Web Mining: A Framework

Surbhi Sharma*[1], Sudhir Kumar Sharma*[2]

[1,2]Institute of Information Technology and Management, New Delhi-58, India

[1]surbhisharma1304@gmail.com and [2]sharmasudhir08@gmail.com

Abstract—*With the availability of huge amount of data on the World Wide Web, it became a fertile place for data mining research. The growth of web data will no doubt continue to grow in coming years. So to analyses all the data in a manner to produce or extract information reflecting user behavior, interaction, demands, and to optimize search results, the concept of web mining is used. Web mining is basically a technique of data mining. Web mining is come under the applications of the data mining approaches to order discover analyze the patterns extracted from the web. The main objective of web mining is to develop intelligent tools to make easy for the user to extract, filter, find and evaluate useful information. Nowadays, data available on the web has become an essential part of organization. Data is produced in huge amount, as a result of interaction of several users and web. The output can be extracted to generate knowledge so that later it can be applied in various applications. Analysis of web site content and patterns obtained by user navigation is valuable for business and research community. The aim of research in web mining is to develop and apply new techniques to mine and extract valuable knowledge or information from the web pages. Due to the diversity and unstructured form of web data, discovering of targeted or unexpected knowledge/information automatically is the important and challenging task. So the focus of this paper is to provide a more evaluative update of web mining research and techniques available. This paper, provide the reviews for concept of Web mining, the type of Web mining and different techniques used in each type. This paper discusses the current trends and challenges in this research area.*

Keywords—*Clustering, Association rules, Pattern discover, Hyperlink analysis, Pattern analysis, Web access logs*

I. INTRODUCTION

Today, World Wide Web (WWW) becomes a large repository of information which is be used to store, disseminate, retrieve information and manipulates data every day. The data available on web includes text, tables, multimedia, audios, videos, hyperlinks, metadata, structured records, web logs etc. Due to the factors such as dynamic, high dimensionality, diverse, huge of web data, Web Researchers faced many problems like multimedia alignment, temporal issues, scalability, etc. The information on the web increases in such a way that it becomes as endless as ocean. This intense growth of information can be in the form of structured or semi structured data. So, to manage this highly evolving content and high data dimensionality, it become necessary to develop new approaches and methods in order to organize the data and extract some relevant information according to our requirements and applications.

Data mining techniques are applied to web data that refers web data mining or web mining. Web mining includes analysis and extraction of relevant information from the data available on World Wide Web. Anyone can easily deluge with data because of unstructured, heterogeneous and partial structured data available on the web.

So, mining the web have become important and challenging task for data mining and data management professionals [1]. Web mining is further divided as Web content mining, web structured mining and web usage mining. Each classification has its own tools and algorithm and used to serve different purposes. As the data will increase on web, this technology plays a main role in extracting knowledge from the web.

The objective of this paper is to provide the reviews for concept of Web mining, the type of Web mining and different techniques used in each type in the last decade. This paper discusses the current trends and challenges in this research area.

This paper is divided into five different sections. Section 2 describes Web Mining. Section 3 describes Web Content Mining. Section 4 discusses Web Structured Mining. Section 5 describes Web Usage Mining. The conclusion is presented in the last section 6.

II. WEB MINING

Data mining can be viewed as a result of the natural evolution of information technol- ogy. There is evolution in the database system by introducing data warehouse

which helps in collecting data, creating information, managing them and analyze them. The data can be in any form like storage, search, retrieval, logs, transactions, etc. [2]. Web mining is one of the important application of data mining techniques. Web mining helps to discover and extracts useful knowledge from the unstructured or semi-structured data available on the web.

Web miming can be classified into three categories. Fig. 1. shows Web Mining Cate- gories. The brief introduction are as follows:

- Web Content Mining: It deals with the content of web pages like text, audio, sound and other multimedia to extract valuable information.
- Web Structured Mining: It deals with structure of a linking of web pages inside a website to discover web graph pattern and generate some useful information.
- Web Usage Mining: It deals with the web logs records to find the user interaction with the web.

