International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 3 Issue 10, October- 2014

Web Mining Classification: a Survey

S. Kalaichelvi
Student M.E. Second Year
Department Of Computer Science And Engineering
K.S.Rangasamy College Of Technology
Tiruchengode

Abstract:- **The World Wide Web is huge, unstructured, universal and heterogeneous. In recent years the growth of the World Wide Web exceeded all expectations. Today there are several billions of HTML and XML documents, pictures and other multimedia files available via internet and the number is still rising. But considering the impressive variety of the web, retrieving interesting content has become a very difficult task. Web usage mining is one of the technique of web mining is very useful to discover knowledge from secondary data obtained from the interaction from users with the web. The web usage mining is very essential for effective website. Web usage mining is mining of usage data captured through various logs stored on server, client or proxy. In this paper, tells basic idea about web usage mining.**

1. INTRODUCTION

1.1. WEB MINING

"Data mining is the process of analyzing data from different angles and summarizing it into useful information. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases."[1].

The term web mining was introduced by Etzioni in 1996 to denote the use of data mining techniques to automatically discover web documents and services, extract information from web resources, and uncover general patterns on the Web. Web mining is the application of data mining techniques to discover patterns or trends followed by the user from the Web [2] and extract knowledge from Web data, including Web documents, hyperlinks between documents, usage logs of web sites, etc. It is required as only small portion of information on web is relevant and giving user what he wants is important. It's again necessary to reduce time loss experienced by users while browsing for the required information.

Two different approaches are taken in initially defining Web mining. First is a 'process-centric view', which defined Web mining as a sequence of task. Second is a 'data-centric view', which defined Web mining in terms of the types of Web data that is being used in the mining process. To make use of the web in several ways. For example, finding relevant information, discovering new knowledge from the web, personalized web page synthesis, learning about individual users etc. Web mining techniques provides a set of techniques which provide solutions to different problems. However web mining techniques are not the only tools to handle these problems. Other related techniques from different research areas such as database (DB), information retrieval (IR) and natural language processing (NLP) can also be used. When we see web mining in terms of data mining it have three interest of operations say clustering (e.g. finding natural groupings of users, pages, etc.), associations (e.g. which URLs tend to be requested together) and sequential analysis (e.g. the order in which URLs tend to be accessed). As in most real world problems the clusters and associations in web mining do not have clear cut boundaries and often overlap considerably.

The web mining research relates to several research communities such a Database, Neural Networks, Information Retrieval and Artificial Intelligence. The most recognized approach is to categorize Web Mining into three areas. They are Web Content Mining, Web Structure Mining and Web Usage Mining. Web Content Mining focuses on the discovery/retrieval of the useful information from the web contents/data/documents. Web Structure Mining emphasizes to the discovery of how to model the underlying link structures of the web. Web Usage Mining is relative independent, but not isolated, category, which mainly describes the techniques that discover the user's usage pattern and try to predicate user's behaviors.

1.2. WHY WEB MINING

The requirement of web mining is used to store information on World Wide Web (WWW). The information is growing rapidly in the web so this gives what users want.

1.3. AREAS OF WEB MINING

There are three main thrust areas of web mining. Patterns followed by the users are evaluated by these three techniques of Web Mining and then these patterns are analyzed to get a user desired output. The taxonomy of web mining is depicted in figure 1.

1.3.1. WEB CONTENT MINING

Web Content Mining is the process of extracting useful information from the contents of Web documents. Content data corresponds to the collection of facts a Web page was designed to convey to the users. It may consist of text, images, audio, video, or structured records such as lists and tables. Application of text mining to Web content has been the most widely researched. Issues addressed in text mining are, topic discovery, extracting association patterns, clustering of web documents and classification of Web Pages. Research activities on this topic have drawn heavily on techniques developed in other disciplines such as Information Retrieval (IR) and Natural Language Processing (NLP). While there exists a significant body of

work in extracting knowledge from images in the fields of image processing and computer vision, the application of these techniques to Web content mining has been limited.
It studies the search and retrieval of information on the web. Web content mining future can be divided as web page content mining and search result mining. It has to do with the retrieval of content available on the web into more structure forms as well as its indexing for easy tracking information locations. Web content may be unstructured (plain text), semi-structure (HTML documents), or structured (extracted from databases into dynamic web pages). Such dynamic data cannot be indexed and consist what is called "the hidden web". A research area closely related to content mining is text mining.

1.3.2. WEB STRUCTURE MINING

Web structure mining is the process of using graph theory to analyze the node and connection structure of a website [3]. It focuses on the structure of the hyperlinks (inter document structure) within the web. The goal of web structure mining is to categorized the web pages and generate the information such as the similarity and relationship between them, taking the advantage of their hyperlink topology. Then it focuses on the identification of authorities. This can be further divided into two kinds based on the kind of structure information used.

