International Journal of Trend in Scientific Research and Development (IJTSRD)
Volume 3 Issue 5, August 2019 Available Online: www.ijtsrd.com e-ISSN: 2456 – 6470

Study on Web Content Extraction Techniques

Aye Pwint Phyu, Khaing Khaing Wai

Department of Information Technology Support and Maintenance,
University of Computer Studies, Mandalay, Myanmar

How to cite this paper: Aye Pwint Phyu | Khaing Khaing Wai "Study on Web Content Extraction Techniques" Published in International Journal of Trend in Scientific Research and Development (ijtsrd), ISSN: 2456-6470, Volume-3 | Issue-5, August 2019, pp.2235-2238, **https://doi.org/10.31142/ijtsrd27931**

IJTSRD27931

Copyright © 2019 by author(s) and International Journal of Trend in Scientific Research and Development Journal. This is an Open Access article distributed under the terms of the Creative Commons Attribution License (CC BY 4.0) (http://creativecommons.org/licenses/by/4.0)

ABSTRACT

Nowadays, the explosive growth of the World Wide Web generates tremendous amount of web data and consequently web data mining has become an important technique for discovering useful information and knowledge. Web mining is a vivid research area closely related to Information Extraction (IE). Automatic content extraction from web pages is a challenging yet significant problem in the fields of information retrieval and data mining. Web Content mining refers to the discovery of useful information from web content such as text, images videos etc. Web content extraction is the process of organizing data instances into groups whose members are similar in some way. Content Extraction helps the user to easily select the topic of interest. Web Content Ming technology is useful in management information system. Web content mining extracts or mines useful information or knowledge from web page contents. This paper aims to study on web content extraction techniques.

KEYWORDS: *Web Data Mining; Web Content Mining*

1. INTRODUCTION

The web is recognized as the largest data source in the world. The nature of such data is characterized by partial or no structure, and even worse there exist no standard data schema for the even low-volume structured data.

Web mining aims to extract useful knowledge from the web by using a variety of techniques that have to cope with the heterogeneity and lack of a unique and fixed way of representing information. Several data mining methods are used to discover the hidden information in the Web. However, web mining does not only mean applying data mining techniques to the data stored in the web. The algorithms have to be modified to better suit the demands of the web. New approaches should be used better fitting to the properties of web data. An important aspect in Web Content Mining is played by the automation of extraction rules with proper algorithms. Among many popular topics in web data mining, extracting information architecture or content structures for a web site has attracted many research attention in recent years. Automatic content extraction from web pages is a challenging yet significant problem in the fields of information retrieval and data mining.

The problem arises particularly on the World-Wide Web, because Web pages are often decorated with side bars, branding banners and advertisements [2]. Many important automated and manual text analysis tasks are handicapped by this mixture of core content and only peripherally related information on a single page. Web content mining extracts or mines useful information or knowledge from web page contents. Identifying the part of actual content, or clipping web pages, has many applications, such as high quality web printing, e-reading on mobile devices and data mining. Although there are many existing methods attempting to address this task, most of them can either work only on certain types of web pages, e.g. article pages or has to develop different models for different websites [3].

Clustering is the process of organizing data instances into groups whose members are similar in some way. A cluster is therefore a collection of data instances which are "similar" to each other and are "dissimilar" to data instances in other clusters. The k-means algorithm is the best known partition clustering algorithm. It is perhaps also the most widely used among all clustering algorithms due to its simplicity and efficiency. Given a set of data points and the required number of k clusters, k-means algorithm iteratively partitions the data into k clusters based on a distance function. The rapid growth of World Wide Web has been tremendous in recent years. With the large amount of information on the Internet, web pages have boosted the development of information retrieval and data mining applications. However, the web pages as the main source of data consists of many parts which are not equally important. Besides the main content, a web page also comprises of noisy parts such as advertisements, headers, footers, that can degrade the performance of information retrieval applications. Therefore, an approach to identify and extract main content is needed to alleviate this problem [4].