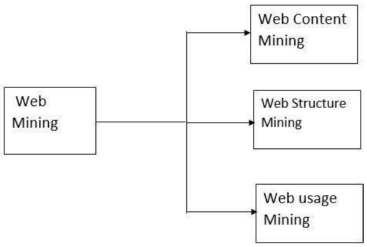

Fig. 1. Web Mining Categories

Technically, the focus of web content mining is on the content of web pages like what type of content is shown, how the information is conveyed, etc., while Web structure mining tries to discover web graph of each website that is the linking structure of the hyperlinks. On the basis of structure of the hyperlinks, Web structure mining will classify the Web pages and produce the information like relationship among different web- sites, how websites are similar to each other, etc. Web usage mining deals with the user interaction with the web and extracting valuable information.

III. WEB CONTENT MINING

Web Content Mining is the process of mining and extracting useful information from the web documents and then indexing them in order to retrieve quickly and users can find information easily. The content of web documents may comprise of text, images, audio, video, sound, structured records such as list and tables and other multimedia data. The data on the web documents can be in form of semi structured or unstructured data. The group of facts that a web page is designed is called content data. It generates the interesting patterns about the user needs. One of the technique to mine text is called text mining. Text documents are related to machine learning, text mining and natural language processing. Text mining involves extraction of information in order to study pattern recognition, word frequency distributions, annotation, and text analysis and pro- duce some valuable knowledge. Text mining deals with natural language texts either stored in semi-structured or unstructured formats. The information extracted can be used to derive summaries of the document.

There are three approaches for web content mining to mine data. Fig. 3. shows Web Content Mining Approaches. The brief introductions are given in the nest section of paper.

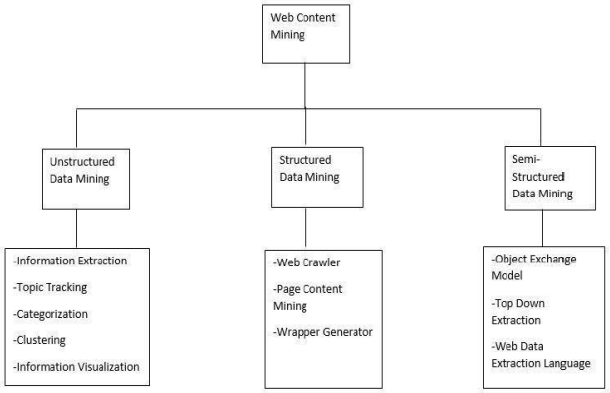

Fig. 4. Web Content Mining Approaches

3.1 Unstructured data mining

Information Extraction. Web pattern matching is used to extract information from unstructured data. This technique is very useful for the huge volume of text. It follows a procedure in which the first step is to detect the keywords and phrases and second step involves discovering of connection of phrases and keywords in the text [3]. In this way, useful data is extracted from which information is mined and then using different approaches missing information is found to complete the information. In Information Extraction, the unstructured data is converted into some structured form [4].

Topic Tracking. As the name suggests, this technique is used to find the documents that are related to the interest of the user. This technique studies the user profile and keeps the record of those documents which are accessed and visited by the user. Yahoo has applied topic tracking, in which a keyword is discovered from user and anything related to that keyword will notify the user. Basically, this technique is used by two fields namely, medical and educational field respectively. In education field, user can easily find latest course or any information related

to work. In medical field, doctors can easily get to know latest treatment and news in their respected fields [3] [5].

Categorization. In this technique, the first step is to count the number of words and their meaning in the document and then find the suitable predefined head topic. Rank is then specified to the document according to the head topic. Web pages with massive content on a given topic are set to rank first. Thus, this technique discovers the head theme and placed the web pages in predefined groups accordingly [3] [4].

Clustering. In clustering, there are no predefined topics. The topics are defined on the basis of data extracted from the content of web pages. Then the grouping of web documents is done on the basis of these topics and similar documents get grouped together. In this way, important documents are not excluded from search result and help the users to choose their topics in which they are interested [3].

3.2 Structured data Mining

Web crawler. Web crawlers are like type of computer program which pass over from the HTML structure in the web. Anyone can use this technique to extract and gain the information available on the web. The very big example is of search engines which use web crawlers to gather information about content available on the web documents. Web crawlers are further divided into two types namely, inter and external web crawler. Internal web crawler passes over internal structure of website and external web crawler pass over different websites [3].

Page Content Mining. Page content mining technique used to find only structured data from the web pages and then mine them to gather information. Rank is given to these pages. Also, the search engine ranked the web pages by comparing web page rank [6].

