© 2018 IJSRSET | Volume 4 | Issue 2 | Print ISSN: 2395-1990 | Online ISSN : 2394-4099
National Conference on Advanced Research Trends in Information and Technologies (NCARTICT-2018), Department of IT, L. D. College of Engineering, Ahmedabad, Gujarat, India In association with
International Journal of Scientific Research in Science, Engineering and Technology

Research Trends in Web and Web Usage Mining

Rachana Parikh*, Rachit Adhvaryu, Komil Vora
Information Technology, V.V.P. Engineering College, Rajkot, Gujarat, India

ABSTRACT

Web Technology is evolving very fast and Internet Users are growing much faster than estimated. The website users are using a wide range of websites leaving back a variety of information. This information must be used by the websites administrator to manipulate their websites according to the users of the websites. This actually is Data Mining. Web Mining is one of the applications of Data Mining. Web mining plays an important role in the decision making in the corporate, education and research environment. Modern developments in digital media technologies have evolved a huge amount of data transmitting over the web and with this huge data storage is required for easy and feasible access. Web Usage Mining is a mining of usage of websites and the information used and delivered on the websites. It is a technique to extract information from the web which includes web documents, hyperlinks between the documents and web usage logs. In this paper we describe the detailed survey of web mining, different techniques of web usage mining and its importance.

Keywords: Data Mining, Internet, Web, Web Mining, Web Usage Mining

I. INTRODUCTION

The World Wide Web (WWW) has lots of information and this information is increasing in a large amount daily. It is a very complex task to filter such information. A web or a website is a collection of web pages generally made using HTML and some programming languages which contains images, text, hyperlinks and similar digital data. The need to understand and use large, complex, information-rich data sets is common to virtually all fields of business, science, and engineering [1]. The technique to extract useful knowledge residing in these data and to act on that knowledge is becoming increasingly important in today's competitive world. Web Mining is the application of Data Mining techniques to retrieve information from the web which includes web documents, hyperlinks between the documents and web usage logs. Also the entire process of applying a computer-based methodology for discovering and retrieving knowledge from web documents is a web mining [1]. As the web data is updated every second, it is not compulsory that every user will get the same data whenever it is retrieved.

II. WEB MINING

Web Mining is classified into 3 main mining techniques [2]. The taxonomy of web mining is as follows:

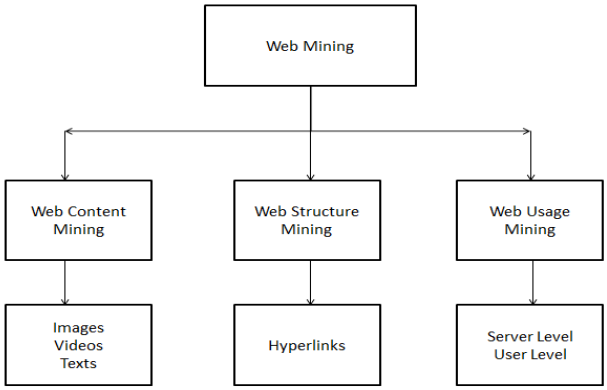

Figure 1. Web Mining Taxonomy

A. WEB CONTENT MINING

Web Content Mining is the way of retrieving important information from the web documents. The information can be in the form of images, videos, audios, texts and hyperlinks. A text mining and developing application for the same is one of the favourite topics for the researchers. Research in web content mining consists resource discovery from the web, categorizing and clustering of documents and extraction of informations from the webpages [2]. Here is the brief detail of web content mining techniques.

a) Image Mining

Image Mining is the technique which is used to detect unusual patterns and extract important and useful information from the images stored on the web or the large database. Thus image mining mainly deals with defining relationships between different images from the web and large databases [4]. Image mining is used in various fields like medical diagnosis, space research, remote sensing, agriculture and industries. The images include maps, geological structures and even it is used in the field of education.

b) Video Mining

Mining video is more complex than mining image data. Video is the collection of moving images like animation. There are 3 types of videos: 1) the produced (movies, news videos etc), 2) the raw (traffic, surveillance etc) and 3) the medical video (x-ray, cardiogram etc). The information from the video can be a) detecting trigger events (movement of vehicle and people), b) determining typical patterns of activity, generating person or object centric views of activity, and c) classifying activities into named categories (walking, sleeping etc), clustering and determination of interactions between two entities[5]. Video mining can also be classified into pixel based, statistics based, feature based and histogram based.

