International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395 -0056
Volume: 04 Issue: 04 | April -2017 www.irjet.net p-ISSN: 2395-0072

Recent Trends and Novel Approaches in Web Usage Mining

Sahaj Chavda,
Student, B.Tech in CE, Indus University, Ahmedabad

Saurabh Jain,
Student, B.Tech in CE, Indus University, Ahmedabad

Nikunj Panchal,
Student, B.Tech in CE, Indus University, Ahmedabad

Manisha Valera,
Assistant Professor, Dept. Of CE, Indus University, Ahmedabad

---

Abstract - *Web mining is an application of data mining which has become a significant area of research due to huge amount of World Wide Web services in recent years. Web Usage Mining is that area of Web Mining which deals with the extraction of interesting knowledge from sorting info produced by web servers. Web mining is an exciting discipline in the domain of data mining as well as in classification. Identifying the usage patterns of users is very vital in use the information available in the World Wide Web. This paper is a work on the future trends of web mining and trying to give a brief idea regarding web mining concerned with its techniques, tools and applications.*

Key Words: Data Mining, Web mining, Web Usage mining, web content mining, Data pre-processing, Web Structure Mining.

1. INTRODUCTION

The Web Mining is the set of techniques of Data Mining applied to extract some helpful knowledge and contained information from Web data. As more organizations be dependent on the Internet to conduct daily business, the study of Web mining techniques to get useful knowledge has become progressively important. Web mining enables one to discover Web pages, text documents, multimedia files, images and other types of resources from web. Web mining is an important area in data mining where we extract the interesting patterns from the contents. Web Mining consists of 3 processes namely Web Content Mining, Web structure mining and Web Usage Mining. Web content mining deals with the raw data that is available on the web. The web structure mining mainly deals with the structure of the web sites. Web Usage mining involve mining the usage characteristics of the users of web applications.

❖ Classification Of Web Mining Techniques

- Web Content Mining
- Web Structure Mining
- Web Usage Mining

1. Web Content Mining

Web Content Mining is that part of Web Mining which focuses on the raw information available in Web pages. Source data mainly consists of documented data in Web pages. Typical applications of web Content mining are content-based categorization and content-based ranking of Web pages. Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page content. Web content mining defines the discovery of useful information from the web contents. [2]Basically, the web content consists of several types of data such as textual, image, audio, video, metadata as well as hyperlinks. Web content mining data may be structured or unstructured even though such of web is unstructured. It is the process of retrieving the information from the web into more structured forms and indexing the information to recover quickly or finding valuable information from web content or web documents.

2. Web Structure Mining

The process of discovering structures information from the web documents are called as web structure mining. This mining can be performed either document level or hyperlink level. Structure mining or structured data mining is the process of finding and extracting useful information from semi-structured data sets. Graph mining, sequential pattern mining and fragment mining are special cases of structured data mining. Web structure mining is uses graph theory to analyze the node and connection structure of a web site. Web structure mining stabs to discover the model underlying the link structures of the web [2]. This model is based on the topology of the hyperlinks with or without the description of the links. According to the type of web structural data, web structure mining can be divided into two kinds:

- Extracting patterns from hyperlinks in the web: a hyperlink is a structural element that connects the web page to a different location.
- Mining the text structure: study of the tree-like structure of page structures to describe HTML or XML tag usage.

3. Web Usage Mining

Web usage mining is the process of extracting useful information from various web logs i.e. users history. 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. This is the process of finding out what users are Looking for on the Internet and usage of web pages. Web usage mining is used to discover the interesting usage patterns form the usage data. [2] This includes server data (IP address), Application server data (web logic), and Application level data (events). Usage data captures the identity or source of Web users along with their browsing behavior at a Web site. Web usage mining involves of three phases namely preprocessing, pattern discovery and pattern analysis. There are different techniques available for web usage mining with its own advantages and disadvantages.