Wrapper Generator. Wrapper generator generates the information on the basis of sources. As web pages are ranked by search engines according web page rank, the web pages are searched on the basis of query [6].

3.3 Semi-Structured Data Mining

Object Exchange Model. In this method, the useful information is get discover from semi structured data and then gather in the groups and this information gets stored in Object Exchange Model. This method is very useful to understand the information structure available on the web accurately. The very helping feature of this model is that there is no need to describe the structure of an object in prior, the model itself describes the object [6, 7].

Top down Extraction. In this technique, compound objects are discovered from abundant web sources and then defragment them until atomic objects get extracted. [3]. **Web Data Extraction Language**. In this technique, the relevant data is stored in the form of table after converting content of web pages in the structured form and then take this data to the users [6].

IV. WEB STRUCTURED MINING

Web structure mining describes the structure of a particular website, how the web pages are connected with each other via hyperlinks. Fig. 5. Shows Web Graph Structure Web structure is used to produce web pattern graph [8]. The web graph pattern mainly consists of web pages and web documents as nodes and hyperlinks act as edges that connect two related web page.

Web page contains HTML tags due to which web pages can organized in a tree structure format based on Document Object Model (DOM) and it helps in the research of link analysis. Link mining has important tasks on the basis of links like classification, cluster analysis, cardinality and sturdiness of link. The study of the hyperlink structure is also called hyperlink analysis [9]. Hyperlinks provide connection to web pages to go on a location either in same webpage or different web page. There are two categories of a hyperlink namely, inter document and intra document. Intra document hyperlink connects different parts of the same page and on other hand, inter document hyperlink connects two different pages.

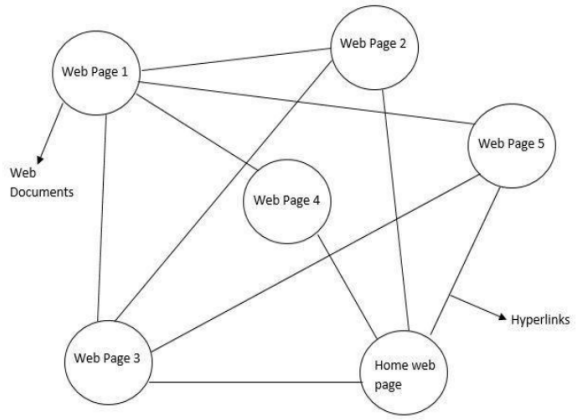

Fig. 6. Web Graph Structure

4.1 Hyperlink Analysis

With the hyperlinks, additional information about a website can be extracted to opti- mize the search result. Hyperlink analysis is a technique used to evaluate relationship (connection) between nodes (web page). There are many algorithms which are used for hyperlink analysis. The two important algorithms are page rank and Hyperlink Induced Topic Search (HITS).

V. WEB USAGE MINING

Web usage mining is the process of tracking behavior of users online by extracting useful information from server logs. For this reason, it is also known as web log mining. User access data is collected from the web. Several users surf the web, follow some pattern and analyzing these patterns enable to find the way user interacts with the web. Thus, this technique is used to predict the behavior of the user. Based on the how the user interacts with the websites, web usage mining copes with the order how one can make personalized web pages or enhanced search engines. The web log data is stored at different locations like web server, web proxy server and client browser. Huge amount of data is stored at location. The data can be in form of semi structured and unstructured data which contain lots of noisy data, errors, missing attributes, failed re- quest message, incomplete data and irrelevant data. With the help of web usage mining techniques web log data is analyzed.

5.1 Web Servers Logs

There are four types of web server logs. Fig. 7. Shows Web Server Logs

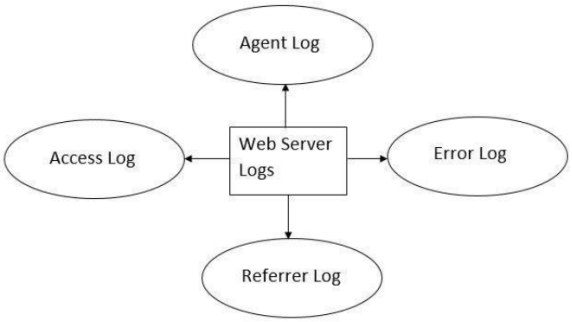

Fig. 8. Web Server Logs

Access Log. This log stores the information of user activities on web and has many attributes like click event, visits, search and access of the user.

