Web Data Mining: Exploring Hidden Patterns, its Types and Web Content Mining Techniques and Tools

Harmeet Kaur, Sonal Chawla

Abstract- ***The abundance of web data has made it an utmost important source for Web data mining. Web data mining takes WWW data as input and after analysis and discovery, the output i.e. extracted information is used by an organisation. It helps the organisation in taking simpatico decisions for better survival in future. The objective of this paper is four folds. Firstly this paper gives a basic introduction of Web data mining. Secondly, it explains Web data mining categories, thirdly it discusses Web content mining techniques and tools in brief and finally a comparison between various tools available for Web Content Mining.***

Keywords: Web Content Mining, Structured data, unstructured data.

I. INTRODUCTION

In the present scenario if you are a naïve user or a computer tech expert, you will refer to Internet for flood of information. But sometimes the user is not satisfied with the outcome of his search and he himself has to do the detailed search by looking into all the web pages (most of which are irrelevant). Hence we need to manage the data on web to increase its efficiency or in other words, Web Mining is the need of time. The huge web data is unstructured and scattered. So before mining it, we need to convert the unstructured documents into some structured format. A large data set is required to explore the web precisely.

II. WEB DATA MINING OVERVIEW

The main task of Web data mining is to extract hidden patterns and relevant information from web data. Depending upon the nature of the data, the main areas to explore consist of the following three categories: Web content mining, Web structure mining and Web usage mining as shown in the figure 1.

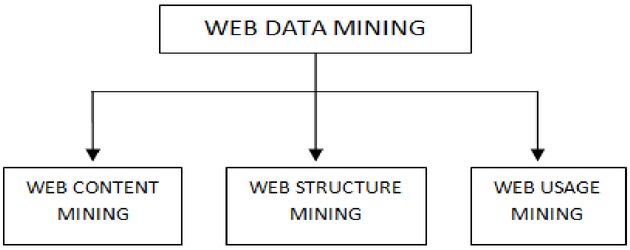

Figure1: Web Data Mining Types

Web content mining is used to mine the contents of the web documents. Content mining explores the text, multimedia contents like audio, video, images, which are embedded in or linked to the web pages. By this mining, the relevance of the content is checked and can be improved greatly. In Web structure mining, the structure and links of web documents are mined to extract useful information. The main purpose for structure mining is to extract previously unknown relationships between pages. This enables an organisation to link the information of its own website to navigate and cluster information into site maps. The hyperlinks present within a website provide useful information regarding the connection between different documents. The web can be considered as a directed graph whose nodes are the documents and the edges are the hyperlinks between them. The graph structure of web can provide a valuable source of information for various web mining tasks [1]. The data for Web usage mining are the user logs, which can be extracted from different sources like server level extraction, client level extraction and proxy level extraction. This consists of the following tasks as mentioned in fig 2.

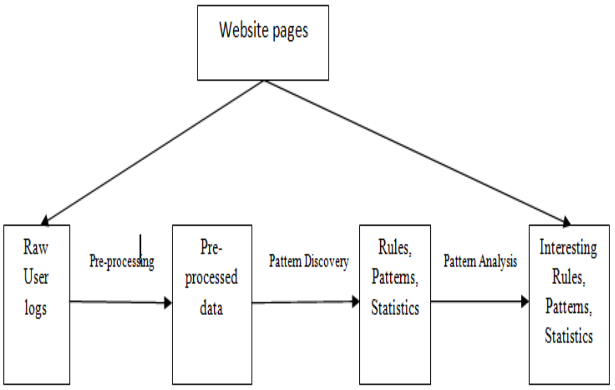

Figure 2: Web Usage Mining Process

Content Pre-processing: It consist of converting the image, text, scripts and other files like multimedia into firms that are useful for such mining. Usually this comprises performing content mining such as classification or clustering.