A. Hyperlinks: A Hyperlink is a structural unit that connects a location in a Web page to different location, either within the same Web page or on a different Web page. A hyperlink that connects to a different part of the same page is called an Intra-Document Hyperlink, and a hyperlink that connects two different pages is called an Inter-Document Hyperlink.

B. Document Structure: In addition, the content within a Web page can also be organized in a tree-structured format, based on the various HTML and XML tags within the page. Mining efforts here have focused on automatically extracting document object model (DOM) structures out of documents

1.3.3. WEB USAGE MINING

Web Usage Mining is the application of data mining techniques to discover interesting usage patterns from Web data in order to understand and better serve the needs of Web-based applications [4]. Usage data captures the identity or origin of Web users along with their browsing behavior at a Web site. It discovers and analyzes user access patterns. The term web usage mining was introduced by Cooley et.al. in 1997 and in according with their definition: web usage mining is the automatic discovery of user access patterns from web servers. Web usage mining is the process of identifying browsing patterns by analyzing the user's navigational behavior. This information takes as input the usage data i.e. the data residing in the web server logs, recording the visits of the users to a web site. Web usage mining itself can be classified further depending on the kind of usage data considered:

A. Web Server Data: The user logs are collected by Web server. Typical data includes IP address, page reference and access time.

B. Application Server Data: Commercial application servers such as Weblogic, StoryServer have significant features to enable E-commerce applications to be built on top of them with little effort. A key feature is the ability to track various kinds of business events and log them in application server logs.

C. Application Level Data: New kinds of events can be defined in an application, and logging can be turned on for them generating histories of these specially defined events. It must be noted however that many end applications require a combination of one or more of the techniques applied in the above the categories.

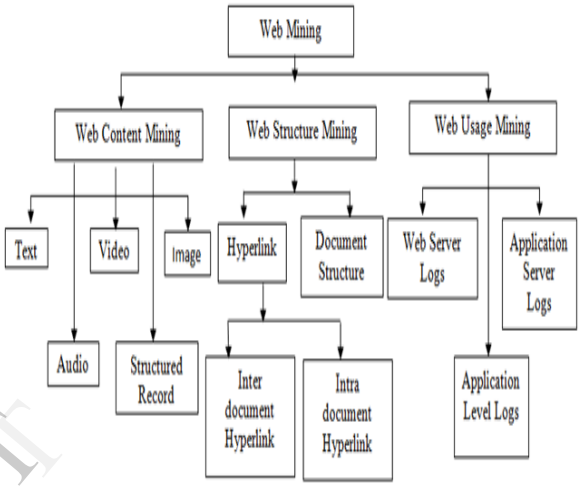

Figure 1. Taxonomy of web mining

2. STRUCTURE OF DATA IN WEB LOGS [5]

The log files are text files that can range in size from 1KB to 100MB, depending on the traffic at a given a website. The data will be taken for any particular website at given time. There are various fields in the log data which includes

- IP address: This is the IP address of the machine that contacted our site
- Username : This is the user that requested that website
- Timestamp: It is the timestamp of the visit
- Access request: It is the request made
- Result status code: This is whether URL was successfully returned or not. A number is saved stating whether request was successfully answered or not
- Bytes transferred: The number of bytes transferred after request was responded to by the server
- Referrer URL: This is the page referred by the user
- User agent: It is the software that the user is using to access the website. It is actually browser used by the user

3. PHASES TO PERFORM WEB USAGE MINING [6]

The phases to perform web usage mining are depicted in figure 2.

3.1 PREPROCESSING

It is a process of preparing data so that it can be used for Pattern Discovery and analysis. It includes Cleaning of Server Log files accompanied by identification of user sessions and user habits. It consists of

- Data field extraction
- Data Cleaning
- User identification
- Session identification

3.2 PATTERN DISCOVERY

After the data is preprocessed, this data is utilized for discovering homogeneous patterns [7]

3.2.1 STATISTICAL ANALYSIS

Statistical techniques are the most powerful tools in extracting knowledge about visitors to a Web site. The analysts may perform different kinds of descriptive statistical analyses based on different variables when analyzing the session file. By analyzing the statistical information contained in the periodic Web system report, the extracted report can be potentially useful for improving the system performance, enhancing the security of the system, facilitation the site modification task, and providing support for marketing decisions [8].

