IJARCCE

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

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

An Exploratory Review of Web Content Mining Techniques and Methods

Narendra Parmar[1], Dr. Vineet Richhariya[2], Jay Prakash Maurya[3]

Computer Science & Engineering Department, Lakshmi Narayan College of Technology, Bhopal, India[1, 2, 3]

Abstract: Web content mining is a branch of web mining which involves the task of searching in the internet, extracting valuable information and filtering related data. It uncovers the useful information from the internet web content, including text, audio, graphics, video and multimedia. Web content mining has become complicated when it has to mine structured, semi-structured, unstructured and multimedia data. Web data extraction is a field of research in web content mining has now become a specialized work for modern internet based business, data science and technology. In this paper, the latest review of web content mining, its properties, significance and importance in data science and data mining is discussed through a theoretical and analytical process. Finally, it is concluded with many future aspects and applications.

Keywords: Web Mining, Web Content Mining, Data Mining, Web Data Analysis.

I. INTRODUCTION

The World Wide Web has plenty of information and it is continuously increasing in volume and complexity. It is very extraordinary task to extract relevant information from enormous amount of web data [1]. The types of data used for web content mining contain both text and graphical based data. Web content mining is basically divided into two branches, first is webpage content mining and second is search result mining. The searching process is executed by content itself in webpage content mining. However, the search result content mining finds results from the previous searched results [1][2].

When some specific keyword or web page is searched, then the number of results and links are exhibited. However, all the data which is exhibited on the web is irrelevant. So retrieving the required relevant data efficiently and effectively on the Web is becoming a challenging task [2]. The web users search the query terms or keywords into a search engine and search engine exhibits a set of page links which is related to the query key or terms.

For a web page, if the user searches the relevant pages further, then user prefers those relevant pages. Here, the relevant web page is one which addresses the same previous topic as its original page, but it is not necessarily identical. Web data is continuously updating at every moment so it is obvious that the data or the web page which is gained by the user will be gained another time in the other structure and disorder. The relevant web data can be gained by some specific methods and techniques [2][3]; these are:

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

The figure 1 represents the web mining classification and its substructure.

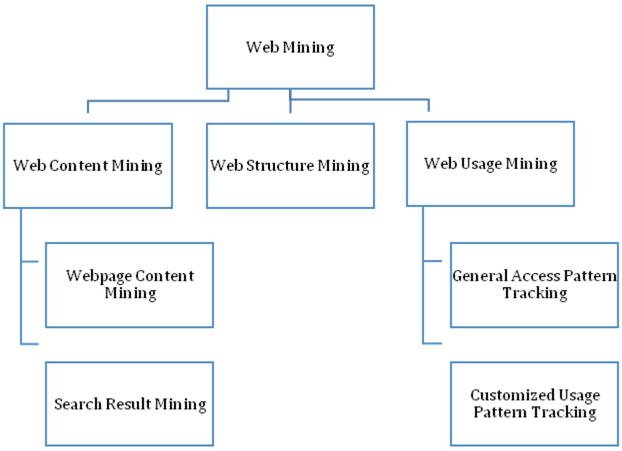

Figure 1: Web Mining Taxonomy

The web content mining as a part of web mining has many significance features with reference to business and statistical analysis.

Web Content Mining: It is the mining, data extraction and integration of valuable data, information and useful knowledge from the web page content [4]. It represents the discovery of valuable information from the web page documents [4]. In web content mining, the contents data are text, audio, video, graphics, metadata and hyperlinks. The web content mining also discriminates personal web home pages with other pages [4]. Research in the web content mining includes resource detection from the web, document classification and clustering, and extracted information from web pages [5].

Web Structure Mining: It emphasizes on the data that represents the structure of the web content. It is categorized into two types, first intra-page structure and second, inter-page structure [6]. Intra-page structure presents the existence of web links within the page itself

DOI 10.17148/IJARCCE.2016.5535

and separate web page will not be opened in this type. Inter-page structure includes the link of one web page within the other web page [6].

Web Usage Mining: The web content mining point out the detection and extraction of user access patterns and behavior from the web page usage logs [7]. It concentrates on several data mining techniques and methods to comprehend and analyze the search patterns. The result of web usage mining can be enhanced by analyzing web page content. Some system integrates the web page clustering into the search pattern log and the cluster labels are also used as a web page content extractor [7].