In order to improve the performance of web mining and information retrieval from web pages, content extraction techniques have been developed to remove such noise. Generally, content extraction improves performance, and is essential for many real world applications. With the ever changing style and design of modern web pages, different approaches are continually needed to keep up with the changes. Many researches in web mining proposed many methods to extract web contents, but they are fail to handle real-time dynamic data. There is no approach has managed

to achieve 100% accuracy in extracting all relevant text from websites so far. There is much work left to be done in the website content extraction.

2. Literature Review

The previous researchers proposed many methods or extraction of information from World Wide Web (WWW). In order to help analyze the content of web pages, many researchers have been developing methods to extract the desired information from a web page. The research papers studies a set of problems that are faced during web data extraction. Web content extraction algorithms are important to extract useful contents from web sources.

The authors [5] proposed a method, Paragraph Extractor (ParEx), clustering HTML paragraph tags and local parent headers to identify the main content within a news article. Their research presented a new method called ParEx to evaluate the content within a website and extract the main content text within. Building upon previous work with the text-to-tag ratio and sliding window clustering techniques, the presented ParEx method focused only on the paragraph tags of each website, making the assumption that most websites will have their main article within a number of *p* tags. Two primary requirements were found to optimize the success of ParEx. Firstly, websites must use *p* tags to store their article content. Secondly, websites that have limited or no user comment sections.

The authors [6] proposed a method for web content extraction that consists of web document selection phase, web cube creation phase, web document pre-processing phase and presentation phase. In their paper, they primarily focus on mining useful information from the web content data. In particular, they consider the issues of web content mining in the web information repositories context: in relational databases, the data are structured and are very well arranged in table using set of attributes and rows. In case of web repositories, web documents are unstructured. It is not that much easy to apply data mining system directly to search the entire WWW to discover required knowledge based on user query. In web content mining, a list of web documents are selected from WWW to select useful information. In their model, they mine based on key words, which are present in the documents.

The authors [7] proposed a method which gives the informative content to the user using DOM tree approach. By using DOM tree, their method intends to filter out non-informative content from the web pages. Their proposed approach concentrates on web pages where the underlying information is unstructured text. The technique used for information extraction is applied on entire web pages, whereas they actually seek information only from primary content blocks of the web pages.

The authors [8] also proposed a hybrid approach which is a combination of content extraction and rule generation and it is applied to extract informative content. Their approach involves generation of automatic rules instead of manual hand-crafted rule insertion. The rules generated are used to infer informative content from simple HTML pages. Initially DOM tree is constructed to demonstrate a visual content of the Web page with richer features. Feature extraction is applied between and <td> tags. Machine learning methods like Decision tree classification and Naïve Bayes classification are applied to generate the rules and create a well formed document. Rules generated are used for extracting the informative content from the Web pages.

The paper [9] proposed a semi-supervised web news extraction technique that uses unsupervised clustering technique and supervised classification technique. Their proposed technique is implemented in MATLAB after clustering process. Their results have been built on the basis of confusion matrix generated by SVM and then recall, precision, accuracy, and F1-Measure is calculated. Their proposed method is for extracting news from the internet web pages by using the idea of clustering neural genetic approach. Extracting web news content is some other process for examining a pattern in the web information that is communicated as regular appearance in the contribution of web pages.

The authors [1] proposed the automatic identification of informative sections of web pages. Here four simple yet powerful algorithms called Content extractor, Feature Extractor, K-Feature Extractor an L-Extractor were proposed to identify and separate content blocks from non-content blocks. Feature Extractor is based on the characterization and uses heuristics based on the occurrence of certain features to identify content blocks. K-Feaure Extractor is a special modification of Feature Extractor which perform better in a wide variety of web pages. Content Extractor identifies non-content blocks based on the appearance of the same block in multiple web pages. L-Extractor uses various block features and train a support vector (SV) based classifier to identify an informative block versus a non-informative block. First, the algorithm partition the web page into blocks based on heuristics. Second, the algorithm classifies each block as either a content block or non-content block. It has the advantage that both K-Feature Extractor and Content Extractor produce excellent precision and recall values and runtime efficiency. It also reduces the complexity and increases the effectiveness of the extraction process. It has the disadvantage that it will increase the storage requirement for indices and the efficiency of the markup algorithm are not improved.