3.2.2 ASSOCIATION RULES

In the Web domain, the pages, which are most often referenced together, can be put in one single server session by applying the association rule generation. Association rule mining techniques can be used to discover unordered correlation between items found in a database of transactions [9]. The authors of [8] pointed that in the term of the Web usage mining, the association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold. The support is the percentage of the transactions that contain a given pattern. The Web designers can restructure their Web sites efficiently with the help of the presence or absence of the association rules. When loading a page from a remote site, association rules can be used as a trigger for pre-fetching documents to reduce user perceived latency.

3.2.3 CLUSTERING

Clustering analysis is a technique to group together users or data items (pages) with the similar characteristics. Clustering of user information or pages can facilitate the development and execution of future marketing strategies [9]. Clustering of users will help to discover the group of users, who have similar navigation pattern. It's very useful for inferring user demographics to perform market segmentation in E-commerce applications or provide personalized Web content to the individual users. The clustering of pages is useful for Internet search engines and Web service providers, since it can be used to discover the groups of pages having related content.

3.2.4 CLASSIFICATION

Classification is the technique to map a data item into one of several predefined classes. In the Web domain, Web master or marketer will have to use this technique if he/she want to establish a profile of users belonging to a particular class or category. This requires extraction and selection of features that best describe the properties of a given class or category. The classification can be done by using supervised inductive learning algorithms such as decision tree classifiers, naïve Bayesian classifiers, k-nearest neighbor classifier, Support Vector Machines etc [8].

3.2.5 SEQUENTIAL PATTERN

This technique intends to find the inter-session pattern, such that a set of the items follows the presence of another in a time-ordered set of sessions or episodes. It's very meaningful for the Web marketer to predict the future trend, which help to place advertisements aimed at certain user groups. Sequential patterns also include some other types of temporal analysis such as trend analysis, change point detection, or similarity analysis [8].

3.2.6 DEPENDENCY MODELING

The goal of this technique is to establish a model that is able to represent significant dependencies among the various variables in the Web domain. The modeling technique provides a theoretical framework for analyzing the behavior of users, and is potentially useful for predicting future Web resource consumption.

3.3 PATTERN ANALYSIS

Once the patterns are discovered then these patterns is evaluated and analysis is performed on these patterns and result generated is given to neural network for further processing. Pattern Analysis is a final stage of the whole Web usage mining. The goal of this process is to eliminate the irrelative rules or patterns and to extract the interesting rules or patterns from the output of the pattern discovery process. The output of Web mining algorithms is often not in the form suitable for direct human consumption, and thus need to be transform to a format can be assimilate easily. This can be done with the help of some analysis methodologies and tools. There are two most common approaches for the patter analysis. One is to use the knowledge query mechanism such as SQL, while another is to construct multi-dimensional data cube before perform OLAP operations [10]. All these methods assume the output of the previous phase has been structured. There are more techniques coming out in recent years, such as visualization etc. This is also a fertilized research area. Although there are quite a few commercial analysis applications available and many more are free on the Web, most of them are dislike by users, considered too slow, inflexible, difficult to maintain and limited in the functionality. To develop the efficient, flexible, and powerful tools, lots of work needs to be done for both researcher and developer.

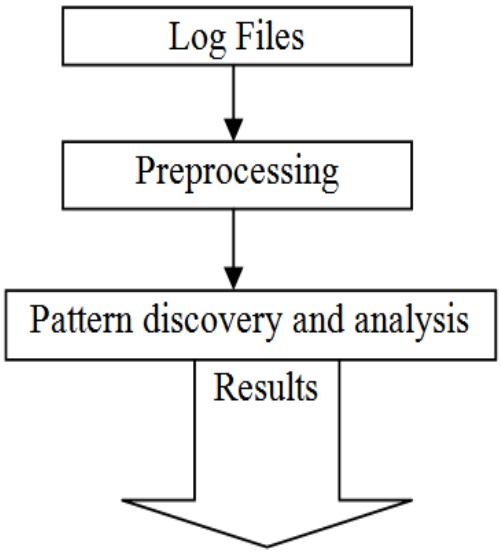

Figure 2. Web usage mining process.

4. WEB USAGE MINING ARCHITECTURE:

The WEBMINER is a system that implements parts of this general architecture [11, 12]. The architecture divides the Web usage mining process into two main parts. The first part includes the domain dependent processes of transforming the Web data into suitable transaction form. This includes preprocessing, transaction identification, and data integration components. The second part includes the largely domain independent application of generic data mining and pattern matching techniques (such as the discovery of association rule and sequential patterns) as part of the system's data mining engine. The overall architecture for the Web mining process is depicted in figure 3.

