ISSN:0975-9646

Rutvija Pandya et al, / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 6 (3) , 2015, 2256-2258

Elicitation of Information from Online Libraries using Web Content Mining Methods and Applications

Rutvija Pandya [#1], Jayati Pandya [#2]

[#1] *Diploma Computer Engineering Department, Gujarat Technological University*
Atmiya Institute of Technology and science, Rajkot, Gujarat, India
[#2] *K.P.Dholakiya Infotech College, Amreli, Gujarat, India*

***Abstract*— Web Data Mining is techniques that move towards web resources to collect required information. It allows users to utilize that mined information to promote their organizations. Web content mining helps user to extract useful information from the text, images and other forms of content. It extracts the features of a data. Labelling process is used to identifying and naming the attributes. Gained information can be used for the survey of the data. Web content mining analyses customers' view on the data from various sources. In this paper, research represented on web content mining techniques which are used to retrieve pattern and different applications of web content mining. The paper shows some methods used for retrieve of data from online libraries.**

***Keywords*— Web Content mining, web document types, Mining techniques**

I. INTRODUCTION

Usage of Web is day by day increasing in human's life. To satisfy user's expectation, number of information also increases on web. Daily updation of web information is needed to fulfil the requirements of the user

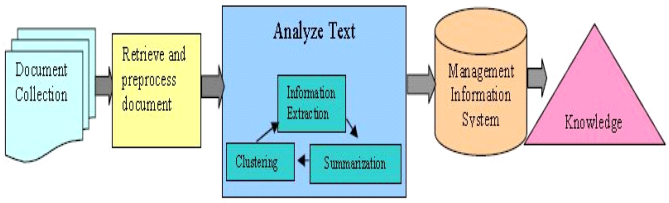

Fig 1 Process flow of Web Mining

Web Mining is based on knowledge discovery from web. Web mining is useful to extract the information, image, text, audio, video, documents and multimedia. Using web mining it becomes easy to extract all features and information of data from web. Web mining discovers the hidden data in web log and satisfies users with their required information.

II. WEB MINING CATEGORIES

There are three categories used for mining the web. These all categories are shown in the figure 2.

a) Web Content Mining
b) Web Structure Mining
c) Web Usage Mining

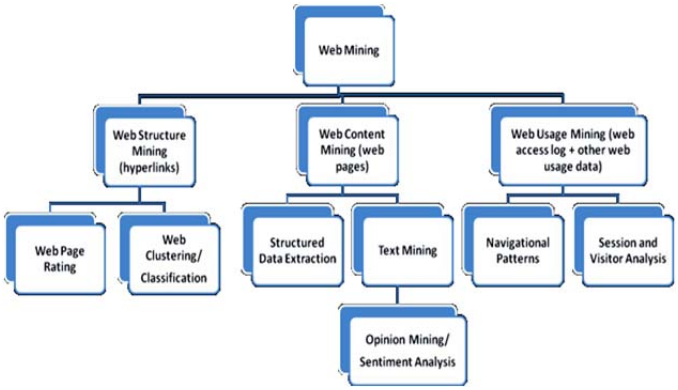

Fig 2 categories of web mining

A. *Web Content Mining*

Web content mining retrieves useful information from the web documents. It contains the generation of wrappers. Wrapper is a set of extraction rules to extract the data from the web pages, this can done either manually of automatically. [7] To get the information from web, user open many pages and these pages becomes the content of web. The accurate result is defined the result pages are content mining. The data may be integrated form of images, texts, audios or videos etc. Web content mining encloses the required data from the web, document classification and clustering, and extract the information from web pages.

B. *Web Structure Mining*

Web structure mining process emphasizes on structures information of web documents. The structure of the web document represents that using hyperlinks user can navigate on different content within the same page or user can navigate from one page to the other page.

C. *Web Usage Mining*

Web Usage Mining is used to find the pattern from user access data from the web usage logs. It is automatically generated by the data which is stored in server access logs, cookies, user profile, meta data, and site structure. Web usage mining technique is used to predict the user's behaviour and generate the meaningful pattern from that.