Pattern Discovery: It consists of methods and algorithm developed from several fields such as statistics, data mining, machine learning and pattern recognition.

Manuscript Received on December 2014.

Harmeet Kaur, Asst. Prof., MCM DAV College for Women, Chandigarh, Panjab, India.

Dr. Sonal Chawla, Chairperson, Department of Computer Science and Applications, Panjab University, Chandigarh, Panjab, India.

It comprises many techniques like statistical analysis, association rules, clustering, classification, sequential patterns, dependency modelling etc.

Pattern Analysis: The purpose of pattern analysis is to filter out interesting rules or patterns from the set found in the pattern discovery phase [2].

III. WEB CONTENT MINING TECHNIQUES

Conventional method of searching was purely based on content. But with the help of Web Content Mining, the search engines have become smarter now. The Web Content varies in three ways: Firstly, it could be unstructured such as free text or in semi structured form such as HTML documents or in purely structured form such as data in tabular form. Different technique is applied on each type of content [3].

3.1 UNSTRUCTURED DATA MINING TECHNIQUES

Most of the web pages are in the form of free text. In this technique the data is searched and retrieved. We further need to use some tools/techniques to squeeze out relevant data/information from that data.

3.1.1 TOPIC TRACKING

This technique tracks the topic of interest of each user and finds the related documents. So, the next time user logs in, advertisements/messages are displayed based on his history of surfing. Topic tracking can be applied in many areas. In the area of Medical, doctors can be updated with the latest treatments, symptoms, etc. In the area of education, new researches can be viewed. A Venture can keep a track on its rival, to keep himself on the edge in the market [3].

3.1.2 CATEGORIZATION

In this technique, the documents are placed into a predefined set of groups. Then it counts the number of words in the document, irrespective of its contents and ranks them. It decides the main topic from the count and the word with maximum count becomes first ranked [3].

3.1.3 CLUSTERING

Clustering is a technique to group objects into clusters based on some properties. Text based clustering is based on the content. If the contents are similar, this implies that the documents are pretty similar. In Graph based Clustering, the web documents can be considered as nodes and the edges will represent the relationship between these nodes. The edges carry a weight, which denotes the degree of that relationship [4].

3.2 SEMI-STRUCTURED DATA MINING TECHNIQUES

Semi-structured data is the loosely structured data like XML based web pages. This type of data arises when source or the environment does not impose a rigid structure on the data and when data is combined from several heterogeneous sources. Semi-structured data has the following properties:

Records do not necessarily have the same number of fields and fields can be different.

Fields do not have to be in a specific order.

The need for mining semi-structured data comes from the growing number of sources of semi-structured data. E.g., in case of Internet we want to know how users use the website and how its usability can be increased [5]. Following is a technique to mine the semi-structured data:

3.2.1 STORED

STORED i.e. Semi-structured TO Relational Data is a declarative query language. This technique uses RDBMS to store, query & manage semi-structured data. A Relational schema is chosen, and then the stored mapping translates the semi-structured data instance into that schema [6].

3.3 STRUCTURED DATA MINING TECHNIQUES

Structured data refers to the information with higher degree of organization. It contains the data records and presented as web pages depending on some template, for ex. in the form of table or form.

3.3.1 INTELLIGENT WEB SPIDERS

Web Spiders, also known as Crawlers, crawls across the www. Web Crawler creates a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide faster search. Crawlers can also be used to automate the maintenance of web sites like deleting the obsolete hyperlinks, verifying HTML code, etc. Web spiders can be used for building up search databases, personal searches, web site backup etc [7].

IV. WEB CONTENT MINING TOOLS

With the abundance of web data, web content mining tools help us to extract hidden patterns or information in an easy and systematic manner. Following are some of the Web Content Mining tools available:

4.1FMiner: FMiner is an easy to use web data extraction tool with an intuitive visual project design tool. Simply select your output file format and record your steps on FMiner as you walk through your data extraction steps on your target web site. FMiner's powerful visual design tool captures every step and models a process map that iterates through the target site pages to capture the information you've identified [8].

