Volume 3, Issue 11, November 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Web Content Mining: Its Techniques and Uses

Govind Murari Upadhyay, Kanika Dhingra
(Assistant Professor)
IITM, Janakpuri, New Delhi, India

Abstract: The focus of this paper is to bring in light the value of Web Content Mining. The paper gives an insight into its techniques, processes and its applications in the current cut-throat business environment as well in research and extracting contents for educational purposes. It further explains how using web content mining plays an integral role by getting rich set of contents and uses those contents in the decision making in the corporate environment, education and research.

Keywords- Content mining, www, extraction pattern, text mining

I. INTRODUCTION

The World Wide Web has lot of information and continues to increase in size and complexity. It is very herculean task to search relevant information from huge amount of data. The data used for web content mining includes both text and graphical data. Content mining is divided into two parts, one is webpage content mining and other is search result mining. In webpage content mining web is search via content. The search result content mining searches from the previous search result. When you search any specific key word or any web page, number of links or result is displayed. But all the data which is displayed on the web is not relevant. So efficiently and effectively retrieve required data on the Web is becoming a challenge. The user issues the query terms (keywords) to a search engine and the search engine returns a set of pages that may be related to the query topics or terms. For a page, if the user wants to search the relevant pages further, he/she would prefer those relevant pages to be at hand. Here, a relevant Web page is the one that addresses the same topic as the original page, but is not necessarily semantically identical. On web data is updated at every second so it is not necessary that a data or the web page that is retrieved by the user will be retrieved another time in the same structure or order. The relevant data can be retrieved by some specific techniques, those are:

- Web Content Mining:
- Web structure Mining
- Web usage Mining

Let us have a bird's eye-view on each of the above three mining techniques. Later on we will focus on the web content mining, its significance and features in this research paper.

- Web Content Mining:

Web content mining is the mining, extraction and integration of useful data, information and knowledge from Web page content. It describes the discovery of useful information from the web documents. In web content mining the content may be text, image, audio, video, metadata and hyperlinks etc. Web content mining also distinguishes personal home pages with other web pages. Research in web content mining encompasses resource discovery from the web, document categorization and clustering, and information extraction from web pages.

- Web Structure Mining:

This kind of mining emphasizes on the data which describes the structure of the content. It is classified into two types namely intra-page structure and inter-page structure. Intra-page structure means the existence of links within a page. No separate page will be opened in this case. Inter-page structure involves the connection of one page with the other page.

- Web Usage Mining:

It refers to the discovery of user access patterns from the web usage logs. It focuses on various data mining techniques to understand and analyze search patterns.

II. WEB CONTENT MINING TECHNIQUES

The focal point of this research paper shall be "WEB CONTENT MINING".

The concept of "WEB CONTENT MINING" involves techniques for summarizing, classification and clustering of the web contents. It can provide useful and interesting patterns about user needs and contribution behavior. *It* targets the knowledge discovery, in which the main objects are the traditional collections of text documents and, more recently, also the collections of multimedia documents such as images, videos, audios, which are embedded in or linked to the Web pages. It is mainly based on research in information retrieval and text mining, such as information extraction, text classification and clustering, and information visualization.

Some of the prominent web content mining techniques are as follows:-

A. Unstructured data mining techniques
B. Structured data mining techniques
C. Semi structured data mining techniques
D. Multimedia data mining techniques

A. Unstructured data mining techniques

One of the techniques for web content mining is unstructured. Number of the web pages is in the form of text. According to this technique the data is searched and retrieved. It is not necessary that the data which is retrieved is meaningful data, it may be unknown information. We have to use some tools or techniques to get relevant data/ information from that data.