3. Web Content Extraction Techniques

Traditional technique of searching the web was via contents. Web Content mining is the extended work performed by search engines. Web Content mining refers to the discovery of useful information from web content such as text, images videos etc. Two approaches used in web content mining are Agent based approach and database approach. The three types of agents are Intelligent search agents, Information filtering/Categorizing agent, Personalized web agents. Intelligent Search agents automatically searches for information according to a particular query using domain characteristics and user profiles. Information agents used number of techniques to filter data according to the predefine instructions. Personalized web agents learn user preferences and discovers documents related to those user profiles. In Database approach it consists of well-formed database containing schemas and attributes with defined domains. Web content mining becomes complicated when it has to mine unstructured, structured, semi-structured and multimedia data. Web content mining identifies the useful information from web contents/data/documents. Many

pages are open to access the information on the web. Web content mining is the process of identifying user specific data from text, image, audio or video already available on the web. It can provide useful and interesting patterns about user needs and contribution behavior [10]. Figure 1 describes the web content extraction techniques.

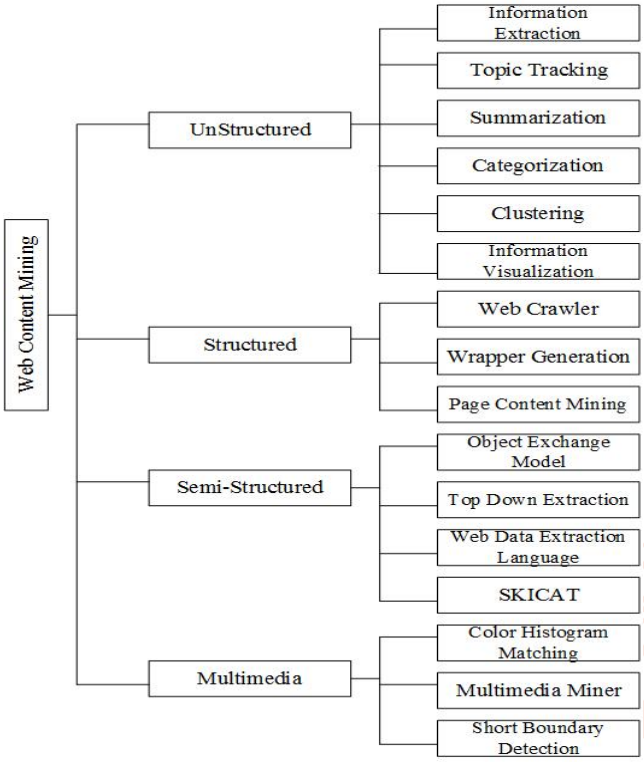

Figure1. Web Content Extraction Techniques

A. Unstructured Web Content Mining:

Aim of the proposed algorithm is to maximize the network. In unstructured data mining technique, Content mining can be done on unstructured data such as text mining of unstructured data gives unknown information. Text mining is extraction of previously unknown information by extracting information from different text sources. Content mining requires application of data mining and text mining techniques. Basic Content Mining is a type of text mining. Some of the techniques used in text mining are Information Extraction, Topic Tracking, Summarization, Categorization, Clustering and Information Visualization.

Information Extraction is used to extract information from unstructured data, pattern matching. It traces out the keyword and phrases and then finds out the connection of the keywords within the text. It utilizes feature extraction and key term indexing to build a graphical representation. This technique is very useful when there is large volume of text. Information extraction is the basis of many other techniques used for unstructured mining. Information extraction can be provided to KDD (Knowledge Discover in Database) module because information extraction has to transform unstructured text to more structured data. First the information is mined from the extracted data and then using different types of rules, the missed out information are found out. Information Extraction that makes incorrect predictions on data discarded [11].