III. WEB CONTENT MINING TECHNIQUES

Some of the important web content mining techniques are:

a) Unstructured data mining techniques
b) Structured data mining techniques

c) Semi structured data mining techniques
d) Multimedia data mining techniques

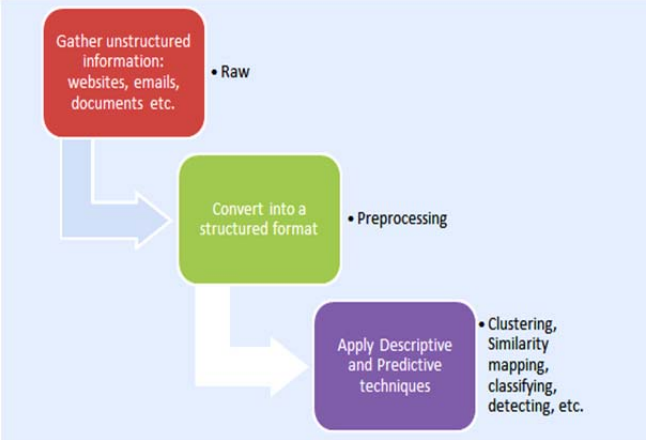

Fig 3 web content mining techniques

A. *Unstructured data mining techniques*

In Unstructured data mining technique web pages are in form of text. In this technique the data is searched and then retrieved data may be meaningful data or unknown information. From that data, some tools or techniques are used to get useful information.

1) *Text Mining for Web Documents*

Text Mining is a technique which is used to retrieve the information from web pages. To design HTML web pages multiple tags are used. These tags have the information within the page but this variety of tags may cause a problem if they are not processed correctly. These HTML web pages are highly unstructured. To get meaningful information from this unstructured collection of data with high accuracy, different tools like IEPAD, decision tree, etc are used.

2) Topic Tracking

In Topic tracking technique registered user can search the topic of their interest. If there is any new research regarding their register topic comes in existence, the information is send to them. Let suppose in online library system, if someone is registered with “text mining “topic then whenever any new content is arrived in library with this topic, user get messages. This technique has few limitations too. For example, if a user registered for “text mining”, user will receive many news stories on mining for minerals, and very few that are really on text mining.

B. *Structured data mining techniques*

For Structured data extraction from web pages, a program called wrapper is used. Structured data are the data records which are retrieved from database and displayed as table or form within web pages. Extraction of such data records is useful to obtain and integrate data from multiple sources.

1) Intelligent Web Spiders

A crawler is a program that reads the pages from different sites. It is also known as Spider. For providing fast searches, it creates a copy of all the visited pages by a search engine, and index that downloaded pages. This program is also used for maintaining the task of websites automatically. Spiders use different algorithms to search required information from World Wide Web. Spiders have different applications like building up search databases, web site backups etc.

C. *Semi structured data mining techniques*

This kind of data is neither raw data, nor organized data. Semi structured data on the web and databases often lack a regular structure. Files that are semi-structured may contain records, but not organized in a recognizable structure. Semi-structured data is a point of association for the Web and database communities: the first deals with documents, the later with data. [3] For semi-structured databases, the structure of an individual data item encodes an important part of its semantics. For frequent substructures important aspect of mining semi-structured databases are mining patterns and models in the structure of the data. Rather than using traditional retrieving techniques, Rich conceptual model is needed to make these databases more accessible to users.

D. *Multimedia data mining techniques*

Multimedia data mining used to find useful pattern from multimedia data such as audio, video, image, etc. It is used to improve decision making for different domains such as meetings, broadcast news, sports, movies, medical data, as well as personal and online media collections. The multimedia mining involves two basic steps: Extract of required features from the data and Select data mining methods to identify the useful information. [1]

IV. Emerging Techniques Of Web Content Mining For Online Libraries

In earlier days the techniques used for the information extraction is based on the HTML document. To retrieve the information, Searching is done using tree methodologies. In an era of digital information and WWW's growing popularity, digital libraries offer a huge range of text and multimedia information. The amounts of digital sources of information are escalating and make accessing material easy for users. Geographically distributed users can access the contents of electronic libraries. This content includes networked text, digital books, images, maps, sounds, videos, scientific research, business, hypertext, and multimedia compositions.