Architecture of Web usage mining

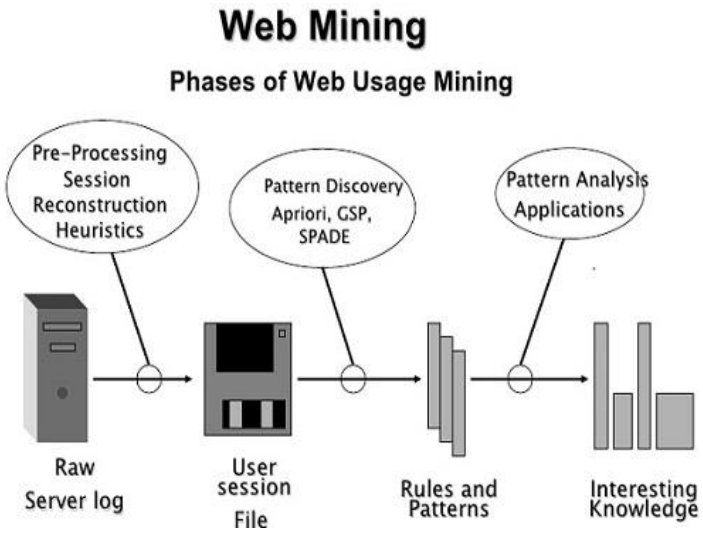

Fig -1: process of Web Usage Mining

1. Steps in Web Usage Mining

A. Pre-Processing
B. Pattern Discovery
C. Pattern Analysis

A. DATA PRE-PROCESSING

The main steps of web usage mining process are Data Pre-processing, Pattern discovery and Pattern analysis. Among them preprocessing is considered as a more intricate and time consuming process due to different nature of log data. It has been observed that preprocessing of log file takes more time than other phases of web usage mining process. Data preprocessing transforms data into a format that will be more easily, and efficiently processed for the purpose of the user. The key task of data preprocessing is to select identical data from the original log files, prepared for user navigation pattern discovery algorithm. The stage of data preprocessing includes data cleaning, user identification and session identification.

1. DATA CLEANING

Data cleaning is the process of detecting and correcting (or removing) corrupt or mistaken records from a record set, table, or database and refers to identifying incomplete, incorrect, wrong or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or rough data. [2] The first step of data preparation is data cleaning or filtering. It is very important as there have many unnecessary entries in the log files. The second step is the removal of graphical contents (audio, video, images) as they are downloaded with the requested page even if they are not clearly requested by the users .Graphical files are easily identified due to its file extension (jpg, gif etc.). The third step is to remove log entries created by web robots (sometimes referred as web spiders or web Crawlers). Robots are a special type of software which is used by various search engines to update its indexed pages by accessing pages of a particular website in a periodic time Interval.

2. USER IDENTIFICATION

User identification is to identify who accesses the website And which pages are accessed. [2]Once HTTP log files have been cleaned, next step in the data preprocessing is the identification of users. User identification is one of the difficult task due to existence of local/external proxy servers, cache systems, cooperate firewalls and shared internet. Different methods for this are:

1) By converting IP address to domain name.

2) The web server unsystematically assigns an ID to web Browser while it connects first time to the site.

This is called cookies. The Web browser sends the same ID back to the Web server, effectively telling the Web site that a specific user has returned. Cookies help the Web site developer to easily identifying individual visitors, which results in a greater understanding of how the site is used. IP address is logged into log file when a user hits a page. This address can be used to distinguish different users. But in case of proxy server when several users request a particular page then web site server logged same IP address (Proxy server IP) into the log file.

3. SESSION IDENTIFICATION

Once user is recognized there is need to identify sessions. Session is set of requests done by single user for defined Period to a particular web site. Session identification, which encodes the navigational behavior of the users, is very important in usage mining. A user session is a sequence of web pages that the user visits in a single website access. Session identification can be performed using time interval

between successive log entries. [2] If two entrances from the same user are separated by an interval longer than a threshold they considered as different session. Sometimes threshold considered as 30 minutes time interval. Another way to identify session is using a time out to identify the End of the session.

B. PATTERN DISCOVERY

After data preparation phase, the pattern discovery method Should be applied. In this stage, pre-treated information is analyzed to extract Valuable patterns. Statistical methods and machine learning are used to mine patterns. Pattern discovery deals with the sorted set of data items presented as part of a sequence [2]. Using this consecutive pattern mining, users can easily recognize the web paths that users commonly follow on a web site. The aim of this research work discovers the patterns which are most applicable and interesting by using a Web usage mining process .The web log files serves as an input to this process. Pattern discovery draws upon the methods and algorithms developed from the numerous fields such as data mining, pattern recognition, machine learning, and statistics. The task for learning the patterns offer some techniques as statistical analysis, association rules, sequential pattern analysis, clustering and so on.

STATISTICAL ANALYSIS

Statistical techniques are the most common method to extract knowledge about guests to a web site. By analyzing the session file, one can perform different kinds of descriptive statistical analyses (frequency, mean, median) on variables such as page views, viewing time and length of a navigational path. For example e-Trade developed a website in German language for Germany and scrapped it because German people were visiting the English site rather than the German site. Many web traffic analysis tools produce a periodic report containing statistical information such as the most frequently accessed pages, average view time of a page or average length of a path through a site. This type of knowledge can be potentially handy for refining the system performance, enhancing the security of the system, facilitating the site modification task, and providing support for marketing decisions. There are lots of commercial tools available for statistical analysis.

