Techniques for Web Usage Mining

Janhavi Bhalerao, Ratna Kendhe, Lahar Mishra

Department of Computer Science, NMIMS University, Mumbai, India.

Abstract- Web mining merges two areas of research: the World Wide Web and data mining. Web mining is applying data mining methods to estimate patterns from the data present on the web. This helps in improving web based services. It also has various commercial uses in the areas of artificial intelligence, business support services, personalization of web services and so on .Effective user behavior patterns can be found by applying mining techniques like clustering and association rules to web log data. This paper provides the overall perspective of web mining. It mainly focuses on the application of various data mining techniques to web data to obtain patterns of web usage.

Index Terms- Association rules mining, Clustering, Web Mining, Web Usage mining.

I. INTRODUCTION

The Web has become one of the most extensive platforms for exchanging or retrieving information. As it has become easier to publish documents, there is a rise in the number of users and as the information grows, searching for relevant information is turning into a time-consuming operation. Web mining aims to extract and mine useful knowledge from the Web.

The coverage of Web information is very wide. Almost all types of data is present on the web and it may or may not be linked. This information changes constantly. It is vital to keep up with these changes since these changes are directly linked to various industries.

Web mining is the process of examining data sets collected from various sources methodically and in detail, in order interpret it to get useful information. These data sets may consist of web log data .Researchers have classified web mining into 3 types, namely , web structure, content and usage mining[1]. This classification is based on the type of data to be mined.

`

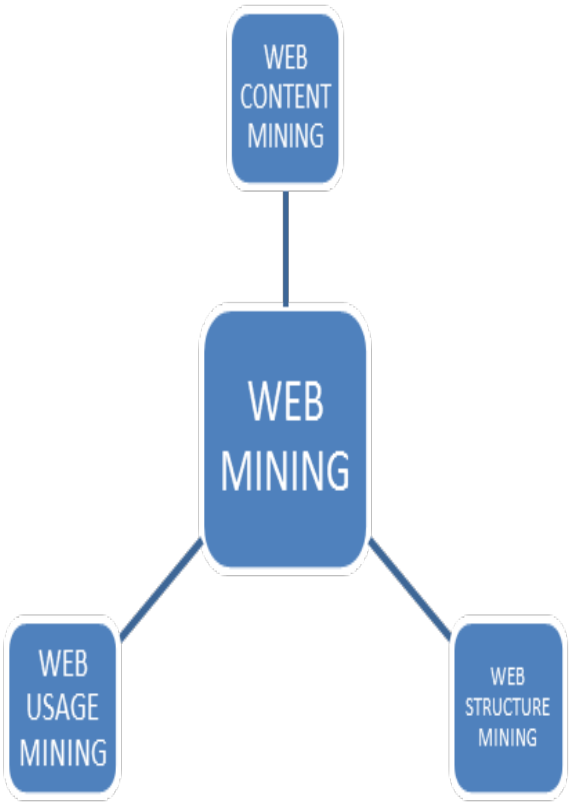

Figure 1: Classification of Web Mining.

II. TYPES OF WEB MINING

A. *Web content mining*

Web content mining is the process of obtaining useful patterns between data available on web like images, audio files and text. The natural language processing (NLP) and information retrieval (IR) are the technologies which are used to implement this.
Sometime, when the input data for mining is text based, the process is known as web text mining. Multimedia mining is the type in which information to be mined is from online multimedia resources.
Web content mining from two approaches: unstructured data and semi- structured data [2].The HTML(Hypertext Markup Language)or the hyperlink structures are used for the semi-structured data.

B. *Web structure mining*

Web structure mining is the method of using graphs and hyperlinks to obtain information of the website architecture and connection. Hyperlinks are the structures which link the web pages to a place.
There are two types of web structure mining: in the first type the patterns are discovered from the hyperlinks of the web pages, and in the second type the patterns are discovered from the document structure. Web structure mining gives the structure of a website and the relation between multiple web sites or web pages.

