IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 5, Issue 4, April 2016*

# A Review on Data Extraction using Web Mining Techniques

**Vidya Shree S I[1], Pooja M R[2]**

PG scholar, Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India[1]

Associate Professor, Department of Computer Science and Engineering, Vidyavardhaka College of Engineering, Mysuru, Karnataka, India [2]

**Abstract:** Today internet is full of structured or unstructured information and this information influences people or society directly or indirectly. With the rapid growth of internet technologies, the web is considered as a world's largest repository of knowledge. Web data processing is the method of handling high volume of data which is not so easy. Therefore, for the user benefits, web mining is used to identify the patterns by understanding the customer behaviour and evaluating a particular website based on the information stored in web log files. Web Mining is evaluated by using data mining techniques, like classification, clustering, and association rules. In this paper, we have given complete review on how to extract the useful data by using web mining techniques.

**Keywords:** Web mining, Classification, Clustering, Association rule.

## I. INTRODUCTION

Internet is a shared global computing network which is a platform for both web services and World Wide Web. Web is an immense source of data which is freely available for the users to access. It is a collection of documents, text files, audios, videos and other multimedia data. As there are different types of data it should be organized in a proper way so that users can access it efficiently. This is where the data mining comes to picture. In data mining, data is extracted in terms of patterns or rules from huge amount of data. Among many applications of data mining techniques, web mining is one such application which automatically discovers and extracts potentially useful and previously unknown information or knowledge from the web.

Web mining [1] also called as web log mining is a part of both information extraction and information retrieval. It supports machine learning because it improves the classification of text. The main aim of web mining is to extract information i.e it is the integration of information which is gathered by traditional data mining techniques with the information gathered over World Wide Web. Web mining is decomposed into following subtasks:

1. Resource Finding: Process of retrieving he services and documents that is either online or offline from the text sources available on the web.
2. Information selection and preprocessing: It selects and preprocesses specific information from the web sources automatically.
3. Generalization: The general patterns are uncovered automatically at individual Web sites as well as across multiple sites by using data mining and machine learning techniques.
4. Analysis: The mined pattern will be validated and interpreted.
5. Visualization: The final result will be in visual form for the users to understand in an easy way.

Web mining is divided into three categories [2] depending on type of the data which is extracted as shown in Fig.1.

1. Web content mining (WCM) is used to extract valuable information from web documents, such as HTML or XML documents.
2. Web structure mining (WSM) is a process of discovering structured information from websites, such as hyperlinks on web pages
3. Web usage mining (WUM) reflects the user's behaviour which catches the meaningful patterns such as web logs or clickstreams from one or more web localities.

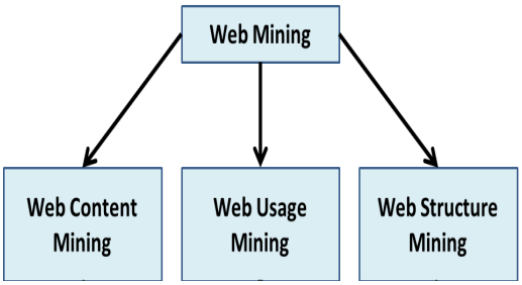


Fig 1 Web Mining Categories

## II. WEB CONTENT MINING

Web Content Mining also known as web text mining is the process of mining, scanning and extraction of text, videos, graphs and pictures from the website or web documents. It may be structured or unstructured / semi structured even though much of web is unstructured. The information is retrieved from the web into more structured forms and index the information to retrieve quickly or finding

valuable information from web content or web documents. The result which we got after the mining may be a structure documents or unstructured ones.

There are two different points of view of web content mining: the information retrieval view to improve filtering and finding the information to the users and the database view to manage the web data. Two approaches are used in web content mining. They are: the database approach and the agent based approach

(i) Agent Based Approach – focuses on searching relevant information from World Wide Web and helps in organizing the collected information. There are three types of agents: Intelligent search agents automatically searches for information along with a query, Information filtering agents which filters the data and personalized web agents which discovers the documents related to the user profiles.

(ii) Database Approach – It consists of databases which contain attributes, tables and schema with defined domains. It retrieves the semi-structured data from web documents.

2.1 Web Content Mining Techniques
Web content mining has different techniques to extract data [3]: unstructured mining, structured mining and semi-structured mining

2.1.1 Unstructured Data Mining
Text document is the form of unstructured data. The data which is available on the web is of unstructured format. Applying the data mining techniques to the unstructured data is known as knowledge discovery in texts.