Agent Log. Agent log store the details of user's online interest like type of browser a user uses, browser version, types of applications downloaded by the user, etc.

Error Log. This log is used to store the information about links or pages on which a user clicks but the page is enable to open and shows failure like error 404 not found.

Referrer Log. The information about the URLs of websites that link to web pages are stored in the referrer log. If a user clicks on a link from a website to go to the other website, then URL of that website gets stored in this log [20].

5.2 Phases of web usage mining

The process of web usage mining is given in Fig. 5. There are three phases of web usage mining. The brief introduction is as follows:

Data Pre-processing. This phase retrieves the raw data and processes it to make it relevant and organize the data to produce useful information. For better efficiency and scalability, data is go through many steps like data cleaning, integration, transformation, reduction and discretization [14].

Discovery of pattern. In this phase, algorithms and rules are applied to extract the pattern formed. Classification, clustering, association rules and sequential analysis are some techniques used to discover pattern.

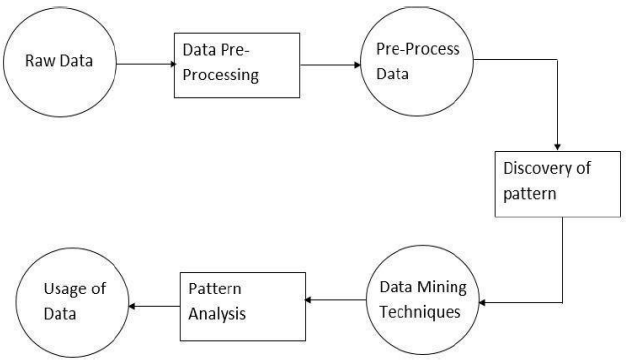

Fig. 9. Process of Web Usage Mining

Analysis of pattern. After pattern discovery, the pattern is checked and analyzed to generate valuable knowledge. Various techniques are there for pattern analysis to ex- tract useful knowledge and finally get the useful pattern used by user and use this in- formation for commercial strategies [15].

This category of web mining has several tools and approaches to analyze the behavior of the user. It mainly uses data mining algorithms such as association rule mining, sequential rule mining and clustering.

5.3 Association Rule

Association rule is the most basic and widely used method. It is used to find the association and correlations among large set of web pages that are frequently access together in the user browser session. This rule shows how frequently an item-set occurs in a transaction. These rules are statements in the form M ==> N where (M) and (N) are the set of available items in a set of transactions. The rule of M ==> N states that, transactions that contain items in X, may also include items in Y [17]. For example,

WebPage1, WebPage2, WebPage3 ==> WebPage4

In this example, if a user visits webpage1, webpage2 and webpage3 then the user will most likely to visit webpage4 as well. Apriorialgorithm having its set of rules can be used that is used to extract pattern by applying some rules on the frequent occurrence of web pages by user.

5.4 Sequential analysis

. Sequential analysis method is used to find the frequent navigation performed by the user. Sequential analysis is the analysis of navigation performed by the user. This method uses data mining techniques to analyze the sequential data and then extract the patterns. It is used to extract interesting sequences and their subsequence and then group them together. To measure the interesting subsequence, various criteria are there such as number of occurrences, length, frequency, etc. In sequential analysis, evaluation of data is processed when they are collected and it is stopped according to the predefined rule (known as stopping rule) when required results are observed. MIDAS (Mining Internet Data for Association Sequences) algorithm can be used for extracting sequential patterns in order to provide marketing intelligent behavior for ecommerce scenario [18].

5.5 Clustering

Clustering is the technique that group together the abstract objects into cluster of similar objects that is, Clustering is the process of grouping objects together in such a way that the objects having similar characteristics, rely in the same group are identical and those belonging to different groups are not identical. It is an unsupervised machine learning- based algorithm that consists of a group of data points into clusters so that the objects having same characteristics included in the same group. There are different methods and techniques used for cluster analysis.

Clustering identifies the user with identical behavior so it can help in personalizing the website. Clustering can be done in two ways, first as usage clustering, in which clustering is done on the basis of those users that have same browsing pattern and second as page clustering, in which clustering is done on the basis of web pages containing same content.

