Extracting Informative Web Content Using Web Content Mining Techniques

Deven M. Kene
Department of Computer Science
VidyabharatiMahavidyalaya
Amravati, Maharashtra
Email: dmk2006@rediffmail.com

Dr.Pradeep K. Butey
Head,
Department of Computer science
Kamala Nehru Mahavidyalaya
Nagpur, Maharashtra
Email: buteypradeep@yahoo.com

Abstract—Extraction of Information has become an important task for discovering useful knowledge or information from the WebPages. ButOn the Internet web pages contain several items that cannot be classified as the "informative content". Maximum clients and endusers search for the informative content, and don't interest in noninformative content. As a result, the need of Informative Web Content Extraction from web pages becomes important. With the informative content, web pages commonly have blocks that are not the main content blocks and are called the non-informative blocks or noise. Content in noise blocks will seriously harmful for information extraction, web mining, web searching. So identifying the main content block is a key issue. In order to improve the performance of extraction of information, cleaning of Web pages becomes critical. The main objective of this paper is to eliminate the non-informative content blocks from a Web Page and extracting informative web content using web content mining techniques. In the paper we proposed techniques, we address the problem of extracting the relevant content from web pages.

Keywords-*content extraction, information extraction, web content mining, web segmentation, Informative blocks*

*****

I. INTRODUCTION

Today internet and World Wide Web (WWW) has become the main and first source of information for everyone, but the World Wide Web has been exponentially increases. It resulted in difficulties for individual user to process all these information. Web contents such as multimedia data, structured i.e. XML documents, semi-structured i.e. HTML documents and unstructured data i.e. plain text [1] offer important information to the users and therefore be termed as informative contents. Other useful information on the Web is often accompanied by contents such as navigation bars, banner advertisements copyright notices etc [2] which can be termed as non informative contents. For web viewers and Web site design owners the non-informative blocks are functionally useful, but they often slow down automated information gathering [3]. Therefore these blocks are termed as the noisy blocks. Also, from the users' perspective only part of the information is useful for a particular application and the remaining information are noises. In the web page apart from the main or relevant content other noisy content is also so it degrades the performance of information retrieval applications. For improving the performance of traditional information extraction, it is necessary to differentiate valuable information from noisy content. Information contained in these noisy blocks can seriously hamper Web data mining task. Eliminating these noisy blocks is thus of great importance. A web page usually contains various content such as contacts, information at the bottom, navigation bars, advertisements, or just some decoration components which are not related the topic of the web page. In information extraction, contents in these parts are all noise information. Visitors to Web pages are only interested in the main content and have no use for the noisy content. Only the main content block of a web page contains the information we wants, we call this block *informative block*. So an accurate detection of informative block of a web page surely will improve the performance of information extraction. To distinguish between the informative and non-informative content in a web page, it needs to segment the web page into semantic blocks. There are several kinds of methods for Web Page SegmentationPresent days, Web page segmentation is a studied topic for research. Web page segmentation is a process to divide a Web page into visually and semantically cohesive pieces. Web page segmentation has a no. of benefits and potential Web applications. Web Page segmentation is of extreme importance when information is to be extracted omitting the noise from web pages [4]. A web page needs to be partitioned into blocks such as main content block, navigation block, advertisements, etc. It is practicable to separate these areas automatically for several useful applications. Techniques belonging to the Web Content Mining such as classification and clustering, separation of block of web pages and removal of noisy blocks enable one to produce much better result for extracting useful information. The web page segmentation is a task in which break down the structure of web page into smaller segments, in the segmentation consist of DOM-based segmentation, location based segmentation and vision based segmentation. The main focus is to extracts the informative content like text, images and multimedia from web pages .

II. LITERATURE REVIEW

EICD (Entropy based informative Content density) is approach developed by Manjusha Annam and G P Sajeev [5] build on Entropy. The Proposed EICD algorithm initially analyses higher text density content. Further, theentropy-based analysis is performed for selected features. The key idea of EICD is to utilize the information entropy for representing the

 ISSN: 2321-8169
30 – 33

knowledge that correlates to the amount of informative content in a page.