4.2Screen Scraper: Screen Scraper extracts data from websites and delivers it to the user in any format like Excel sheet, XML, database or Website. It can handle any type of website, including sites that use AJAX & the scraping process can run for weeks or months without interruption. It can be used to scrape medical data, financial data, e-commerce data, real estate data & many other types of data [9].

4.3RapidMiner: RapidMiner provides software, solutions & services in the field of advanced analytics including predictive analysis, data mining & text mining. It automatically analyses data including database & text on a large scale [10].

4.4Web Content Extractor: Web Content Extractor is a professional web data extraction software designed to greatly increase productivity & effectiveness of the web data scraping process. It offers a friendly wizard driven interface without any hassle of code. The extracted data can be exported to any format [11].

4.5Mozenda: Mozenda is a faster data extraction tool with a very easy to use interface. It provides an cloud computing environment in which you can scrape, store & manage your data from any machine at any point of time. With Mozenda's optional anonymous proxy feature user can utilise thousand of IP addresses to successfully gather information [12].

V. COMPARISON OF WEB CONTENT MINING TOOLS

The comparison of the five tools is done on five features:

Table1: Comparison of Web Content Mining Tools

TOOL	FEATURE				
	CLOUD COMPU-TING	USER FRIENDLY	AUTOMATI-ON OF EXTRACTI-ON	DATA PUBLIS-HING	ANONYMO-US PROXY FEATURE
SCREEN SCRAPER	Y	Y	Y	Y	Y
MOZENDA	Y	Y	Y	Y	Y
RAPIDMINER	Y	Y	Y	Y	N
WEB CONTENT EXTRACTOR	N	Y	Y	Y	Y
FMINER	N	Y	Y	Y	N

The above table signifies that Screen Scraper and Mozenda have all the five features. Whereas Rapidminer does not provide anonymous proxy feature. The Web Content Extractor is lacking in cloud computing feature. The FMiner is the most short of features as it is unable to provide cloud computing and anonymous proxy feature.

VI. CONCLUSIONS

In this paper we discussed Web data mining. Based on types of data, we explained its types: Web Content mining, Web structure mining and Web usage mining. After this, we discussed various techniques for Web content mining followed by web content mining tools. From the above discussed tools, Screen Scraper and Mozenda come out to be the best tools with respect to the five features mentioned above. To further bifurcate our conclusion, it came out that Mozenda is more user friendly than Screen Scraper. Also Mozenda is easy to learn and provides more ways of customer support (like phone, email & video chat). Hence Mozenda is the finest tool for web content mining.

REFERENCES

1. Johannes Furnkranz, Web Structure Mining Exploiting the Graph Structure of the World Wide Web.
2. Jaideep Srivastava, Robert Cooley, Mukund Deshpandey, Pag Ning Tan, Web Usage Mining: Discovery and Application of Usage patterns from Web Data, ACM SIGKDD Explorations Newsletter, January 2000, Volume 1 Issue.
3. Faustina Johnson & Santosh Kumar Gupta, Web Content Mining Techniques: A Survey, International Journal of Computer Applications (0975 – 888) Volume 47– No.11, June 2012 44
4. Magdalini Eirinaki, Web Mining: A Roadmap.
5. Mining Semi-Structured Data Theoretical and Experimental Aspects of Pattern Evaluation E.H. de Graaf
6. Alin Deutsch, Mary Fernandez, Dan Suciu, Storing Semi-structured Data with STORED
7. Govind Murari, Upadhyay, Kanika Dhingra, Web Content Mining: Its Techniques and Uses, Volume 3, Issue 11, November 2013 ISSN:2277 128X
8. www.fminer.com
9. www.screen-scraper.com
10. www.rapidminer.com
11. web content extractor help
12. www.Mozenda.com