- **Text Mining for Web Documents:-**

Text Mining is a subset of the domain of data mining techniques. Retrieval of information from HTML web pages in itself is a challenging task. This is due to the fact that HTML web pages have multiple tags which are required to identify information and secondly because the web pages are highly unstructured. The rich variety of tags may pose a problem in case they are not processed correctly. **However, there is no need to worry. With the advent of modern tools like IEPAD, SVM, Decision trees the results that are coming out is of much high accuracy.**

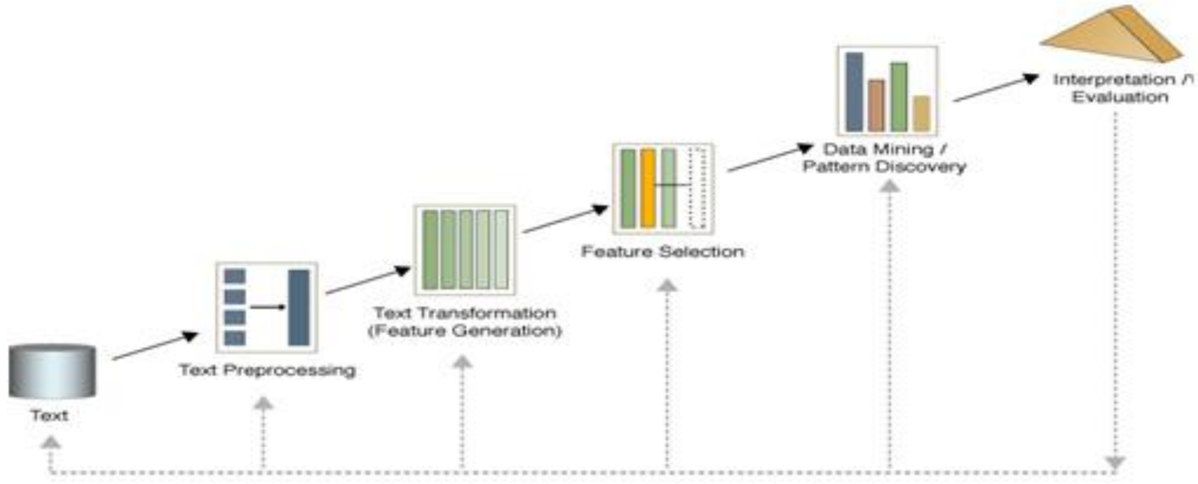

Fig: Text Mining for Web Documents

- **Topic Tracking**

Topic tracking is the technique by which a registered user can track the topic of his/her interest. He/she have to register with the topic, whenever any of the update regarding the interest of the user he will intimate by the message. Let suppose we register with any provider company then any of the relevant our area is determined informed to us. Same to other fields like as medical if there is any new research come in existence the information is send to the doctors. Other example for topic tracking is that if we select the competitors name then if at anytime their name will come up in the news then this information will be passed to the company. But this technique have some limitations like, if I am getting the information from the job portal related to computer science, but the updates may come from the other fields like as mathematics etc. for the same kind of job.

B. Structured data mining techniques

Structured data extraction is a progress of extracting information from web pages. A program for extracting such data is usually called a wrapper. Structured data are typically the data records retrieved from underlying database and displayed in the web pages following some templates. Sometime, the template is a table. Sometime, it is a form. Extracting such data records is useful because it enables us to obtain and integrate data from multiple sources (Web sites and pages) to provide value-added services, e.g., customizable Web information gathering, comparative shopping, meta-search, etc.

- **Intelligent Web Spiders:-**

Web spiders are prominently known as crawlers which look for the information across the WWW. Web crawlers are mainly used to create a copy of all the visited pages for later processing by a search engine that will index the downloaded pages to provide fast searches. Crawlers can also be used for automating maintenance tasks on a Web site, such as checking links or validating HTML code. Spiders use different algorithms like breadth-first search, genetic algorithms etc to explore for information. Spiders have number of applications like building up search databases, personal searches, web site backups etc.

