, gif files or sound files. The improved data quality also improves the analysis on it. If a user request to view a particular page along with

server log entries the scripts and graphics are downloaded with an HTML file. Also Check the Status codes in log entries for successful codes.

B. User Identification: The identification of individual users who access a web site is an important step in web usage mining Process. Various methods are to be followed for this. The simplest method is to assign distinct user id to distinct IP addresses. If the user's IP address is same as previous entry and user agent is different, then the user is assumed as a new user. If the page that is requested is not directly reachable from any of the pages till visited by the user [4], then the user is identified as a new user in the same address.

C. Session Identification: The set of pages visited by the same user within the duration of one specific visit to a web-site is considered as a session of user. There are more than one session associated with same user also. The one method depends on time and another one on navigation in web topology used for identification of sessions. In Time Oriented Heuristic [1], there are two methods in which one method based on total session time and the other based on single page stay time. The set of pages which are visited by a user at a specific time is called page viewing time. The second method depends on stay time on page which is calculated with the difference between two timestamps. These methods are not reliable because users may involve in some other work after opening the web page. While in Navigation-Oriented Heuristic, the thing which is considered is webpage connectivity. If a web page is not connected with page which is opened previously in a session, then it is considered as a new session. Both the methods are used by many applications.

3. Pattern Discovery:
Once transactions of user have been identified, techniques of data mining are performed for pattern discovery in web usage mining process. These methods represent the ways that often appear in the data mining study such as discovery of association rules and sequential patterns and clustering and classification etc. Classification is a supervised learning process. In this, the data item mapped into one of several predefined classes. It can be done by using inductive learning algorithms such as naive Bayesian classifiers, decision tree classifiers, Support Vector Machines etc. Clustering is a technique of grouping users which exhibit similar browsing patterns. Such patterns are useful for inferring user count in order to perform market study in Ecommerce or provide personalized web content to web pages. By using this method, web marketers can predict future visit patterns which can help in placing advertisements aimed at certain user groups.

4. Pattern Analysis:
The last stage of web usage mining Process is Pattern Analysis. The patterns which are mined are not suitable for interpretations. So it is important to sort out patterns or rules which are not interesting from the set found in the pattern discovery phase. The tools are provided to help the transformation of information into knowledge in this phase. The exact analysis is governed by the application for which web mining is done. The SQL is the most common method of pattern analysis. While another method is to load usage data into a data cube in order to perform OLAP operations.

III. REVIEW OF LITERATURE

- **Paper by B. Uma Maheswari and Dr. P. Sumathi titled: "A New Clustering and Preprocessing for Web Log Mining".**

In this paper, the algorithms are proposed for data preprocessing and clustering. The main problem given is to get a dataset which is reliable for mining. Hence the data should be retreated and accessing behavior of users is to be constructed as transactions. The reliability of a transaction is an important. By using Cookies or authentication mechanism, users are identified. But users are not attracted by these types of sites due to privacy concerns. The two heuristics are mentioned for the acknowledgement of requests to different visitors. It has undergone various steps such as data cleaning, user identification, session identification and clustering. Also the transactions which show the user's behavior are constructed exactly in preprocessing step by calculating the Reference Lengths of user access by means of byte rate. Also by using Maximal Forward Reference and Reference Length algorithm Time Window concept is combined to find pages carries contents. By using Web clustering the objects of various types can be clustered into different groups for various purposes. Also by using the theory of distribution in Dempster-Shafer's theory [1], the belief function similarity measure in this algorithm adds to the task of clustering the ability to capture the uncertainty among Web user's navigation performance. In this Paper, the algorithm lacks in scalability problem.