In Algorithm for extraction of core content, pattern matching technique are used [6]. The approach uses devised algorithm that applies regular expressions identify the correct pattern for extracting the actual text contents from these news documents. Proposed approach deals with news web pages of any size and extracts core contents with efficiency and high accuracy.

Mohammad Mehind Yadollahi and MasoudAsadpourproposed an AWS (Automatic webpage segmentation) technique [7]. AWS, which classifies the main content of a given webpage using a feature set consisting of structural and shallow text features. We benefit DOM tree of web pages for feature extraction. AWS consists of three main steps:Feature Extraction, Feature Selection, Classification.

Webpage segmenter technique proposed by AbdelghnyOrogatand Hameda [8]. Webpage Segmenter Using Block Function Tree has been described as a new approach for web page segmentation. It uses the fact that every block in the webpage plays a specific function that web designer designs the block for this function. The Segmentation process is divided into two phases: The first one is building the Block Function Tree and the second one is analyzing the resulting Block Function Tree.

Hybrid approach proposed byMadhura R. Kaddu and Dr.R.B.Kulkarnito Extract Informative Content from online web pages [9]. In this the web page is converted into DOM tree and features are extracted. Use this features in the machine leaning method like decision tree classification and dynamic rules are created. By using these rules informative content is extracted from the web pages. Further the rules which are created in the automatic extraction technique used as hand crafted rule for content extraction from the web pages without using machine learning method. Here the informative content like relevant text, images, multimedia are extracted from the web pages and these web pages are taken from the Internet i.e. Online. And also consider dataset for extracting relevant content. The propose work generates effective rules, achieve automaticity and efficiency. In the hybrid approach initially automatic rules are generated by using machine leaning method if rules are not present. Otherwise use the extracted rules to extract informative content from web pages without using ML technique. The important thing into this approach is to learn pattern or rules and use this rules for efficiently extracting the informative content from web pages.

The Debnath S et al proposed four algorithms to identify primary content blocks. First the algorithm segments the web pages into web page blocks by considering different HTML tags. Secondly differentiate the primary content blocks from non-informative content blocks. The algorithm decides whether block B is content or not and also compare block B with stored blocks to decide whether block B is same as stored blocks.

Weka-LibSVM approach proposed by Kevin Joy Dsouza and Zaheed Ahmed Ansari [10].In this presented work classification of multi-domain documents is performed byusingweka-LibSVM classifier. Here to transform collected training set and test set documents into term-document matrix (TDM), the vector space model is used. In classifier TDM is used to generate predicted results. The results emerged from weka with its GUI support using TDM have quick response time in classifying the documents.

Priyanka B dastanwala and Vibha Patel proposed a Text mining methods for extractinginformative content [11]. Text Mining alsoreferred as the process of deriving information from text. In this approach SVM and LDA method are used for solving problem of content extraction. Data cleaning and reparation, seed words generation, Fuzzy Keyword match methods, LDA method and support vector machine steps are used.

LBDA (Layout based detachment approach) approach proposed by Dr.AnnaSaroVijendran andC Deepa[12] . The proposed approach extracts the main content from the web pageand removes the irrelevant information like header, footercontents, navigation bars, advertisements and other noisy images. The proposed methodology uses the following techniques: tagtree parsing to get the analysis structure, block acquiring pagesegmentation method to remove unwanted tags, and dataextraction to retrieve the necessary contents. It can eliminatenoise and extract the main content blocks from web pageeffectively and display the essential content to the users. Theperformance is evaluated based on the following metrics likeprecision, recall, accuracy, execution time and memory usage. The implementation results obviously show that our proposedLBDA approach is performed better than the existing heuristicapproach. In this approach we proposed a novel methodology layout based detachment approach to extract the content from the web pages. Web page content extractions are more vital to retrieve the contents of web pages, particularly in unstructured web.

VIPS approach focus on segmentations of web pages and informative block detection algorithms [13]. Row –column splitting indicator helpsus provide an easy to use partition granularity value which solves the difficulty of choosing an appropriate Degree of coherence value in VIPS algorithm.