C. *Web usage mining*

Web usage mining is the process of estimating the user patterns of web browsing based on the web log data stored in client, proxy servers etc. The future behavior of the user on the web can be estimated using this kind of mining. In web usage mining, the mining techniques are applied on web log files. Web log files are files, which consist of data like who has visited the site, their location and their activity on the web site. They can be taken from servers or through proxy servers. It may also be present in the user's computer. There are various types of log files like: access log, error log, agent log and referrer log file.
The web usage mining process is divided into different stages. The first stage is the pre processing stage. In this, the data on which the mining techniques are applied is first cleaned and made ready for the further stages of usage mining.

III. STAGES OF WEB USAGE MINING PROCESS

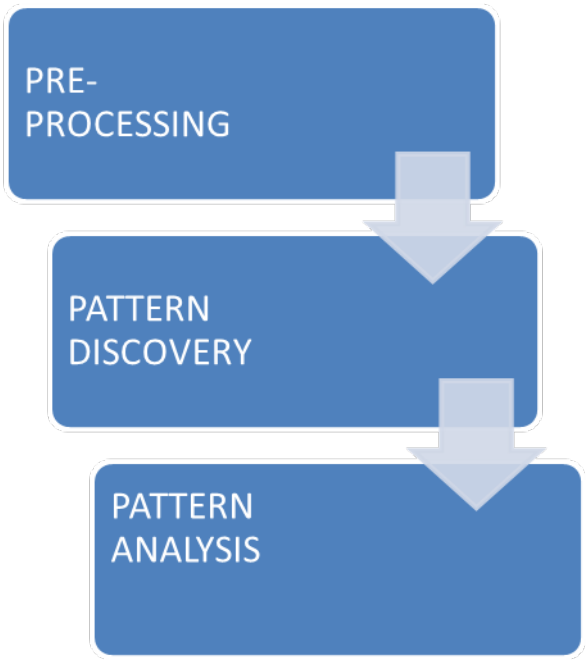

Figure 2:Stages of Web Usage Mining .

Web Usage Mining is composed of three different stages: Pre-processing, Pattern Discovery and Pattern Analysis [3].

A. *Pre-processing*

The data on which the web data mining algorithms are applied has to be made ready for it.The unprocessed raw data has to be converted to usable form so that data abstraction can be implemented. Data abstraction is the process of providing only the essential information and hiding the unnecessary background details. The data is collected from the client, server or proxy servers.
Pre-processing can be of usage pattern, content or structure.Data abstraction is implemented using the user identification algorithm and Data Cleansing of Web Log File algorithm. These algorithms take the web server log file as an input and give the log database as an output.
User and session identification is also done as a part of the process of cleaning of raw data. The aim of that is to categorize a user that accesses a web page and is likely to visit the web page multiple times into individual sessions. The last step of pre-processing is formatting. In this step the sessions obtained are formatted properly. This is followed by analyzing and discover of the patterns in the processed data.

B. Pattern Discovery

Pattern discovery is the key component of web usage mining. It merges the algorithms and techniques from data mining, machine learning and pattern recognition.
A variety of methods to find out hidden data information are implemented on Web server logs. Pattern discovery involves application of methods like association rules, clustering and sequential patterns.

- *Association Rule Mining Algorithm*
  - Association rules are used to find out the pages on web which are accessed together. This helps in foretelling which pages may be accessed by the user in future. The pages which are accessed together are put in a single server session. There is a specific support value based on which the web pages which are accessed together are found out. This happens if the support value is greater that the specific value.Association rules are useful in prefetching the data or web pages so that they are ready for later use. This reduces the time delay and latency. The websites can be restructured easily by using these rules from the access logs.
  - The setbacks of Association rule mining are that a lot of rules are produced and those rules may not be relevant. Based on parameters like minimum support and confidence, there may be incorrect estimations of these rules. This may lead to faulty results of pattern estimation. Clustering is done on the data to overcome this drawback. This leads to reduction of input data to be small for association rule mining algorithm. This avoids the rules produced to be incorrect or irrelevant.
  - If association rule mining algorithm is applied to clustered data, this algorithm will overcome the limit of the association rule mining algorithm. This limit is that irrelevant rules and wrong rule predictions may be generated on the basis of parameters like minimum support and confidence. The accuracy and results of the pattern discovery are improved using a combination of association rule mining and the clustering techniques.