- **Paper by V. Chitraa and Dr. Antony Selvdoss Davamani titled: "A Survey on Preprocessing Methods for Web Usage Data".**

In this Paper, it is given that for discovering patterns, the sessions are to be constructed efficiently. This paper also reviews existing work done in the preprocessing stage. The results of mining can be used to improve the design of the website and increase satisfaction which helps in various applications. The raw log files contain unnecessary details like images which are accessed, failure entries etc., which will affect the accuracy of pattern discovery Phase and analysis Phase. So the preprocessing stage is an important work in mining to make efficient pattern analysis. To get accurate mining results the session details of users are to be known. In this Paper, the survey was performed on a selection of web usage methods used in preprocessing proposed by research community. The preprocessing stages like session identification and path completion on which more concentration is done and this paper also have presented various works done by different researchers. In future scope it is given that to create more efficient session reconstructions with help of graphs and mining the sessions by using graph mining as quality sessions gives

more accurate patterns for user's analysis. Also in this paper, a brief overview of various data mining techniques for discovering patterns and pattern analysis is discussed. Finally a preview of various applications of web usage mining is also presented.

- **Paper by Theint Theint Aye titled: "Web Log Cleaning for Mining of Web Usage Patterns".**

In this paper, it is given that the most time consuming process is data preparation process. This paper presents two algorithms that is one for field extraction and other for data cleaning for data preprocessing. The entry in a log contains different fields which need to be separate out for the processing and this is done in field extraction algorithm. The process of dividing fields from the single line of the log file is known as field extraction. The server used various characters which work as separators. The most generally used separator character is 'or', 'space' character. Also the errors and inconsistencies will be detected and removed in data cleaning algorithm to improve the data quality. Hence the system given in this removes accesses to such items which are not relevant and failure requests in data cleaning. The advantage of using this proposed algorithm is that it speed up extraction time when interested information of user is retrieved and pages accessed by the user is discovered from log data.

- **Paper by Shahnaz Parvin Nina, Md. Mahamudur Rahaman, Md. Khairul Islam Bhuiyan, Khandakar Entenam Unayes Ahmed titled: "Pattern Discovery of Web Usage Mining".**

In this paper, the authors to give a clear understanding process of the data preparation and discovery of patterns. This paper provides a clear idea about the pattern discovery in the web usage mining process. The various algorithms are proposed in this paper for data preprocessing like data Preparation, user identification and session identification. Then by applying pattern discovery method they find some result based mostly used OS, on mostly used browser, etc.

IV. CONCLUSION

We studied all these discussed papers which are related to Web Usage Mining. The Data Preprocessing is very necessary. Due to this all irrelevant entries in the dataset are deleted. And only remains the important data which is used for the next phases in the web usage mining process such as pattern discovery and pattern analysis. The Interesting patterns are sorted out in pattern discovery phase and analyzed for better use in last phase that is pattern analysis phase of web usage mining process.

REFERENCES

[1] Uma Maheswari, Dr. P.Sumathi, A New Clustering and Preprocessing for Web Log Mining, World Congress on Computingand Communication Technologies, IEEE, 2014

[2] Theint Theint Aye, Web Log Cleaning for Mining of Web Usage Patterns, International conference on Computer Research and development, 2011, IEEE, Volume 2.

[3] Navin Kumar Tyagi, A.K. Solanki, Sanjay Tyagi, An Algorithmic Approach to Data Preprocessing in Web Usage Mining, International Journal of Information Technology and Knowledge Management, Volume 2, No. 2, July-December 2010.

[4] Suresh R.M., Padmajavalli R, An Overview of Data Preprocessing in Data and Web usage Mining, IEEE, 2006.

[5] Chitraa, Dr. Antony Selvdoss Davamani, A Survey on Preprocessing Methods for Web Usage Data, (IJCSIS) International Journal of Computer Science and Information Security, Vol. 7, No. 3, 2010.

[6] Shahnaz Parvin Nina, Md. Mahamudur Rahaman, Md. Khairul IslamBhuiyan, Khandakar and Entenam Unayes Ahmed, Pattern Discovery of Web Usage Mining, International Conference on Computer Technology and Development, Volume 2, IEEE,2009.

BIOGRAPHY

Ankit Rathi was born on 5th June1991.He is currently pursuing Post graduation studies in the final year of Master of technology (2014-2016) in Computer Science and Engineering at Shri Ramdeobaba College of Engineering and Management, Nagpur under Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur, Maharashtra State, INDIA. He is graduated from Prof. Ram Meghe College of Engineering and Management, Amravati.