Data cleaning is the first step performed in the Web usage mining process. Some low level data integration tasks may also be performed at this stage, such as combining multiple logs, incorporating referrer logs, etc. After the data cleaning, the log entries must be partitioned into logical clusters using one or a series of transaction identification modules. The goal of transaction identification is to create meaningful clusters of references for each user. The task of identifying transactions is one of either dividing a large transaction into multiple smaller ones or merging small transactions into fewer larger ones. The input and output transaction formats match so that any number of modules to be combined in any order, as the data analyst sees fit. Once the domain-dependent data transformation phase is completed, the resulting transaction data must be formatted to conform to the data model of the appropriate data mining task. For instance, the format of the data for the association rule discovery task may be different than the format necessary for mining sequential patterns. Finally, a query mechanism will allow the user (analyst) to provide more control over the discovery process by specifying various constraints.

helps to produce applications that can more effectively and efficiently utilize the Web of knowledge for humankind.

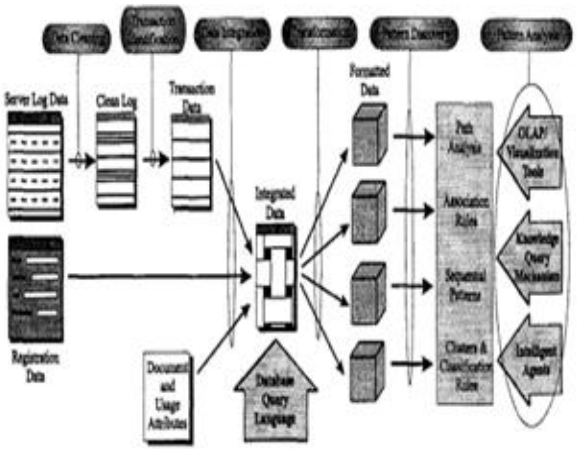

Figure 3. General Architecture for Web Usage Mining

5. PROBLEMS FACED WHILE PERFORMING WEB USAGE MINING [13]

- Processing of logs that is cleaning of log files
- Cleaning of log files that is removing data that is not relevant
- Identification of user sessions
- Identification of user habits

6. CONCLUSION

The Web has become the world's largest knowledge repository. Extracting knowledge from the Web efficiently and effectively is becoming increasingly important for a variety of reasons. The hidden Web, also known as the invisible Web or deep Web, has given rise to another issue facing Web mining research. The hidden Web refers to documents on the Web that are dynamic and not accessible by general search engines. Most documents in the hidden Web, including pages hidden behind search forms, specialized databases, and dynamically generated Web pages, are not accessible by general Web mining applications. However, without appropriate knowledge representation and knowledge discovery algorithms, it is just like a human being with extraordinary memory but no ability to think and reason. Hence believe that research in Web mining is promising as well as challenging and it will

REFERENCES

[1] http://www.anderson.ucla.edu/faculty/jason. frand/teacher/technologies/palace/datamining.htm
[2] http://en.wikipedia.org/wiki/Web_mining
[3] http://en.wikipedia.org/wiki/Web_mining/ web_ structure_mining
[4] http://en.wikipedia.org/wiki/Web_mining/ web_ usage_mining
[5] http://www.web-datamining.net/usage/
[6] Sonali Muddalwar Shashank Kawar (2012), "Applying artificial neural network in web usage mining", Vol 1 Issue 4, International Journal of Computer Science and Management.
[7] Anshuman Sharma (2012), "Web usage mining using neural network" International Journal of Reviews in Computing.
[8] R. Cooley. Web Usage Mining: Discovery and Application of Interesting Patterns from Web data. PhD thesis, Dept. of Computer Science, University of Minnesota, May 2000.
[9] R. Cooley, B. Mobasher, and J. Srivastava. Data preparation for mining World Wide Web browsing patterns. Knowledge and Information Systems, 1(1), 1999.
[10] O. Zaiane, M. Xin, J. Han. Discovering Web Access Patterns and Trends by applying OLAP and Data Mining Technology on Web Logs. In Advances in Digital Libraries, pages 19-29, Santa Barbara, CA, 1998.
[11] R. Cooley, B. Mobasher, and J. Srivastava. "Web mining: Information and pattern discovery on the world wide web". Technical Report TR 97-027, University of Minnesota, Dept. of Computer Science, Minneapolis, 1997.
[12] B. Mobasher, N. Jain, E. Han, and J. Srivastava. "Web mining: Pattern discovery from world wide web transactions". Technical Report TR 36-050, University of Minnesota, Dept. of Computer Science, Minneapolis, 1996.
[13] Ketki Muzumdar, Ravi Mante, Prashant Chatur, (2013)" Neural Network Approach for Web Usage Mining" Volume-2, Issue-2, International Journal of Recent Technology and Engineering (IJRTE).