Information Extraction
Pattern matching is used to extract the information from unstructured data. It includes the process of tracing the keywords and phrases and later finding the connection of keywords within text. Then information is mined from extracted data and transforms unstructured text to structured text by discarding the incorrect predictions.

Topic Tracking
This includes the checking of the documents viewed by the user and the user profile is analysed. Depending on the user interest the documents are predicted. It is mainly used in medical and education field. Sometimes it provides the information not related to the user topic which is a drawback.

Summarization
Using this technique the confusion of the user to decide whether the read the topic or not will be avoided. This is because this technique will reduce the length of the document by maintaining only the important points. It uses two methods extractive method and abstractive method.

Clustering
It includes grouping the similar documents to form clusters. This is done on fly basis rather than predefined topics basis. This technique helps user to select the topic of their interest. This technology is used in management information system.

Information Visualization
If the documents are having any similarities between them, it can be found out through visualization. It utilizes feature extraction and key term indexing. Large textual materials are represented as visual hierarchy or maps where browsing facility is allowed.

2.1.2 Structured Data Mining
Data in the form of list, tables and tree is structured data which is easy to extract compared to unstructured data.

Web Crawler
These are the computer programs which can traverse the hypertext structure in web. Usually search engines uses these web crawlers to collect the information available on the web pages. External and internal web crawlers are the two types of web crawler.

Wrapper Generation
This will provide the information based on the capability of sources. The sources are what query they will answer and the output types. Here web pages are retrieved according to the query by using page rank value.

Page content mining
It is a technique which works on the pages ranked by traditional search engines. By comparing the page content ranking, the pages are classified.

2.1.3 Semi-Structured Data Mining
Sometimes source does not impose rigid structure on data which is referred as semi-structured data.

Object Extraction Model
From the semi-structured data the relevant information is extracted and it is collected in a group which is stored in Object Extraction Model (OEM). The user can understand the structure of the information accurately from this model.

Top down Extraction
It includes the process of extracting complicated objects from rich sources and decomposing into less complex objects until atomic objects are extracted.

Web data Extraction Language
It helps in converting web data to structured data and delivered to end users.

## III.WEB STRUCTURE MINING

Web structure mining is the study of data interconnected to the structure of a particular website. It is the process of analysing the node and connection structure of a web site by using the graph theory. The main purpose for structure mining is to extract previously unknown relationships between Web pages. A useful source for extracting information is a Web structure which can be done either at intra-page level or inter-page level. According to the type of web structural data [4], web structure mining can be divided into two kinds: hyperlinks and document structure.

3.1 Document structure
Web page is organized in tree structure format based on HTML or XML tags. By using Document Object Model

IJARCCE
ISSN (Online) 2278-1021
ISSN (Print) 2319 5940
*International Journal of Advanced Research in Computer and Communication Engineering*
*Vol. 5, Issue 4, April 2016*

(DOM) the documents are extracted automatically.

3.2 Hyperlinks
Hyperlink is one of the components which connect webpage to different location. If it is connected to a different part of the same page then it is called intra-page hyperlink and it is a document structure level. A hyperlink that connects two different pages are called inter-page hyperlink which is of structure level. By using the hyperlinks the patterns can be extracted in the web.

Link mining is the research of structure analysis. The research of the hyperlink level is also called as hyperlink analysis which is used to retrieve useful information from the web. The main reason for developing link mining is to understand the social organization of the web.

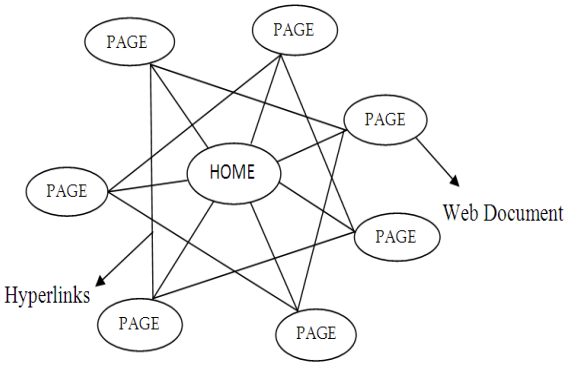


Fig 2 Web Graph Structure

Web structure mining is used in search engines like Google, Yahoo, etc. This mining technique reduces irrelevant search results. It also indexes the information on the web which causes low amount of recall with content mining.

## IV. WEB USAGE MINING

It is the area of web mining which deals with extraction of knowledge of interest from logging information produced by web servers, web clients and proxy servers.

Hence it is also called as web log mining. It includes extraction of information from different web log files to find which of the pages are mostly accessed by the people, who are accessing what type of web sites and for what purpose and which pages are accessed together or in serial order.