c) Audio Mining

As audio is the continuous media like video, the techniques and tools for audio processing and mining information is similar to video mining. Audio data can be in the form of radio, speech or spoken languages. Also television news has audio which are integrated with the videos [6]. Mining audio data requires conversion of audio to speech for better processing. Very few works has been carried out in this field.

d) Text Mining

The most trending research in the field of web content mining is text mining. The text mining refers to the text representation, classification, clustering, information extraction and search for hidden patterns. Text mining is the process of extracting useful information from the text and converting to automated discovery of knowledge [2]. It is natural extension of data mining or applying data mining techniques on a specific domain.

B. WEB STRUCURE MINING

This type of mining focuses on the data which describes the structure of the content of the web page. It is classified into two types: 1) intra-page structure: links within the page, 2) inter-page structure: links between 2 web pages [2]. This can be classified into two types based on structure of information:

a) Hyperlinks

A hyperlink is a structural unit that connects one web page with other web page either within same location or different location. A hyperlink connecting web pages in the same location is called intra-document hyperlink and a hyperlink connecting web pages at different locations is called inter-document hyperlink [2].

b) Document Structure

The content within a web page can also be organized in a tree structure based on HTML and XML Tags used to create a web page. Mining can be done to identify document object model (DOM) structures automatically from the documents [2].

C. WEB USAGE MINING

Web usage mining is one techniques of data mining to retrieve interesting and useful patterns from the web logs [3]. Web logs stores the identity or the origin of web users along with the browsing behaviour on the web site. Web Usage Mining can be grouped based on the type of usage logs:

a) Web Server Data

User logs are collected by the web server which includes IP address, page references and the time accessed by the user [3].

b) Application Server Data

Application servers are used to track various types of business events which can be used to improve the performance for any business firms [3]. For e.g. E-commerce websites uses such servers to know the events, business policies developed by their competitors.

c) User Level Data

User level data is the software developed using the information available from the web server and application server data [3]. It is an end user application which is used for various purposes.

D. Tools for Web Mining

Web mining tools helps the users to download essential information from the web. It collects the accurate and necessary information for the user which can be helpful in mining. The different tools are:

a) Automation Anywhere

It is a tool which is used to find web data very easily;. It is unique Intelligent and Smart Automation Application used for quick automation of any complex tasks [3].

b) Web Info Extractor

These tools is used to collect web content, constantly updating data and analysing data like images, videos, texts etc. [3].

c) Screen-Scraper

It is a tool used in searching databases and document structure. It provides a graphical interface allowing the user to navigate through URLs, data elements and hyperlinks and extract useful information from it [3].

d) Mozenda

This tool is used to extract and manage web data. User is allowed to setup tools at different places which can store and publish data at a regular interval of time [3].

e) Web Content Extractor

This tool is used to retrieve information from various websites like online auctions, online shopping, business directories, financial sites etc. The data can be represented in the form of excel, HTML, XML or any other script [5].

III. WEB USAGE MINING

Number of internet users in India is growing by 150K every month or 1.8 Million new users every year. India is the fastest growing online market in the world with 75% of the users being below the age of 35. The recent survey of Internet Users all over the world is described in image below: [15]