ASSOCIATIONS RULES

Association rule generation can be used to relate pages that are most often referenced together in a single server sessions [10]. In the context of web usage mining, association rules refer to sets of pages that are accessed together with a support value exceeding some specified threshold. Association rule mining has been well studied in Data Mining, especially for basket transaction data analysis. Many association rule algorithms have been used, such as Apriority, Partition [11]. Aside from being applicable for e-Commerce, business intelligence and marketing applications, it can help web designers to restructure their web site. The results about the usefulness of such rules in supermarket transaction or in web application have not been reported. People also put some constraints over the mining process, and prune the extracted rules. The association rules may also serve as heuristic for pre fetching documents in order to reduce user-perceived latency when loading a page from a remote site. In electronic CRM, an existing customer can be retained by dynamically creating web offers based on associations with threshold support and/or confidence value [12].

CLUSTERING

Clustering is a technique to group together a set of items having similar characteristics [10]. Clustering can be performed on either the users or the page views. Clustering analysis in web usage mining intends to find the cluster of user, page, or sessions from web log file, where each cluster represents a group of objects with common interesting or characteristic. User clustering is designed to find user groups that have common interests based on their behaviors, and it is critical for user community construction. Page clustering is the process of clustering pages according to the users' access over them. Such knowledge is especially useful for inferring user demographics in order to perform market segmentation in e-Commerce applications or provide personalized web content to the users. On the other hand, clustering of pages will discover groups of pages having related content. This information is useful for the Internet search engines and Web assistance providers. In both applications, permanent or dynamic HTML pages can be created that suggest related hyperlinks to the user according to the user's query or past history of information needs. The intuition is that if the probability of visiting page, given page has also been visited, and is high, then maybe they can be grouped into one cluster. For session clustering, all the sessions are processed to find some interesting session clusters. Each session cluster may be one interesting topic within the web site. Mobasher et al [13] generated recommendations from URL clusters to build an adaptive web site by using ARHP (Association Rule Hypergraph Partitioning).

C. PATTERN ANALYSIS

Pattern analysis is the last step in the overall Web Usage mining process as described in Figure 1. Pattern Analysis is to filter out unexciting rules or patterns from the patterns which discovered in the pattern discovery.The most common form of pattern analysis consists of knowledge query mechanism like SQL. After usage patterns are discovered, techniques and tools are needed to make these patterns explicable for analysts and to maximize the benefits from these patterns. Visualization techniques, such as graphing patterns or assigning colors to different values, can often

highlight overall patterns or trends in the data. Content and structure information can be used to filter out patterns containing pages of a certain usage type, content type, or pages that match a certain hyperlink structure. Techniques Include database querying, graphics and visualization, Statistics and usability analysis. The most used techniques Are:

Visualization:

This is a very effective method to help in understanding user behavior. Several tools have been developed to apply this method; for example, WebKIV, which is a tool that provides structural visualization for small and huge web structures, web navigation, which is for individuals and aggregate user navigation patterns, and result comparison.[8] Another method is to load usage data into a data cube in order to perform OLAP operations.

❖ APPLICATIONS FOR WEB MINING

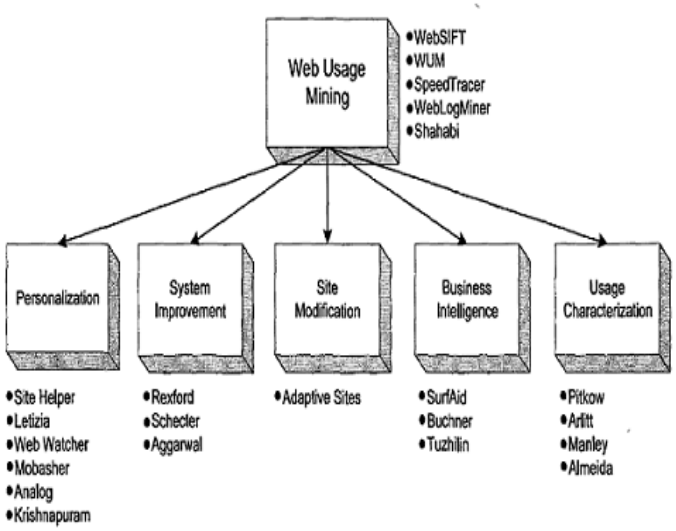

Figure 4: Major Application Areas for Web Usage Mining

The key impartial of web usage mining is to learning user navigation and their use of web resources. This information can be demoralized later to improve the Web site from the users viewpoint. There are various applications for web usage mining in dissimilar areas, and such applications are:

➢ Personalization

Personalizing the Web involvement for a user is the consecrated of many Web-based applications, e.g. individualized marketing for e-commerce [14]. Making dynamic recommendations to a Web user, based on her/his profile in addition to usage performance is very attractive to many applications, e.g. cross-sales and up-sales in e-commerce. Web usage mining is an excellent method for achieving this goal, as showed in [43] Existing recommendation systems, such as [15], do not currently use data mining for recommendations, though there have been some recent application. The Web Watcher, Site helper, Litizia and clustering work by Mobasher ET. al. [] yan ET. al. [] have all concentred on providing web site personalization based on usage information.