Web usage mining applications [5] are based on the data collected from three main sources:

(i) Web servers which collect large amount of information such as name and IP of remote host, date and time of the request, etc. in the log files. There are different types of web server logs: Access logs, Agent logs, Error logs and Referrer logs

(ii) Proxy servers which collects data of group of users accessing huge groups of web servers. Sometimes session reconstruction will be difficult but when there is no caching between proxy server and the clients, the identification of users_sessions is easier.

(iii) Web clients which will track the data by using JavaScript, Java applets or even modified browsers.
This will provide detailed information about actual user behaviours.

4.1 Process of Web Usage Mining
Web Usage Mining process is divided into three phases: Preprocessing, Pattern Discovery and Pattern Analysis.

4.1.1 Data Pre-processing
This step [6] includes identifying users, sessions, and page views and so on. These types of data may be noisy or inconsistent so we have to pre-process them to improve efficiency and scalability which can be done in following steps

1. Cleaning – It refers to removing irrelevant data like additional request and error entries from web log file so that only the useful information will remain for the next step.
2. User Identification – The users who have different ip address are called unique users which will be identified in this step. This is very important to extract the user access characteristics.
3. Session Identification – This step includes the differentiation of the web log entries into different user sessions by a session timeout.

4.1.2 Pattern Discovery
In this phase, different data mining techniques [7] are applied on web log files after preprocessing to discover the useful pattern.

1. Association rule mining – It is a technique which is mostly used for pattern discovery. It works on generating frequent pattern and rules.
2. Clustering – It finds the group of common behaviour users without the knowledge of group definition. Many algorithms are developed and are categorized such as hierarchical methods, partitioning methods, density-based methods and grid-based methods.

4.1.3 Pattern Analysis
The last process of web usage mining is pattern analysis. In this phase irrelevant pattern are removed from the pattern which identified during pattern discovery phase. It includes many techniques such as visualization technique, OLAP technique, data and knowledge querying and usability analysis.

## V. CONCLUSION

This paper has tried to provide a complete review of data extraction by using different techniques of web mining. Web Mining techniques are used for the extraction of web information. When the techniques are used accurately according to the requirements of the user, even a large amount of data can be maintained and clearly extracted. The detailed study and analysis of each web mining technique have been done in this paper. Though various algorithms and techniques have been proposed still work has to be done in discovering new tools to mine the web. This study and review would be helpful for researchers who are doing the researches on web mining domain.


DOI 10.17148/IJARCCE.2016.5449

## REFERENCES

[1]. Dr. S. Vijiyarani and Ms. E. Suganya, "Research Issues in Web Mining" in International Journal of Computer-Aided Technologies Vol.2, No.3, July 2015.

[2]. Khushbu Patel, Anurag Punde, Kavita Namdev, Rudra Gupta, Mohit Vyas, "Detailed Study of Web Mining Approaches – A Survey" in International Journal of Engineering Sciences & Research Technology, February 2015.

[3]. Shipra Saini, "Review on Web content Mining Techniques" in International Journal of Computer Applications Vol 118, No.18, May 2015.

[4]. Aarti M.Parekh, Anjali S.Patel, Sonal J. Parmar, Prof. Vaishali R. Patel, "Web usage mining: Frequent Pattern Generation using Association Rule Mining and Clustering" in International Journal of Engineering Research & Technology, Vol 4, Issue 04, April 2015.

[5]. Anitha Talakokkula, "A survey on Web Usage Mining, Applications and Tools", in Computer Engineering and Intelligent Systems, Vol.6, No.2, 2015.

[6]. Nazneen Tarannum S.H. Rizvi and Prof. Ranjit R. Keole, "A Preliminary Review of Web-Page Recommendation in Information Retrieval using Domain Knowledge and Web Usage Mining" in International Journal of Advance Research in Computer Science and Management Studies, Vol.3, Issue 1, January 2015

[7]. Pooja Kherwa, Jyotsna Nigam," Data Preprocessing: A Milestone of Web Usage Mining" in International Journal of Engineering Science & Innovative Technology, Vol.4, Issue 2, March 2015.

[8]. Srividya, M.,D.Anadhi and M.I.Ahmed."Web mining and its categories-a survey."International Journal of Engineering and Computer Science, IJECS 2.4 (2013).

[9]. Sharma, Kavita, Gulshan Shrivastava and Vijay kumar. "Web mining: Today and Tomorrow". Electronics Computer Technology (ICECT), 2011 3rd International Conference Volume 1,IEEE, 2011.