VI. CONCLUSION

This paper has attempted to give the detailed review on the concept of web mining which acts as a framework to extract pattern and analyze valuable information from the web on the basis of con- tent, hyperlinks and web logs. The main purpose of web mining is discovering useful information from the World-Wide Web and its usage patterns. This paper also dis- cussed its categories- web content mining, web structured mining and web usage mining along with the techniques and methods used in each category of web mining. Web content mining extracts the knowledge, in which the data like text, audio, video, documents, records, tables, etc. of web documents are mined. Web Structure Mining emphasis on analysis of the web pattern graph that is, the link structure of the websites. Web usage mining extracts knowledge from user navigation patterns through web data and also uses secondary data like data generated by the user through surfing the web to find patterns. Web usage mining collects the data from different web logs records to find user visit and access pattern on the web. This paper discussed the current trends and challenges in this research area.

REFERENCES

[1] Jiawei Han, Kevin, Chen-Chuan Chang "Data Mining for Web Intelligence" IEEE International Conference on Data Mining, 2002.

[2] Kosala and Blockeel, ―Web mining research: A survey,‖ SIGKDD:SIGKDD Explorations: Newsletter of the Special Interest Group(SIG) on Knowledge Discovery and Data Mining, ACM, Vol. 2, 2000

[3] Sharma, Arvind Kumar, and P. C. Gupta. "Study & Analysis of Web Content Mining Tools to Improve Techniques of Web Data Mining" International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) Volume 1 (2012)

[4] Srividya, M., D. Anandhi and M. I. Ahmed. "Web mining and its categories– a survey "In- ternational Journaof Engineering and Computer Science, IJECS 2.4 (2013)

[5] Deepti Sharda and Sonal Chawla "Web Content Mining Techniques: A Study."International Journal of Innovative Research in Technology & Science

[6] Johnson, Faustina, and Santosh Kumar Gupta. "Web Content Mining Techniques: A Sur- vey."International Journal of Computer Applications (0975–888) Volume (2012)

[7] Srividya, M., D. Anandhi and M. I. Ahmed. "Web mining and its categories–a survey."In- ternational Journal of Engineering and Computer Science, IJECS 2.4 (2013).

[8] Joy Shalom Sona, Prof. Asha Ambhaikar" A Reconciling Website System to Enhance Ef- ficiency with Web Mining Techniques" International Journal Of Scientific & Engineering Research Volume 3, Issue 2, February-2012 1 ISSN 2229-5518

[9] Mamta M. Hegde, Prof. M.V.Phatak, "Developing an approach for hyperlink analysis with noise reduction using Web Structure Mining", International Journal of Advanced Research in Computer Engineering & Technology Volume 1, Issue 3, May2012

[10] Q. Lu, and L. Getoor. Link-based classification. In Proceedings of ICML-03, 2003

[11] N. Duhan, A.K. Sharma and K.K. Bhatia, PageRanking Algorithms: A Survey, Proceedings of the IEEE International Conference on Advance Computing, 2009.

[12] T.Nithya, International Journal of Advanced Research in Computer and Communication Engineering Vol. 2, Issue 8, August 2013

[13] Ashutosh Kumar Singh, Ravi Kumar P, "A Comparative Study of PageRanking Algorithms for Information Retrieval",International journal of electrical and computer engineering 4:7:2009

[14] Mitali Srivastava, Rakhi Garg, P. K. Mishra," Preprocessing Techniques in Web Usage Mining: A Survey" International Journal of Computer Applications (0975 – 8887) Volume 97– No.18, July 2014

[15] Amit Pratap Singh1, Dr. R. C. Jain 2," A Survey on Different Phases of Web Usage Mining for Anomaly User Behavior Investigation" International Journal of Emerging Trends & Technology in Computer Science (IJETTCS)Volume 3, Issue 3, May – June 2014 ISSN 2278-6856

[16] Dr.S. Vijiyarani1 and Ms. E. Suganya2, International Journal of Computer-Aided Technologies (IJCAx) Vol.2, No.3, July 2015

[17] Nasrin JOKAR, Ali Reza HONARVAR, Shima AgHAMIRZADEH, Khadijeh ESFANDIARI, Bulletin de la Société des Sciences de Liège, Vol. 85, 2016, p. 321 – 328