➢ System improvement

Performance and further service talents are crucial to user satisfaction from services such as databases, net- works, etc. Similar qualities are expected from the users of Web services. Web usage mining provides the key to under-standing Web traffic behavior, which can in chance be used for developing policies for Web caching, network transmission [17], load balancing, or data distribution. Security is an highly growing concern for Web-based services, especially as electronic commerce continues to grow at an exponential rate [18]. Web usage mining can also provide patterns which are useful for detecting intrusion, fraud, attempted break-ins, etc.

➢ Pre-fetching and caching

The outcomes produced by Web Usage Mining can be abused to improve the performance of Web servers and Web-based applications. [7] Typically, Web Usage Mining can be used to develop suitable pre-fetching and caching schemes so as to reduce the server response time.

➢ Security

The position of web application and its security growing day by day, but outdated networks fails to provide security for web application Web usage mining can provide patterns that are useful in Trouble, attempted break-ins, fraud, etc. [6]

➢ Site Modification

The charm of a Web site, in terms of both content and structure, is fundamental to many applications, e.g. a product catalog for e-commerce. Web usage mining provides detailed feedback on user behavior, providing the Web site designer information on which to base redesign choices. While the results of some of the projects could lead to re- designing the structure and content of a site, the adaptive Web site project (SCML algorithm) [19;20] focuses on repeatedly changing the structure of a site based on usage patterns discovered from server logs. Clustering of pages is used to determine which pages should be directly linked.

➢ Business Intelligence

Information on how customers are using a Web site is critical information for marketers of e-tailing businesses.

Buchner et al [21] have presented a knowledge discovery process in order to discover marketing intelligence from Web data. They define a Web log data hypercube that will associate Web usage data along with marketing data for e-commerce applications. They identified four different steps in customer relationship life cycle that can be supported by their knowledge discovery techniques: customer attraction, customer holding, cross sales and customer departure.

❖ TOOLS FOR WEB USAGE MINING

Many different tools are used to execute analysis on collected data, and most of them are based on statistical analysis techniques.[8] The number of commercial tools increased again last year and most of them are included in the Customer Relationship Management (CRM) software, which has solutions for e-commerce. Various tools used for web usage mining are Web Utilization Minor (WUM), Web Site Information Filter System (Web SIFT), KOINOTITES used for Web personalization etc.
Maintain, or give limited results.

❖ FUTURE TRENDS IN DATA MINING

Businesses which have been slow in accepting the process of data mining are now catching up with the others. Extracting important information through the process of data mining is broadly used to make critical business decisions. In the coming period, we can think data mining to become as ubiquitous as some of the more dominant technologies used today. Some of the key data mining trends for the future include like
Multimedia, Ubiquitous, Distributed and Spatial and Geographical Mining.

➢ MULTIMEDIA DATA MINING

This is one of the modern methods which is holding up because of the increasing ability to capture useful data accurately. It involves the extraction of data from different kinds of multimedia sources such as audio, text, hypertext, video, images, etc. and the data is converted into a numerical representation in different formats. This method can be used in clustering and classifications, performing similarity checks, and also to identify associations.

➢ UBIQUITOUS DATA MINING

This method involves the mining of data from mobile devices to get information about persons. In spite of having some challenges in this type such as difficulty, privacy, cost, etc. this method has a lot of opportunities to be massive in various industries especially in studying human-computer interactions.

➢ Distributed Data Mining

This type of data mining is gaining popularity as it involves the mining of huge amount of information stored in different company locations or at different organizations. Highly sophisticated algorithms are used to extract data from different locations and provide proper insights and reports based upon them.