TOP 20 COUNTRIES WITH HIGHEST NUMBER OF INTERNET USERS - JUNE 30, 2017

#	Country or Region	Population, 2017 Est.	Internet Users 30 June 2017	Internet Penetration	Growth (*) 2000 - 2017	Facebook 30 June 2017
1	China	1,388,232,693	738,539,792	53.2 %	3,182.4 %	1,800,000
2	India	1,342,512,706	462,124,989	34.4 %	9,142.5 %	241,000,000
3	United States	326,474,013	286,942,362	87.9 %	200.9 %	240,000,000
4	Brazil	211,243,220	139,111,185	65.9 %	2,682.2 %	139,000,000
5	Indonesia	263,510,146	132,700,000	50.4 %	6,535.0 %	126,000,000
6	Japan	126,045,211	118,453,595	94.0 %	151.6 %	26,000,000
7	Russia	143,375,006	109,552,842	76.4 %	3,434.0 %	12,000,000
8	Nigeria	191,835,936	91,598,757	47.7 %	45,699.4 %	16,000,000
9	Mexico	130,222,815	85,000,000	65.3 %	3,033.8 %	85,000,000
10	Bangladesh	164,827,718	73,347,000	44.5 %	73,247.0 %	21,000,000
11	Germany	80,636,124	72,290,285	89.6 %	201.2 %	31,000,000
12	Vietnam	95,414,640	64,000,000	67.1 %	31,900.0 %	64,000,000
13	United Kingdom	65,511,098	62,091,419	94.8 %	303.2 %	44,000,000
14	Philippines	103,796,832	57,607,242	55.5 %	2,780.4 %	69,000,000
15	Thailand	68,297,547	57,000,000	83.5 %	2,378.3 %	57,000,000
16	Iran	80,945,718	56,700,000	70.0 %	22,580.0 %	17,200,000
17	France	64,938,716	56,367,330	86.8 %	563.1 %	33,000,000
18	Turkey	80,417,526	56,000,000	69.6 %	2,700.0 %	56,000,000
19	Italy	59,797,978	51,836,798	86.7 %	292.7 %	30,000,000
20	Korea, South	50,704,971	47,013,649	92.7 %	146.9 %	17,000,000

Figure 2. Country wise Internet Users

The above image clearly shows that India ranks 2nd and the internet users in India are increasing gradually. With this, the data and usage logs are getting larger and complex to manage and analyze. Web usage mining is the application of data mining techniques to discover usage pattern from Web data, in order to understand and better serve the needs of Web-based applications. Web usage mining consists of three phases: 1. preprocessing, 2. pattern discovery, 3. pattern analysis. Figure 3. represents these 3 phases of web usage mining.

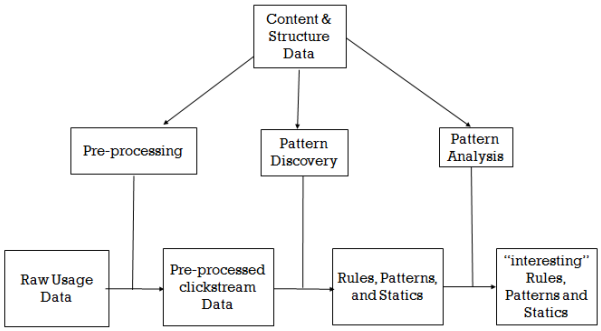

Figure 3. Web Usage Mining

The first is preprocessing state in which user sessions are inferred from log data. The second 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 third stage an information filter bases on domain knowledge and the web site structures is applied to the mining patterns in search for the interesting patterns [7]. The web usage mining process can be classified into two parts. The first part includes transforming the Web data into suitable transaction form. This includes Data Preprocessing and Data Integration components. The second part includes some Data Mining and Pattern Discovery [7].

A. Data Processing and Data Integration

Data preprocessing consists of data filtering, user identification, session/transaction identification, and topology extraction. Data filtering filters out some noise, i.e., unsuccessful requests, automatically downloaded graphics, or requests from robots, to get more compact training data [10]. People use some heuristic rules to identify user, such as IP address, cookies, etc. Data Preprocessing converts the available data sources into the data abstractions.

a) Usage preprocessing:

Usage pre-processing consists of Web pages, such as IP addresses, page references, and the date and time of accesses. Typically, the usage data comes from an Extended Common Log Format (ECLF) Server log [11].

b) Content Preprocessing:

Content pre-processing consists of converting the text, images, scripts and multimedia data into forms that are useful for the web usage mining process [11].

c) Structure Preprocessing:

Web structure mining analyses the link structure of the web in order to identify relevant documents. The structure of a site is created by the hypertext links between page views. The Google Search engine makes use of the web link structure in the process of determining the relevance of a page. The Google search engine achieves good results because while the keyword similarity analysis ensures high precision the use of a probability measure ensures high quality of the pages returned [11].