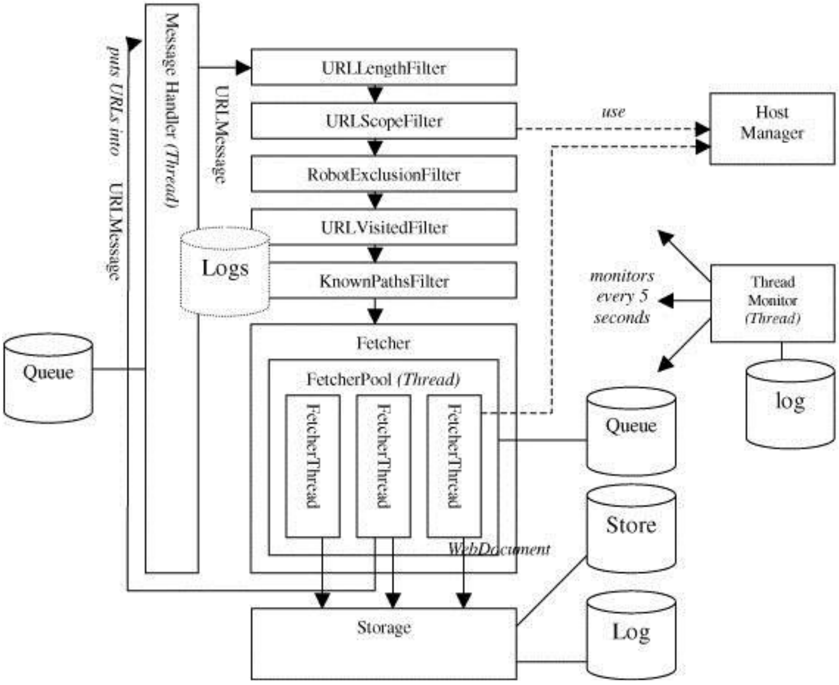

Fig2: Intelligent web spiders

C. Semi structured data mining techniques

Semi-structured data is a point of convergence for the Web and database communities: the former deals with documents, the latter with data. The form of that data is evolving from rigidly structured relational tables with numbers and strings to enable the natural representation of complex real-world objects like books, papers, movies, etc., without sending the application writer into contortions. Emergent representations for semi- structured data (such as XML) are variations on the Object Exchange Model (OEM). In OEM, data is in the form of atomic or compound objects: atomic objects may be integers or strings; compound objects refer to other objects through labeled edges. HTML is a special case of such 'intra-document' structure [2]. Users not only query the data to find a particular piece of information, but he is also keen in knowing better understanding of the query. Because of this variety, semi - structured DBs do not come with a conceptual schema. To make these databases more accessible to users a rich conceptual model is needed. Traditional retrieving techniques are not directly applied on these databases.

D. Multimedia data mining techniques

Multimedia data mining can be defined as the process of finding interesting patterns from media data such as audio, video, image and text that are not ordinarily accessible by basic queries and associated results. The motivation for doing Multimedia data mining is to use the discovered patterns to improve decision making. Multimedia data mining has therefore attracted significant research efforts in developing methods and tools to organize, manage, search and perform domain specific tasks for data from domains such as surveillance, meetings, broadcast news, sports, archives, movies, medical data, as well as personal and online media collections. The main aspects of feature extraction, transformation and representation techniques. These aspects are: level of feature extraction, feature fusion, features synchronization, feature correlation discovery and accurate representation of multimedia data. Comparison of Multimedia data mining techniques with state of the art video processing, audio processing and image processing techniques is also provided [3].

III. Web Content mining tools

With the flood of information and data on the Web, the content mining tools helps to extract the essential information that one would require. Some of them are Screen-scraper, Automation Anywhere 6.1, Web Info Extractor, Mozenda, and Web Content Extractor.[6]

- **Screen-scaper** : Screen-scraping is a tool for extracting/mining information from web sites. It can be used for searching a database, SQL server or SQL database, which interfaces with the software, to achieve the content mining requirements. Screen-scraper present a graphical interface allowing the user to designate URL's, data elements to be extracted and scripting logic to traverse pages and work with mined data.[7]

