International Journal of Innovative Research in Engineering Science and Technology

APRIL 2018 ISSN 2320 –981X

A Review Technique on Web Content Mining - Tools and Algorithms

[1]Ms.Sindhuja.P, [2]Mrs.D.Nithya, [3]Dr.S.Sivakumari

1. Department of Computer Science and Engineering,
Avinashilingam Institute for Home Science and Higher Education for Women,
Coimbatore, India
Sindhujap18@gmail.com

2. CORRESPONDING AUTHOR:
Mrs.D.Nithya
Asst. Professor, Avinashilingam Institute for Home Science and Higher Education for Women,
Coimbatore, India
nithya.apcse@gmail.com

2. CORRESPONDING AUTHOR:
Dr.S.Sivakumari
Prof.&Head, Avinashilingam Institute for Home Science and Higher Education for Women,
Coimbatore, India
hodcseau@gmail.com

Abstract:-Web is a group of inter-related files on one or more web servers. Web mining is the integration of information gathered by traditional data mining methodologies and techniques with information collected over the World Wide Web. Web mining is the application of data mining techniques to solve the problem of extracting useful information from web server. The web mining process can be classified into three types: web content mining, web usage mining and web structure mining. The objective of web content mining is to collect, classifies, establish and provide the best information available on web pages. Web data contains web pages, web logs, web links and objects on the web. Web mining is evaluated by using data mining techniques, namely classification, clustering, and association rules. This review focus on web content mining with its tools and algorithms.

Keywords: Web Mining, Web Content Mining, Structured Mining, Un-Structured Mining, Semi- Structured Mining

1. INTRODUCTION

The major data source in the world is the web. The key objective of web mining is to mine useful information from web data. The amount of information in web is enormous and also simply accessible. It is a multidisciplinary field which includes data mining, machine learning, natural language processing, statistics, databases, information retrieval, multimedia, etc. Web mining is essentially used to capture associated data, creating new data out of the related data, personalization of the data, learning about individual user's .To naturally discover and mine information from WW W Web mining uses data mining techniques [1]. Web mining helps to recognize customer performance and helps to calculate the performance of a web site. Web mining is classified into three types namely Web content mining, Web structure mining and Web usage mining [11]. The process of extracting or discovering useful information from web pages is called Web content mining. It includes image, audio, video and metadata. Web structure mining deals with the hyperlink structure of web. The process of

extracting useful information from server logs is called Web usage mining [12].The complete web mining is divided into five subtasks:

- **Resource Discovery**: It helps in retrieving services and different documents on web.
- **Information selection and preprocessing**: It automatically selects and preprocesses specific data from the web sources.
- **Generalization**: It discovers overall pattern at individual web sites as well as across multiple sites.
- **Analysis:** It validates and interprets the mined pattern [12].
- **Visualization**: It presents the result in visual and easy to understand way.

2. WEB MINING CLASSIFICATIONS

Web mining can be divided into three types depending on the type of data such as Web Structure, Web Content and Web Usage Mining. Fig. 1.1 depicts the classification of web mining. Web data is

- Web content –text, image, records, etc.
- Web structure –hyperlinks, tags, etc.
- Web usage –http logs, app server logs, etc.

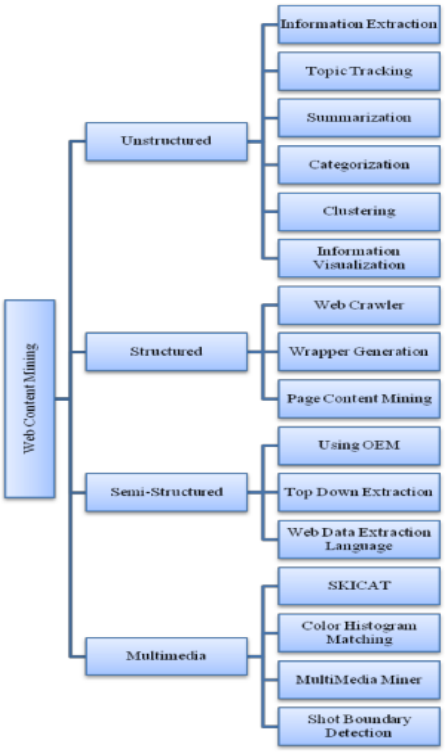

Fig. 1: Web Content mining

Web Content mining

Web Content mining refers to the discovery of useful information from the contents of the web data or documents [1]. Webpage can be in fixed text form or in the form of multimedia document containing table, form, image, video and audio. Web content mining identifies the useful information from the Web contents [10]. For example, web pages can routinely classify and group the web pages based on their topics. The data's in web content mining can be structure or unstructured documents.