II. WEB CONTENT MINING TECHNIQUES

In this section, the main point is discussion of web content mining techniques. The concept of web content mining includes techniques and methods for summarizing, categorization and clustering of the web page contents [8]. It provides valuable and interesting patterns of user requirements and contribution formats. It focuses on the knowledge discovery from web pages, in which the main purposes are the traditional gathering of text documents and also the collections of multimedia documents like graphics, audios, videos and animations which are integrated in or linked into the web pages [8]. It is generally based on research in the information extraction and text mining, like text classification and clustering and the information visualization [9]. Some of the prominent web content mining techniques are as follows: -

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

A. Unstructured Web Content Mining Techniques: It is one of the technique for web content mining. Generally, many web pages are in the form of text and hypertext. According to unstructured data mining technique, the text data is searched and information is extracted [9]. It is not essential that the data which is extracted is a meaningful data, it may be some unknown information. Some efficient tools or techniques are required to get relevant information from web data. The components of unstructured web content mining techniques include information extraction, topic tracking, summarization, categorization, clustering and information visualization [9].
Text mining is a component of web content mining. For web based documents, text mining works as a subset of the field of data mining techniques. The information extraction from HTML web pages is a challenging task in the age of modern internet [6][10]. Because HTML web pages contain multiple tags which are needed to identify information data and also the web pages are usually very unstructured. The many variety of HTML tags can produce a trouble in case when they are not processed perfectly. With the arrival of modern tools and techniques like support vector machine (SVM) and decision trees, the extracted products which are coming out have much higher accuracy [10]. By topic tracking technique, a registered web user can track the keyword and topic of interest. The web user has to register with the selected keyword and topic, whenever there are any updates regarding the topic of interest of the user, then the user is implied by a message. For example, any registered user of a company gets informed by some information relevant to the topic or keyword of user data whatever the user searches out. Same condition is applied to other fields like medical science, pharmacy, engineering and business administration, if there is some new research emerges in existence, then these information messages are send to the related departments [10].

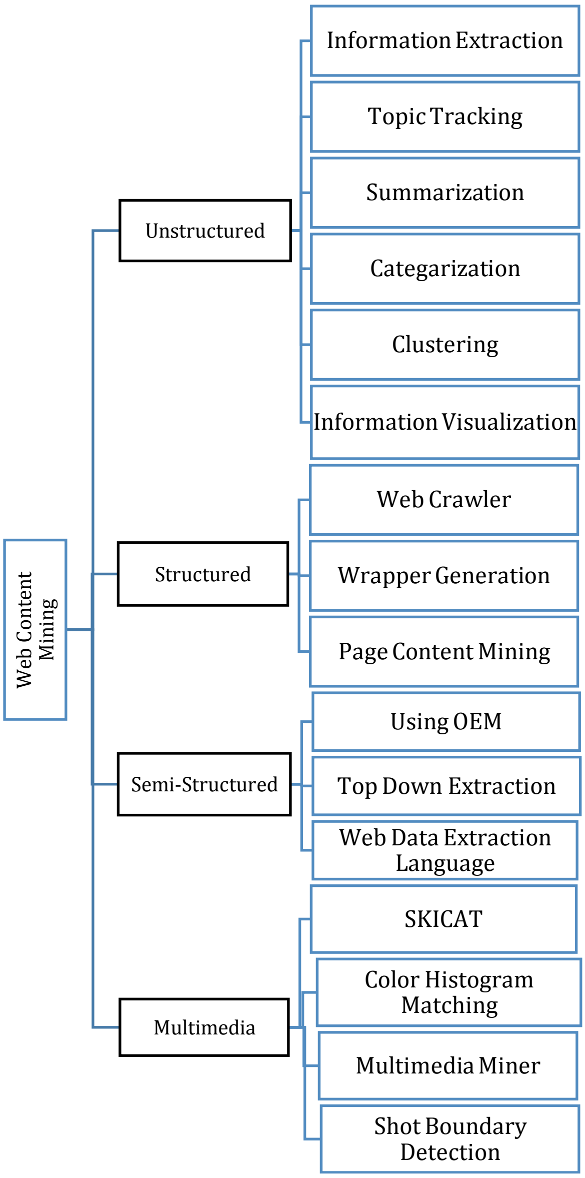

Figure 2: Techniques of Web Content Mining

B. Structured Web Content Mining Techniques: The structured data extraction is the development and progress of information extraction from web pages. The components of structured web content mining include web crawler, wrapper generation and page content mining [11].