The vast amounts of products are available in Online Library systems. Web content mining helps to extract required pattern from its features and different user’s view. This pattern is used in decision making and provides faster access to information. Online library can have unstructured, semi structured, structured or multimedia data. So it depends on data that which mining technique is used to get required pattern. If data is unstructured, Topic Tracking, Information Extraction, Information Visualization, etc technique is used. For semi structured data Object Exchange Model, Top down Extraction, etc is used and for multimedia data SKICAT, Multimedia Miner, Colour Histogram Matching, etc is used.

V. Applications Of Web Content Mining

Web content mining is used to maintain the large information of various fields. Web servers

provide the cloud to extract the information from different locations and make effective use of web mining .Online library systems use the web content mining to generate the pattern and extract the information. Online library systems also extract reviews of a user and this technique of data mining is known as Opinion mining. Web content mining effectively used in Web wide tracking. Web communities can be maintained such as face book. The users of same field of interest can be grouped and they can communicate. Web mining is useful to understand the customer's behaviour. Now days, It is important to maintain the private information and for that Web mining is used.

Web mining can be used in e-learning environment. Using web mining techniques, the process of learning in e-learning environments is improved. Digital libraries are used to distribute valuable information all around the world, and eliminate the necessity to be physically present at different libraries. Digital library performs automated citation indexing using web mining techniques. E-services include e-banking, search engines, on-line knowledge management, social networking, e-learning, blog analysis, and personalization systems. [4]. E-Government is also one of the application fields of web mining. Government affairs can be carried out rapidly over the web. To provide simplicity and speed, Electronic government is one of the most appropriate applications of data mining. The policies of government derived from web content mining can be evaluated in time.

VI. Conclusions

Vast amount of data which is maintained by the web sources can be extracted by the web mining techniques with accuracy and as per the user's requirement. Web mining is a revolution from the problems or difficulties faced in data mining. In this paper, three standard categories of web data mining are introduced. A clear idea about different web content mining techniques depends on type of web data. Online libraries can implement web content mining to generate effective patterns from different type of data for decision making. This paper also includes different applications of web data mining.

Acknowledgment

This research paper is not made possible without the help and support from GOD, Parents and Friends. Firstly I would like to thank GOD and then Parents for their unconditional guidance. Second, I would like to thank my friend for their encouragement for this research. This research gives us the better experience to grow our selves with knowledge. Finally. I sincerely thank all, who provide the advice and support.

References

[1] Sarla More and Durgesh Kumar Mishra "Multimedia Data Mining: A Survey" Pratibha international journal of science, spirituality, business and technology (IJSSBT), Vol. 1, No.1, March 2012

[2] Govind Murari Upadhyay, Kanika Dhingra " Web Content Mining: ItsTechniques and Uses" International Journal of Advanced Research in Computer Science and Software Engineering , Volume 3, Issue 11, November 2013

[3] Kohavi, R., Mason, L., Parekh, R., Zheng, Z. (2004) "Lessons and Challenges from Mining Retail E-commerce Data" Machine Learning, Vol. 57 No. 1-2, pp. 83-113

[4] Arvind Kumar Sharma, 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, Issue 8, October 2012

[5] Rajashree Shettar, Dr. Shobha G ,"Survey on Mining in Semi-Structured Data",IJCSNS International Journal of Computer Science and Network Security, VOL.7 No.8, August 2007

[6] Sovers Singh Bisht, Prof. (Dr.) Sanjeev Bansal, "Optimization of Web Content Mining with an Improved Clustering Algorithm", International Journal of Emerging Technology and Advanced Engineering, Volume 3, Issue 11, November 2013

[7] Ananthi.J "A Survey Web Content Mining Methods and Applications for Information Extraction from Online Shopping Sites" / (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 4091-4094

[8] Jaideep Srivastava, Prasanna Desikan, Vipin Kumar," Web Mining—Concepts, Applications, and Research Directions"

[9] Govind Murari Upadhyay, Kanika Dhingra," Web Content Mining: Its Techniques and Uses", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 11, November 2013

[10] Govind Murari Upadhyay, Kanika Dhingra," Web Content Mining: Its Techniques and Uses", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 11, November 2013

[11] K. Mohammad Mujahid, Mr. I.S.Raghuram, M. Niranjan Kumar, K.V. Chaitanya Krishna,T. Mohaneshwar," WEB MINING: DAY-TODAY", International Journal of Emerging Trends & Technology in Computer Science (IJETTCS), Volume 3, Issue 5, September-October 2014