Web Content mining can be differentiated into two types they are

1. Agent based approach and
2. Database approach.

Agent based approach transforms the data from semi and unstructured data into structured data [1]. Database approach follows the typical database querying method and data mining applications to examine the result [11].

Web structure mining

The objective of web structure mining is to generate structural summary regarding web pages and web sites. It is a mechanism used to identify the relationship between Web pages associated by direct link connection [3]. This structure data is discoverable by the delivery of webstructure plan through database techniques for Web pages.

Web usage mining

Web usage mining emphases explicitly on decision patterns concerning to consumers of a Web based system [8]. It manages with the detection of commercially important information in order to generate customized Web-pages.

3. WEB CONTENT MINING APPROACHES

The method of mining useful information from the contents of Web documents is known as Web Content Mining. Content records relates to the collection of facts a Web page was aimed to transport to the users. It consists of images, text, images, audio, video, or structured records such as lists and tables [3].

Two main approaches used in web content mining are namely Agent based approach and database approach. The three types of agents are intelligent search agents, Information filtering agent, and personalized web agents Intelligent Search agents [10]. Web content mining has the following approaches to mine data

1. Unstructured text mining,
2. Structured mining,
3. Semi-structured text mining, and
4. Multimedia mining

Unstructured Text Data Mining: Content mining can be organized on unstructured data such as text. Mining of unstructured data provides unknown data [1]. The study about applying data mining procedures to unstructured text is termed Knowledge Discovery in Texts (KDT). Some of the techniques used are

·Information Extraction,
·Topic Tracking,
·Summarization,
·Categorization,
·Clustering and
·Information Visualization

Structured Data Mining*:* The Structured data on the Web denotes their host pages. When compared to unstructured texts structured data is cooler to extract [1]. The techniques used for mining structured data are

·Web Crawler
·Wrapper Generation
·Page content Mining

Semi-Structured Data Mining*:* Semi-structured data growing from structured relational tables with numbers and strings to allow the natural representation of complex actual world objects without sending the application writer into expressions [3]. The techniques used for semi structured data mining are as follows

·Object Exchange Model (OEM),
·Top Down Extraction, and
·Web Data Extraction language.

Multimedia Data Mining: The techniques of Multimedia data mining are namely [1];

·SKICAT,
·Color Histogram Matching,

·Multimedia Miner and
·Shot Boundary Detection.

4. WEB CONTENT MINING ALGORITHMS

The two general tasks involved in web mining over which useful information can be extracted [12]. They are Classification and clustering. The various classification algorithms used to obtain the information are:

i) Decision Tree: The decision tree is one of the dominant classification techniques. It takes the input as its structures and the output as decision, which represents the class information. The system can handle multiclass classification difficulty [12].

ii) k-Nearest Neighbor: KNN is measured among the oldest nonparametric classification algorithms.

iii) Naive Bayes: Naive Bayes is an effective classifier built upon the principle of Maximum A Posteriori (MAP).

iv) Support Vector Machine: The most healthy and successful classification algorithms is Support Vector Machines [12]. It is a new classification method for both linear and nonlinear data.

v) Neural Network*:* The popular neural network algorithm is back propagation which achieves learning on a multilayer feed forward neural network. It contains an input layer, one or more hidden layers and an output layer.

5. WEB CONTENT MINING TOOLS

Web content mining tools comforts to download the necessary information. Some of them are Screen-scraper, Automation Anywhere, Web Info Extractor, Mozenda Web Content Extractor and Rapid Miner [2].