- **Automation Anywhere 6.1 (AA)** : AA is a Web data extraction tool used for retrieving web data, screen scrape from Web pages or use it for Web mining. It is a nique SMART Automation Technology for fast automation of complex tasks.[8]

- **Web Info Extractor (WIE)** : This is a tool for data mining, extracting Web content, and Web content analysis. WIE can extract structured or unstructured data from Web page, reform into local file or save to database, place into Web server. It can deal with text, image and other link file.[9]

- **Mozenda** : This tool enables users to extract and manage Web data. Users can setup agents that routinely extract, store, and publish data to multiple destinations. Once information is in Mozenda systems, users can format, repurpose, and mash up the data to be used in other applications or as intelligence.[10]

- **Web Content Extractor (WCE):** WCE is a powerful and easy to use data extraction tool for Web scraping, data mining or data extraction from the Internet. It offers a friendly, wizard-driven interface that will help through the process of building a data extraction pattern and creating crawling rules in a simple point-and click manner. This tool allows users to extract data from various websites such as online stores, online auctions, shopping sites, real estate sites, financial sites, business directories, etc. The extracted data can be exported to a variety of formats, including Microsoft Excel (CSV), Access, TXT, HTML, XML, SQL script, MySQL script and to any ODBC data source. It helps to extract/collect the market figures, product pricing data, or real estate data. [11]

IV. Uses of Web Content Mining

Following are the uses of Web Content Mining:

To gather, categorize, organize and provide the best possible information available on the WWW to the user requesting the information. To determine the relevance of the content to the search query. Improve the navigation of information on the web provides productive marketing. Produce a higher quality of information to the user. Understand customer behavior, evaluate effectiveness of a particular web site, and help quantify the success of a marketing campaign. Business intelligence. Competitive intelligence. Pricing analysis. Product data. Reputation.

V. **Conclusion**

The World Wide Web is the universe of network-accessible information, an embodiment of human knowledge. The web continues to increase in size and complexity with time hence making it difficult to extract relevant information. Thus various Data mining techniques and web content mining tools are used to extract useful information or knowledge from web page contents. By these techniques we can make our search of contents over the web faster and exact. This paper focuses on web content mining tools, techniques and uses of Web Content Mining.

References:

1. Faustina Johnson and Santosh Kumar Gupta Web Content Mining Techniques: A Survey. International Journal of Computer Applications (0975 – 888) Volume 47– No.11, June 2012
2. V. Bharanipriya & V. Kamakshi Prasad WEB CONTENT MINING TOOLS: A COMP ARA TIVE STUDY.
3. Chidansh Amitkumar Bhatt, Mohan S. Kankanhalli Multimedia data mining: state of the art and challenge. Journal Multimedia Tools and Applications archive Volume 51 Issue 1, January 2011.
4. Syed Salman Ahmed, Zahid Halim, Rauf Baig, and Shariq Bashir Web Content Mining: A Solution to Consumer’s Product Hunt, International Journal of Social and Human Sciences 2 2008.
5. Faustina Johnson and Santosh Kumar Gupta Web Content Mining Using Genetic Algorithm Advances in Computing, Communication, and Control Communications in Computer and Information Science Volume 361, 2013, pp 82-93
6. Abdelhakim Herrouz, Chabane Khentout, Mahieddine Djoudi Overview of Web Content Mining Tools Volume 2 , 2013.
7. Screen-scraper, http://www.screen-scraper.com Viewed 19 February 2013.
8. Automation Anywhere Manual. AA, http://www.automationanywhere.com Viewed 06 February 2013.
9. Zhang, Q., Segall, R.S., Web Mining: A Survey of Current Research, Techniques, and Software, *International Journal of Information Technology & Decision Making. Vol.7*, No. 4, pp. 683-720. World Scientific Publishing Company (2008).
10. Mozenda, http://www.mozenda.com/web-mining-software Viewed 18 February 2013.
11. Web Content Extractor help. WCE, http://www.newprosoft.com/web-content-extractor.htm Viewed 18 February, 2013.