The specified program which extracts such type data is basically called the wrapper. The structured data is usually the data records which are retrieved from the underlying database and visualized in the web pages using some themes and templates [6][11]. These templates are tables, forms and some predefined patterns or structures. Extraction of these data records are useful since it enables user to gain and combine data from multiple sources like web sites and pages. The intelligent web spiders are the components of web content mining also known as web crawlers which searches the information across the internet. The web crawlers are generally used to produce a copy of all visited web pages for later processing by the search engine which makes the index of downloaded pages to enable fast searches. Web crawlers can also be applied for automatic maintenance jobs on the web site, like link checking or HTML code validation [6][11]. Spiders use algorithms of various types such as genetic algorithm, breadth-first search, particle swarm optimization and other soft computing techniques to extract the information. Spiders contain many applications like building index of search databases [11].

C. Semi Structured Web Content Mining Techniques: The semi-structured data is central repository for the web databases. The central repository data is growing from tables of structured relational with numbers and strings for enabling the natural visualization of disorganized real-world objects such as papers, books, articles and videos without forwarding the application writer. Instant visualizations for semi-structured data like XML are variations on Object Exchange Model (OEM). Data is in the format of atomic and compound objects in OEM [12]. Atomic objects are integers or strings, in OEM, compound objects direct to the other objects with labeled edges. The web users query the data to search a specific type of information, also users learn better understanding of that query. Due to this variation, semi-structured databases do not arrive with the conceptual schema [12]. For designing these databases more accessible to users a rich conceptual model is needed. Traditional retrieving techniques are not directly applied on these databases [12].

D. Multimedia Web Content Mining Techniques: Multimedia data mining is the process of searching interesting valuable patterns from media data like graphics, audio, video, and text which are not generally accessible by the basic queries and its associated results. The reason for deploying multimedia web content mining is to use the extracted patterns to improve the decision making [11][12]. Multimedia web content mining has therefore paid attention to remarkable research undertakings in developing tools, methods and techniques to search, organize, manage and perform area specific tasks for web data from areas like movies, sports, surveillance, meetings, archives, broadcast news, medical data, personal data and online multimedia data collections. The prominent aspects of multimedia web mining techniques are feature extraction, transformation and visualization techniques. These aspects include level of feature extraction, feature synchronization, feature fusion, feature correlation detection and accurate visualization of multimedia web data. Comparison of multimedia web content mining techniques with research methods of audio processing, video processing and image processing methods and techniques is also provided [10][12].

III.LITERATURE REVIEW

In this section, the tools, techniques and methods with reference to web content mining are analyzed and reviewed. In web content mining, text is a type of data in which the attributes of word are defined in high dimensional attributes. It is complicated to apply the classification techniques to the text. The techniques which are generally used for web content or text classification are follows:

A. **Decision Tree Classification for Web Content Mining:**
The decision tree classification method is used for decision making planning and design and it is now extended to search web contents. Decision tree generally contains root node and branch node. Based on decision tree reasoning algorithm the users divide each node recursively from the root node of the tree [13]. The final conclusive result of decision tree contains branches and each branch illustrates a possible way of decision and its consequences. The decision tree algorithm verifies each partition and selects the best way. The complexity of this algorithm is minimized by pruning techniques like cutting the branches of tree. This searching algorithm also determine the attribute selection and classification and it requires the lowest amount of information. Finally, decision tree classification algorithm selects optimal split of tree, the one with the highest information benefit as a result [13].

B. **Neural Network Classification used in Web Content Mining:**
Neural networks models are extensively used for wide range of application and also used in a wide variety of areas for the classifications which are extended to search the web content effectively. Neural networks classifiers are implemented in the web content and text data with the use of word attributes. The fundamental unit in the neural network is an artificial neuron. Each neuron receives some set of inputs that are signified by the vector $\overline{Xn}$, which is related to the item frequencies for the nth document. In neural network, each neuron has its own set of weights which is used to calculate some function [13]. The linear function is determined as follows.

$$L_n = W.\overline{X_n}$$

Where, L = Linear function.
W = weight.
X_n = set of inputs.