In Topic Tracking according to each user it predicts the other documents related to users interest. Disadvantage of topic tracking is that when we search for topics we may be provided with information which is not related to our interest. For example if user sets an alert for 'web mining' it can provide us with topics related to mineral mining which are not useful for user.

Summarization has been used to reduce the length of the document maintaining the main points. It helps the user to decide whether they should read this topic or not. To understand the key points summarization tool search for headings and sub headings to find out the important points of that document. This tool also give the freedom to the user to select how much percentage of the total text they want extracted as summary. It can work along with other tools such as Topic tracking and categorization to summarize the document.

Categorization technique is used to identify main themes by placing the documents into a predefined set of group. This technique counts the number of words in a document. It does not process the actual information. It decide the main topic from the counts. It ranks the document according to the topics.

Clustering technique has been used to group similar documents. Here in clustering, grouping is not done based on predefined topics. It is done based on fly. Same documents can appear in different group. As a result useful documents will not be omitted from the search results. Clustering helps the user to easily select the topic of interest. Clustering technology is useful in management information system.

Information Visualization Through documents having similarity are found out. Large textual materials are represented as visual hierarchy or maps where browsing facility is allowed. It helps the user to visually analyze the contents. User can interact with the graph by zooming, creating sub maps and scaling. This technique is useful to find out related topic from a very large amount of documents.

B. Structured Web Content Mining:

Structure data mining is most widely used in web content mining. Structured data is easier to extract compare to unstructured data. Text mining of structured mining technique gives known information. The techniques which have been used for mining structured data are referred as Structured Data Mining Technique [12].

Web Crawlers are computer programs which traverse the hyper-text structure in the Web. There are two categories of Web Crawler such as: Internal and External Web Crawler. Internal Crawler crawls through internal pages of the Website which are returned by external crawler. External Crawler crawls through unknown Website. Page Content Mining is structured data mining technique which works on the pages ranked by traditional search engines. By comparing page content rank it classifies the pages.

In Wrapper Generation, it provides information on the capability of sources. Web pages are already ranked by traditional search engines. According to the query web pages are retrieved by using the value of page rank. The sources are what query they will answer and the output types. The wrappers will also provide a variety of Meta information. E.g. Domains, statistics, index look up about the sources.

C. Semi-structured Web Content Mining:

Aim of the proposed algorithm is to maximize the network. The use of Semi-structured data can be felt in the area involving raw data which does not have any fixed format. More and More data sets do not fit in the rigid relational model because of the Individual data items do not have the same structure completely. The techniques used for semi structured data mining are Object Exchange Model (OEM), Top down Extraction, and Web Data Extraction language.

Relevant information are extracted from semi-structured data and are embedded in a group of useful information and stored in Object Exchange Model (OEM). It helps the user to understand the information structure on the web more accurately. It is best suited for heterogeneous and dynamic environment.

In top down extraction, it extracts complex objects from a set of rich web sources and converts into less complex objects until atomic objects have been extracted. Web data extraction language converts web data to structured data and delivers to end users. It stores data in the form of tables.

D. Multimedia Web Content Mining:

It is a part of content mining where high level information and knowledge from large online multimedia sources. Multimedia data mining refers to the analysis of large amounts of multimedia information in order to find patterns or statistical relationships. Once data is collected, computer programs are used to analyze it and look for meaningful connections. This information is often used by governments to improve social systems. It can also be used in marketing to discover consumer habits. Some of the Multimedia Data Mining Techniques are SKICAT, Multimedia Miner, Color Histogram Matching and Shot Boundary Detection [13].

SKICAT is a Successful Astronomical Data Analysis and Cataloging System that produces digital catalog of sky object. It uses machine learning technique to convert these objects to human usable classes. It uses machine learning technique to convert these objects to human usable classes. It integrates technique for image processing and data classification which helps to classify very large classification set.