The usage data collected at the different sources such as Server level, Client Level and Proxy Level represent the navigation patterns of different segments of the overall Web traffic.

i) Server-level Collection:

A Web server log records the browsing behaviour of site visitors. The data recorded in server logs reflect the concurrent and interleaved access of a Web site by multiple users. These log files can be stored in various formats such as Common Log Format (CLF) or Extended Common Log Format (ECLF). ECLF contains client IP address, User ID, time/date, request, status, bytes, referrer, and agent.[10] Tracking of individual users is not an easy task due to the stateless connection model of the HTTP protocol. In order to handle this problem, Web servers can also store other kind of usage information such as cookies in separate logs, or appended to the CLF or ECLF logs. Packet sniffing technology (also referred to as "network monitors") is an alternative method for collecting usage data through server logs. [10]

ii) Client level collection:

Client-side collection can be implemented by using a remote agent (such as Java scripts or Java applets) or by modifying the source code of an existing browser (such as Mosaic or Mozilla) to enhance its data collection capabilities.[9]

iii) Proxy Level Collection:

The Internet Service Provider (ISP) machine that users connect to through a model is a common form of proxy server. A web proxy acts as an intermediary between client browsers and Web servers. Proxy-level caching can be used to reduce the loading of time of a Web page experienced by users as well as the network traffic load at the server and client sides. [10]

B. Data Mining and Pattern Discovery

Data Mining comprises of the following data techniques which are as follows:

a) Association Rules:

Association rule generation can be used to relate pages that are most often referenced together in a single server sessions. 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.[10] Association rule mining has been well studied in Data Mining, especially for basket transaction data analysis. Aside from being applicable for e-Commerce, business intelligence and marketing applications, it can help web designers to restructure their web site. 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 [7].

b) Clustering:

Clustering is a technique to group together a set of items having similar characteristics. 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 behaviours, and it is critical for user community construction. This information is useful for the Internet search engines and Web assistance providers. [7]

c) Deviation/Outlier Detection:

It contains techniques aimed at detecting unusual changes in the data relatively to the expected values. Such techniques are useful, for example, in fraud detection, where the inconsistent use of credit cards can identify situations where a card is stolen. The inconsistent use of credit card could be noted if there were transactions performed in different geographic locations within a given time window. [12]

d) Statistical Analysis:

Statistical techniques are the most common method to extract knowledge about visitors to a web site. By analyzing the session file, one can perform different kinds of descriptive statistical analyses (frequency, mean, median, etc) on variables such as page views, viewing time and length of a navigational path. 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 information can be useful for improving the system performance. [12]

Pattern discovery uses methods and algorithms developed from several fields such as statistics, data mining, machine learning and pattern recognition. The knowledge that can be discovered is represented in the form of rules, tables, charts, graphs, and other visual presentation forms for characterizing, comparing, predicting, or classifying data from the web access log [13]. Several pattern discovery techniques are as follows:

a) Sequential Patterns:

The technique of sequential pattern discovery attempts to find inter-session patterns such that the presence of a set of items is followed by another item in a time-ordered set of sessions or episodes. A new algorithm MiDAS (Mining Internet data for Associative Sequences) for discovering sequential patterns from web log files has been proposed that provides behavioural marketing intelligence for e-commerce scenarios.[13] MiDAS contains three phases: 1. A priori phase is the input data preparation, which consists of data reduction and data type substitution. 2. Discovery Phase discovers the sequences of hits and generates the pattern tree. 3. A posteriori Phase filters out all sequences that do not fulfil the criteria laid in the specified navigation templates and topology network and also pruning is done in this phase. By using this approach, Web marketers can predict future visit patterns, which will be helpful in placing advertisements aimed at certain user groups [13].

b) Dependency modeling:

Dependency modelling is another useful pattern discovery task in web mining. The goal here is to develop a model capable of representing significant dependencies among the various variables in the web domain. As an example, one may be interested to build a model representing the different stages a visitor undergoes while shopping in an online store based on the actions chosen (i.e., from a casual visitor to a serious potential buyer. [14]

c) Pattern analysis:

Pattern analysis is the last step in the overall Web Usage mining process. The motivation behind pattern analysis is to filter out uninteresting rules or patterns from the set found in the pattern discovery phase. The exact analysis methodology is usually governed by the application for which Web mining is done [13]. The most common form of pattern analysis consists of a knowledge query mechanism such as SQL. Another method is to load usage data into a data cube in order to perform OLAP operations. 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 [14].