Screen-Scraper: Screen-Scraping is a tool for mining facts from web sites. It can be used for investigating a database, SQL database, which interfaces with the software [3]. The programming languages such as Java, .NET, PHP, Visual Basic and Active Server Pages (ASP) can be used to contact screen scraper.

Automation Anywhere: It is a Web data mining tool used for improving web data, screen scrape from Web pages [14].

Web Info Extractor: This tool is used for data mining, mining Web content, and Web content exploration [2]. It can mine structured or unstructured data from Web page, change into local file or save to database, set into Web server.

Mozenda: This tool allows users to mine and attain Web data. Operators can design agents that normally mine, store, and publish data to numerous destinations [11]. The two parts of Mozenda's scraper tool are:

i. Mozenda Web Console: It is a Web application that allow user to run agents, view and establish results [15].

ii. Agent Builder: It is a Windows application used to practice data mining project.

Web Content Extractor: It is easy to use data mining tool for Web scraping, data extraction from the Internet [6]. This tool allows users to extract data from various websites such as online stores, etc [13].

Rapid Miner: Rapid Miner is open source software for mining data from web, Comprises inbuilt algorithm [10]. It can produce algorithm by itself.

CONCLUSION

This paper discusses the techniques tools and algorithms of web content mining. Web content mining has been proven as very useful in the business world. Web content mining can also be practical to business use like mining online news site and developing a suggestion system for distance learning. There are many concepts available in web content mining but this paper tried to expose various web content mining strategy and explore some of the techniques.

REFERENCES

[1]Faustina Johnson, Santosh Kumar Gupta Web Content Mining Techniques: A Survey *International Journal of Computer Applications (0975 – 888)Volume 47– No.11, June 2012.*

[2] Herrouz, A., Khentout, C., Djoudi, M. Overview of Visualization Tools for Web Browser History Data, IJCSI International Journal of Computer Science Issues, Vol.9, Issue 6, No3, November 2012, pp. 92-98, (2012).

[3]Anuragkumar, Ravi Kumar Singh *International Journal Of Engineering And Computer Science ISSN: 2319-7242 Volume 6 Issue 1 Jan. 2017, Page No. 20003-20006*
A Study on Web Content Mining
[4] Geeta R. Bharamagoudar, Shashikumar G.Totad*IOSR Journal of Computer Engineering (IOSRJCE) ISSN: 2278-0661, ISBN: 2278-8727 Volume 5, Issue 4 (Sep-Oct. 2012*Literature Survey on Web Mining*).*
[5] Ahmed, S. S., Halim, Z.,Blaig, R. and Bashir, S. 2008. Web Content Mining: A Solution to Consumers Product Hunt. International Journal of Social and Human Sciences 2, 6-11.
[6] Ajoudanian, S. and Jazi, M. D. 2009. Deep Web Content Mining.World Academy of Science, Engineering and Technology 49.
[7] Inamdar, S. A. and shinde, G. N. 2010. An Agent Based Intelligent Search Engine System for Web Mining. International Journal on Computer Science and Engineering, Vol. 02, No. 03.
[8] Bharanipriya, V. and Prasad, K. 2011. Web content Mining Tools: A Comparative study.International Journal of Information Technology and Knowledge Management.Vol. 4. No 1,211- 215.
[9] Dunham, M. H. 2003. Data Mining Introductory and Advanced Topics.Pearson Education.
[10] Nimgaonkar, S. and Duppala, S. 2012. A Survey on Web Content Mining and extraction of Structured and Semistructured data, IJCA Journal.
[11] B. Singh, H.K. Singh, 2010, "Web data Mining Research", IEEE International
Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1-10.
[12] R.Malarvizhi, K.Saraswathi International Journal of Computer Trends and Technology (IJCTT) – volume 4 Issue 8–August 2013 Web Content Mining Techniques Tools &
Algorithms – A Comprehensive Study.
[13] Web Content Extractor help. WCE ,http://www.newprosoft.com/web-content-extractor.htm Viewed 18 February 2013.
[14] Automation Anywhere Manual. AA, http://www.automationanywhere.comViewed 06 February 2013.
[15] Mozenda,http://www.mozenda.com/web- mining-software Viewed 18 February 2013.

IJIREST