Color Histogram matching consists of Color histogram equalization and Smoothing. Equalization tries to find out correlation between color components. The problem faced by equalization is sparse data problem which is the presence of unwanted artifacts in equalized images. This problem is solved by using smoothening.

Multimedia Miner Comprises of four major steps, Image excavator for extraction of image and Video's, a preprocessor for extraction of image features and they are stored in a database, A search kernel is used for matching queries with image and video available in the database. The discovery module performs image information mining routines to trace out the patterns in images. Short Boundary Detection is a technique in which automatically the boundaries are detected between shots in video.

4. Conclusion

Nowadays, web pages become much more complex than before, so content extraction becomes more difficult and nontrivial. Extracting useful or relevant information from web pages becomes an important task. Web page content extraction technology is a critical step in many technologies. The rapid growth of the web in the last decade makes it the largest publicly accessible data source in the world. Web mining aims to discover useful information or knowledge from Web hyperlinks, page contents, and usage logs. Web content mining extracts useful information/knowledge from web page contents. This paper conduct the study on web content extraction techniques. According to the study, web content extraction can be classified in four main part. In the future, the detail analysis of web content extraction techniques will be discussed.

5. REFERENCES

[1] Sandip, prasenjit, Nirmal Pal and C.Lee Giles, 'Automatic Identification of Informative Sections of web pages', IEEE Transactions on knowledge and data Engineering, Vol 7, No 9, 2005.

[2] Sirsat, S. and Chavan, V, 'Pattern matching for extraction of core contents from news web pages'. IEEE Second International Conference on Web Research (ICWR) (pp. 13-18), 2016.

[3] Geng, H., Gao, Q. and Pan, J, 'Extracting content for news web pages based on DOM'. IJCSNS International Journal of Computer Science and Network Security, Vol 7, No (2), pp.124-129, 2007.

[4] Song, D., Sun, F. and Liao, L, 'A hybrid approach for content extraction with text density and visual importance of DOM nodes'. Knowledge and Information Systems, 42(1), pp.75-96, 2015.

[5] Carey, H.J. and Manic, M, 'Html web content extraction using paragraph tags', Industrial Electronics (ISIE), IEEE 25th International Symposium on (pp. 1099-1105), 2016.

[6] Mahesha, S., Giri, M. and Shashidhara, M. S, 'An Efficient Web Content Extraction from Large Collection of Web Documents using Mining Methods' International Journal of Computer Applications, 69(7), 2013.

[7] Gondse, P., Raut, A. and HVPMCOET A, 'Primary Content Extraction Based On DOM', Intl. Journal of Research in Advent Technology, 2(4), pp.208-210, 2014.

[8] Nethra, K., Anitha, J. and Thilagavathi, G., 'Web Content Extraction Using Hybrid Approach', ICTACT Journal On Soft Computing, 4(2), 2014.

[9] Kaur, P. and Bhatia, R., 'Development of Cluster based Supervised Learning Technique for Web News Extraction', International Journal of Computer Applications, 152(5), 2016.

[10] Shu, K., Sliva, A., Wang, S., Tang, J. and Liu, H., 'Fake news detection on social media: A data mining perspective', ACM SIGKDD Explorations Newsletter, 19(1), pp.22-36, 2017.

[11] YesuRaju, P. and KiranSree, P., 'A language independent web data extraction using vision based page segmentation algorithm', arXiv preprint arXiv: 1310.6637, 2013.

[12] Afonso, A. R. and Duque, C. G., 'Automated text clustering of newspaper and scientific texts in brazilian portuguese: analysis and comparison of methods', JISTEM-Journal of Information Systems and Technology Management, 11(2), pp.415-436, 2014.

[13] Gondse, M. P. G. and Raut, A., 'Main content extraction from web page using DOM', International Journal of Advanced Research in Computer and Communication Engineering, 3, p.5302, 2014.