IV. APPLICATIONS OF WEB MINING

There are many applications of web mining. Most dominating applications of web mining are as follows:

a) In the world of online shopping, it is very important to know the customer behavior and experience with the website. The feedback of the experience given by the user helps the website owner to improve their content in an efficient way. The main target of the website owner is that once a customer is visiting the website for a purchase, the user should not move to the other websites [9].

b) Google is the best and widely used search engines. It provides the users to access information from over billions of web pages indexed on its server. The quickness and quality of information provided by the search engines make them the most successful search engines. Web mining helps the search engines to know the behavior of the user, the keywords they search and based on this they give a Page Rank [9].

c) Web usage mining offers users the ability to analyze massive volumes of click stream or click flow data, integrate the data seamlessly with transaction and demographic data from offline sources and apply sophisticated analytics for web personalization, e-CRM and other interactive marketing programs [8].

d) Personalization for a user can be achieved by keeping track of previously accessed pages. These pages can be used to identify the typical browsing behavior of a user and subsequently to predict desired pages [8].

e) Web usage patterns can be used to gather business intelligence to improve Customer attraction, Customer retention, sales, marketing and advertisement, cross sales [8].

f) Mining of web usage patterns can help in the study of how browsers are used and the user's interaction with a browser interface [8].

V. CONCLUSION

As the web data and its usage increases day by day, it is very important to analyse the web data and retrieve the information. Thus web mining, web usage mining and its techniques play an important role in information extraction from the web. In this paper, we have briefly discussed various web and its usage mining techniques, tools and applications. By the techniques and the tools described in the paper, one can use in his/her research and new techniques can be developed for more effective, efficient and faster results. We hope that this primary discussion is the beginning for a fruitful researches and results in future.

VI. REFERENCES

[1]. Arvind Kumar Sharma, P.C. Gupta, -"Exploration of efficient methodologies for the improvement in web mining techniques-A survey", International Journal of Research in IT & Management (ISSN 2231-4334) Vol.1, Issue 3, July 2011.

[2]. G. Srivastava, K. Sharma, V. Kumar,"Web Mining: Today and Tomorrow", in the Proceedings of 2011 3rd International Conference on Electronics Computer Technology (ICECT), pp.399-403, April 2011.

[3]. S. K. Madria, S. S. Bhowmick, W. K. Ng,and E. P. Lim, "Research Issues in Web Mining", in proceding of data mining and knowledge discovery, 1st International conference, DaWk 99, pp 303-312, 1999.

[4]. Ji Zhang, Wynne Hsu and Mong Li Lee "An Information-Driven Framework for Image Mining" Database and Expert Systems Applications in Computer Science, 2001, Volume 2113/2001, 232-242, DOI: 10.1007/3-540-44759-8_24

[5]. Boresczky J. S. and L. A. Rowe, "A Comparison of Video Shot Boundary Detection Techniques", Storage & Retrieval for Image and Video Databases IV, Proc. SPIE 2670, 1996, pp.170-179.

[6]. A. Czyzewski, "Mining Knowledge in Noisy Audio Data", in Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, pages 220-225, 1996.

[7]. J.W. Han, M .Kamber, Data Mining-Concepts and Techniques, Elsevier Science & Technology Books, 2006.

[8]. J. Srivastava, R. Cooley, M. Deshpande, P. Tan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", SIGKDD Explorations, 1(2), 20001, pp.2-23.

[9]. Jaideep Srivastava and Robert Cooley "Usage Mining: Discovery and Applications of Usage Patterns from Web Data", ACM SIGKDD Explorations Newsletter .2000,1(2):12-23.

[10]. Bamshad Mobasher, "Web Usage Mining", Springer Berlin Heidelberg, 2007, 449-483.

[11]. Navin Kumar Tyagi, A.K. Solanki and 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, pp. 279-283.

[12]. Jose Roberto de Freitas Boullosa. "An Architecture for Web Usage Mining".

[13]. Yan Wang." Web Mining and Knowledge Discovery of Usage Patterns". CS 748T Project. February, 2000.

[14]. Cyrus Shahabi, Amir M. Zarkesh, Jafar Adibi, and Vishal Shah, "Knowledge Discovery from Users Web-page Navigation", IEEE RIDE 1997.

[15]. http://www.internetworldstats.com/top20.htm