➢ Spatial and Geographic Data Mining

This is new trending type of data mining which includes extracting information from environmental, astronomical, and geographical data which also includes images taken from outer space. This type of data mining can reveal various aspects such as distance and topology which is mainly used in geographic information systems and other navigation applications

• INFERENCE

The web is a most important medium to conduct business and commerce. Therefore the design of web pages is very important for the system administrator and web designers. These structures have great impact on the number of visitors. So the web analyzer has to analyze with the data of server log file for sensing pattern. In this paper we tried to give a clear understanding of the data preparation process and pattern discovery process. This paper has attempted to provide an up- to-date survey of the rapidly growing area
Of Web usage mining, which is the demand of present technology. In this paper a general outline of Web usage
Mining is presented in introduction section. There is a need to develop tools, which incorporate statistical methods, visualization, and human factors to help better understand the mined knowledge. One of the open issues in data mining is the creation of intelligent tools that can assist in the clarification of mined knowledge. This paper has analyzed the importance of web usage mining, the process of usage mining in detail and the techniques used in pattern discovery. Also, it has described the various applications for web usage mining and the tools that can be used.

REFERENCES

[1] Global Journal of Computer Science and Technology
J Vellingiri, S.Chenthur Pandean
Volume 11 Issue 4 Version 1.0 March 2011
Type: Double Blind Peer Reviewed International Research Journal Publisher: Global Journals Inc. (USA)
Online ISSN: 0975-4172 & Print ISSN: 0975-4350

[2] Mr. Akshay Upadhyay, Mr. Balram Purswani
International Journal of Scientific and Research Publications, Volume 3, Issue 2, February 2013
ISSN 2250-3153

[3] D. Jayalatchumy, Dr. P.Thambidurai
IOSR Journal of Computer Engineering (IOSR-JCE)
E-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 14, Issue 3 (Sep. - Oct. 2013), PP 20-27 www.iosrjournals.org

[4] Tulasi Gayatri Devi, Aparna KS
International Journal for Research in Applied Science & Engineering Technology (IJRASET)
Volume 4 Issue I, January 2016 ISSN: 2321-9653

[5] RACHIT ADHVARYU
JOURNAL OF INFORMATION, KNOWLEDGE AND RESEARCH IN COMPUTER ENGINEERING
ISSN: 0975 – 6760| NOV 12 TO OCT 13 | VOLUME – 02, ISSUE – 02

[6] S.Geetharani, S.Priyadharshini
IJETST- Vol.||02||Issue||03||Pages 1973-1975||March||ISSN 2348-9480

[7] ANITHA TALAKOKKULA
Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol.6, No.2, 2015

[8] M. Aldekhail, International Journal of Computer Theory and Engineering, Vol. 8, No. 1, February 2016

[9] Mitali Srivastava, Rakhi Garg, P. K. Mishra
International Journal of Computer Applications (0975 – 8887)
Volume 97– No.18, July 2014

[10] Jaydeep Srivastava, Robert Cooley, Mukund Deshpande, and Pang Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data (2000). SIGKDD Explorations, Vol. 1, Issue 2, 2000.

[11] Margaret H. Dunham, Data Mining Introductory and Advanced Topics, Prentice Hall, 2003.

[12] A.G. Buchner, M.D. Mulvenna, Discovering Internet Marketing Intelligence through Online Analytical Web Usage Mining, ACM SIGMOD, Vol. 27, No. 4, pp. 54-61, 1998.

[13] Bamshad Mobasher, Robert Cooley, Jaydeep Srivastava, Creating Adaptive Web Sites Through Usage-Based Clustering of URLs,in *Proceedings of the 1999 IEEE Knowledge and Data Engineering sExchange Workshop (KDEX'99)*, November 1999

[14] Brogdvision. http://www.broadvision.com.

[15] Like minds. http://www.andromedia.com.

[16] Bamshad Mobasher, Robert Cooley, and Jaydeep Srivastava. Creating adaptive web sites through usage- based clustering of urls. In Knowledge and Data Engineering Workshop, 1999.

[17] E. Cohen, B. Krishnamurthy, and J. Rexford. Improving end-to-end performance of the web using server volumes and proxy filters. In Proe. ACM SIGCOMM, pages 241-253, 1998.

[18] T. Fawcett and F. Provost. Activity monitoring: Noticing interesting changes in behavior. In Fifth ACM SIGKDD International Conference on Knowledge Dis- covery and Data Mining, pages 53-62, San Diego, CA, 1999. ACM.

[19] Mike Perkowitz and Oren Etzioni. Adaptive web sites: Automatically synthesizing web pages. In Fifteenth National Conference on Artificial Intelligence, Madison, WI, 1998.

[20] Mike Perkowitz and Oren Etzioni. Adaptive web sites: Conceptual cluster mining. In Sixteenth International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 1999.

[21] Alex Buchner and Maurice D Mulvenna. Discovering internet marketing intelligence through online analytical web usage mining. SIGMOD Record, 27(4):54-61, 1998.