C. **Naïve–Baye's Classification for Searching Web Content and Text:**
Naïve-Baye's classification model is also known as generative classification models used in the distribution of

specified documents into each class with the probabilistic access [15]. This approach is now extended in web mining technologies. The models of two classes are most commonly used model in Naïve-Baye's classification method. These models aim on posterior probability of the class, which is based on the probability distribution of the words within the document [12][15]. These models generally work with the huge data set of words; however, it ignores the real position of the words. Based on the occurrence word frequencies, the models can be differentiated [15].

D. **Support Vector Machine Classification for Web Content Mining:**

The support vector machine (SVM) is an accurate learning technique and method for many classification problems and this technique is now applied in the web mining process. This method tries to search an optimal hyper plane in the set of input space in order to search and categorize the web or text data into the binary form [13][14]. For hyper plane, the linear separable space is generally given by.

$$\mathrm{W.X} + \mathrm{b} = 0$$

Where X = the arbitrary object which is to be classified.
W = The vector which is learned from the training set of linearly separable data.
b = A constant which is also learned from the training set of linearly separable data.
The support vector machine (SVM) separates the positive and negative data training sets with the maximum margin.

E. **Association Rule Based Classification for Web Mining:**

In associative classification technique, the rules produced from association rule mining are translated into classification rules. The theory of association rule mining can be implemented into web mining environment to search associations between user web pages visited through the internet in their browsing sessions [13]. A weighted fuzzy association rule mining methods are able to search natural associations between the items by deliberating the importance of their presence in the transaction. Classification based on association generally combines classification and association rule mining [14][15]. The classification association rules are basically the association rules with the class on right side of rules and the conditions on left side of rules. These rules are achieved from available training data sets and exact association classification model is constructed. The classification based association is applied to classify web text documents into the topic hierarchies in web text classification [14][15]. These rules are achieved using the Apriori Algorithm and many final classes are derived from the association rule.

IV. PERFORMANCE MEASURE

The performance metrics that is used to evaluate the performance of web content mining. The first main metrics is throughput. The throughput is the total time required to execute web content data [9]. The other performance metrics for web content mining evaluation are:

Cost Performance: It is measured as the ratio of throughput to cost of web content mining.

$$\text{Cost Performance} = \frac{\text{Throughput}}{\text{Cost}}$$

Scale up: It is capability of the system to manage more web content mining data with integrating more computers while maintaining the performance.

$$\text{Scale up} = \frac{\text{Throughput After}}{\text{Throughput Before}}$$

Latency: It is time to execute web content mining data set of operations.

$$\text{Latency} = \frac{1}{\text{Throughput}}$$

Durability: It is the ability of the system to maintain the information for extensive time period.

$$\text{Durability Ratio} = \frac{\text{Current Reads}}{\text{Total Reads}}$$

Concurrency: It is the ability of the system to provide a service to different users at the same time.

$$\text{Concurrency Ratio} = \frac{\text{Successful Operations}}{\text{Total Operations}}$$

There are many evaluation metrics and models to measure the performance of web content mining and data execution.

V. PROBLEM IDENTIFICATION AND SOLUTION

The problems identified to analyze, process and execute web content mining are basically not the problem but some limitations and shortcoming associated with mining techniques. These limitations and shortcoming can be summarized as follows:

- Web content data sets are usually very large and it takes hundreds of terabytes (TB) size to store it on the database.
- Scaling problem for high dimensional web data.
- The large volume of web content data cannot be mined on a single ordinary server so large number of servers is required.
- It requires perfect organization of hardware and software to process and mine multi-terabyte of web content data sets.
- The limitations of sequential mining process and time series web data with respect to execution time.
- The problems in searching valuable relevant information from huge volume of web data set within a limited time.
- The web data is either over fitting or under fitting.

DOI 10.17148/IJARCCE.2016.5535

- The extraction of new knowledge information from the large web data.
- There is limited customization, limited query and limited customization interface to individual web users.
- The requirement of automated data cleaning.
- Imperfect sampling of web data.

The solution of the above mentioned limitations and shortcoming can be overcome by applying parallelization process to the algorithms, software technology and hardware. Parallelization of web content data can improve the throughput, cost performance and accuracy for huge volume of web content data and it is the fundamental requirement for future web data.