- *Clustering Algorithm*
  - Clustering is the method by which similar data items are identified and grouped together. It helps in discovering, groups of users who have almost identical navigation patterns on the web and also patterns in sets of data. Different types of clustering are partitional Clustering, incremental clustering and hierarchical clustering.
  - It is useful for personalization of web sites for users, e-commerce, online marketing and various other commercial applications. It is also useful for search engines and web services since it can be utilized to get pages having almost identical content.
  - Complete Linkage algorithm is suggested for use, which is a combination of hierarchical complete linkage and incremental leaders algorithms. This composite algorithm enhances the quality of clustering.
  - Hybrid Leaders complete linkage algorithm (LCL) is used to categorize the users on the web into distinct clusters and also to cluster web data logs so that there is no need of reorganising the data objects. The input data is initially split into n clusters .This is followed by formation of a hierarchical structure on the basis of those n sub clusters.
  - The web log data is clustered and the user web usage pattern is estimated. Association rule mining is used to discover relation between sets of data and to interpret patterns in them. This rule is also used to obtain the web pages which are visited repeatedly. APRIORI algorithm is used to identify a particular user's navigation pattern on the web.

- *Kmeans Algorithm*
  - If Kmeans algorithm and vector matrix is implemented together it results in effective clustering of the web users. The first step before clustering process of the web data logs is the formation of the vector matrix consisting user and Uniform resource locater(URL).
  - The K means algorithm is a type of clustering algorithm. A data set is categorized into groups or clusters. Suppose, the number of clusters are assumed to be k in number. K number of centroids are defined for each cluster and they should be placed as far from each other as possible. This is followed by associating each point to its nearest centroid [4].The initial group centroids are kept into space which is depicted by the objects which

are currently being a part of clustering process. The group having the closest centroid gets an object. After this allocation, the centroid positions are recomputed.

- This results in a loop and k centroids change their position until no more changes are done. This is continued till centroids become stable. This results in formation of groups of objects. The K means algorithm is practical and has the ability to manage a growing amount of work in a capable manner .It is very easy to implement.

C. Pattern Analysis

Pattern Analysis is the last stage of web usage mining process. This is done after the output from the pattern discovery stage is obtained. In this stage, repeated association rules or patterns are eliminated and relevant and meaningful patterns are found. Techniques like Structured Query Language (SQL) and On-line Analytical Processing (OLAP) to make the output into suitable format.SQL is a knowledge query mechanism. OLAP is used to build a multi-dimensional data cube. For these methods, the output obtained at previous phase has to be structured.

IV. Conclusion

This paper explores the different techniques of web mining with emphasis on web usage mining. A detailed description of these methods and their advantages is given. The distinction between web mining types is also introduced .Overall the usage mining process is illustrated. Various combination of algorithms like association rule mining and clustering are suggested to produce effective results of discovering web usage patterns. Since the web continues to develop in size and complexity with time, it has become difficult to find information which is to the point. Hence, web data mining techniques are developed to get useful data from the web pages.

Acknowledgment

We would sincerely our professors for the constant mentorship. We are also grateful towards our peers for the encouragement and the constructive criticism.

References

[1]Pooja Sharma, Rupali Bhartiya," An efficient algorithm for improved web usage mining"
[2] R. Kosala, H. Blockeel,"Web mining research: A survey"
[3] Ilampiray.P," Efficient resource utilization of web using data clustering and association rule mining"
[4] JinHuaXu, HongLiu,"Web User Clustering Analysis based on KMeans Algorithm"
[5] Shaily G.Langhnoja, Mehul P. Barot, Darshak B. Mehta, "Web Usage Mining Using Association Rule Mining on Clustered Data for
Pattern Discovery"

AUTHOR

First Author- Janhavi Bhalerao

B tech Computer Science student (Currently pursuing),
Department of Computer Science,
NMIMS University, Mumbai, India.

Email id- janhavi.bhalerao@gmail.com

Second Author- Ratna Kendhe

B tech Computer Science student (Currently pursuing),
Department of Computer Science,
NMIMS University , Mumbai, India.

Email id- ratna.kendhe@gmail.com

Third Author- Lahar Mishra

B tech Computer Science student(Currently pursuing),
Department of Computer Science,
NMIMS University,Mumbai,India.

Email id- laharm@gmail.com