VI. CONCLUSION AND FUTURE RESEARCH

The web is the huge storage of network-accessible information, and knowledge. The web pages are continuously increasing in volume and complexity with time so it is going difficult to extract the valuable relevant information from internet. Thus several web mining techniques, methods and web content mining tools are applied to extract relevant useful information and knowledge from the web page contents. This paper reviews exploratory mining tools and techniques to mine the web contents in the internet. The analysis and theoretical review suggested the improvement of web mining algorithms. The parallelization process of huge volume of web data mining process can improve the performance in future. The parallelization process is the recommendation for future as the web data is continuously growing at rapid speed.

REFERENCES

[1] Brijendra Singh, Hemant Kumar Singh, “Web Data Mining research: A survey”, IEEE International Conference on Computational Intelligence and Computing Research (ICCIC), pp. 1-10, 2010.

[2] Petar Ristoski, Heiko Paulheim, “Semantic Web in data mining and knowledge discovery: A comprehensive survey”, Web Semantics: Science, Services and Agents on the World Wide Web, Elsevier, Vol. 36, pp. 1-22, Jan 2016.

[3] Jonathan Jeffrey, Peter Karski, Björn Lohrmann, Keivan Kianmehr, Reda Alhajj, “Optimizing Web Structures Using Web Mining Techniques”, Intelligent Data Engineering and Automated Learning – IDEAL, Springer, Vol. 4881, pp. 653-662, 2007.

[4] Kshitija Pol, Nita Patil, Shreya Patankar, Chhaya Das, “A Survey on Web Content Mining and Extraction of Structured and Semistructured Data”, IEEE First International Conference on Emerging Trends in Engineering and Technology, pp. 543 – 546, 2008.

[5] Jing Li, C. I. Ezeife, “Cleaning Web Pages for Effective Web Content Mining”, Database and Expert Systems Applications, Springer, Vol. 4080, pp 560-571, 2006.

[6] Krishna Murthy, Suresha, “XML URL Classification Based on their Semantic Structure Orientation for Web Mining Applications”, Procedia Computer Science, Elsevier, Vol. 46, pp. 143-150, 2015.

[7] Bhupendra Kumar Malviya, Jitendra Agrawal, “A Study on Web Usage Mining Theory and Applications”, Fifth International Conference on Communication Systems and Network Technologies (CSNT), pp. 935 – 939, 2015.

[8] Cheng Wang, Ying Liu, Liheng Jian, Peng Zhang, “A Utility-Based Web Content Sensitivity Mining Approach”, IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Vol. 3, pp. 428 – 431, 2008.

[9] Amit Dutta, Sudipta Paria, Tanmoy Golui, Dipak K. Kole, “Structural analysis and regular expressions based noise elimination from web pages for web content mining”, IEEE International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1445–1451, 2014.

[10] Loredana Caruccio, Vincenzo Deufemia, Giuseppe Polese, “Understanding user intent on the web through interaction mining”, Journal of Visual Languages & Computing, Elsevier, Vol. 31, Part B, p. 230-236, December 2015.

[11] Olatz Arbelaitz, Ibai Gurrutxaga, Aizea Lojo, Javier Muguerza, Jesús Maria Pérez, Iñigo Perona,, “Web usage and content mining to extract knowledge for modelling the users of the Bidasoa Turismo website and to adapt it”, Expert Systems with Applications, Elsevier, Vol. 40, Issue 18, pp. 7478-7491, 15 December 2013.

[12] YoonKyung Cha, Craig A. Stow, “Mining web-based data to assess public response to environmental events”, Environmental Pollution, Elsevier, Vol. 198, pp. 97-99, March 2015.

[13] Binu Thomas, G. Raju, “A Novel Web Classification Algorithm Using Fuzzy Weighted Association Rules”, ISRN Artificial Intelligence, Hindawi, Vol. 2013, Article ID 316913, pp. 10, 2013.

[14] Rong Qian, Kejun Zhang, Geng Zhao, “A topic-specific Web crawler based on content and structure mining”, 3rd IEEE International Conference on Computer Science & Network Technology (ICCSNT), pp. 458 – 461, 2013.

[15] Sotiris Kotsiantis, “Increasing the accuracy of incremental naive bayes classifier using instance based learning”, International Journal of Control, Automation & Systems, Springer, Vol. 11, Issue 1, pp 159-166, February 2013.

DOI 10.17148/IJARCCE.2016.5535