In Lou van, S said that, to extract relevant contentfrom web pages existing approach used heuristic methods butexisting heuristic methods use few features and it is quitedifficult to determine the threshold value for the featureparameters. Louvan, S developed hybrid approach consist ofmachine learning and heuristic methods. In machine learningapproach use many features and from many machine learningalgorithm may learn the parameter automatically. In machinelearning part consider two-phase classifier and single-phaseclassifier. In this classification DOM tree is generated fromHTML page in which each HTML element is consider as anode. Then many features are extracted for node and nodes areclassified.

III COMPARATIVE ANALYSIS TABLE OF WEB CONTENT MINING TECHNIQUES

S. No	Techniques/ Approaches	Proposed Approaches	Previoused Approaches
1	EICD (Entropy based informative Content density	EICD Algorithm	DOM Tree
		Weighted DOM Tree	SST approach
		Text-Density algorithm	Segmentation based approach
2	Algorithm for extraction of	DOM Tree	Tag based approach

	core content	Pattern Matching approach	A novel template approach
		Web content extractor	ECON approach
3	AWS(Automatic webpage segmentation)	Feature Extraction	Wrapper based approach
		Feature selection	Template based approach
		Classification	DOM Tree
4	Webpage Segmenter	Segmenter Block Function Tree	vision based segmentation
		DOM Tree	Text based segmentation
5	Hybrid Approach	Automatic Extraction Technique	Info-Discoverer
		DOM Tree	Generic approach
6	Weka -LibSVM	Term document matrix	Vector Space Model
		vector space model	Clustering
		classification	
7	Text Mining Methods	Latent Dirichlet allocation(LDA)	Seed words generation
		Support vector machine	Data cleaning
8	VIPS Approach	VIPS method	Heuristic rules
		Informative block dection algorithm	SST approach
9	LBDA(Layout based detachment approach	DOM Tree parsing	EXALG algorithm
		Vision based Tree	content extraction
10	Clustering Techniques	Graph Based Algorithm	Web Content Extractor
		Text based clustering	Screen Scraper

IV PARAMETERS AND FACTORS USED IN THE APPROCHES

In the above table we give comparatives analysis of web content mining 10 techniques for extraction of informative blocks. Here we mansion the parameters and factors used in various approaches.

1. In EICD technique we used parameters like No. of characters of node, No. of informative pages and No. of informative tags. The Text Density and Entropy factors are considered.
2. In Algorithm for extraction of core content approach the parameters used are precision and recall. Accuracy and pattern factor areconsidered.
3. In AWS technique, F-measure and precision are used. Performance and Effectiveness factors are considered.
4. In Webpage segmenter, goodness and Discrepancy are used. The factors consider in this method are Accuracy and Performance.
5. In Hybrid Approach we used Retrievaltimeand Recall parameters are used. The Accuracy and Efficiency factors are consider.
6. In Weka–LibSVM technique AREF file format and Count Matrix parameters are used. Performance and Accuracy factors are considered.
7. In Text Mining Methods we used parameters like seed Words and Roc curve. The factors considered are Accuracy and Performance.
8. In VIPS Approach Degree of Coherence and Granularity parameters are used. The Accuracy and Efficiency factors are considered.
9. In LBDA (Layout based detachment approach) No. of Relevant terms and precision parameters are used. The Time and memory factors are considered.
10. In clustering Techniques precision and recall parameters are used. The Time and memory factors are considered.

V. CONCLUSION

In this papers we gives various approaches and techniques for Extracting informative web content using web content mining techniques. The informative contents like text, images and multimedia are extracted from web pages using the above technique. Content extraction is useful for the human users as they will get the required information in a time efficient manner. In this study we give the techniques for the extraction of informative content blocks and elimination of non informative blocks based on the idea of web page segmentation.

REFERENCES

[1] S.Balan and P.ponmuthuramalingam, "A study of various Techniques of web content mining Research issues and Tools," International Journal of Innovative Research and Studies, pp. 508-516, May 2013.

[2] S.H. Lin and J.M. Ho, "Discovering informative content blocks from web documents," Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining, KDD ', pp. 588–593, 2002.

[3] Christian Kohlschütter and Wolfgang Nejdl, "A Densitometric Approach to Web Page Segmentation," ACM SIGKDD, 2008.

[4] Lan Yi, Bing Liu, Xiaoli Li, "Eliminating Noisy Information in Web Pages for Data Mining," ACM SIGKDD, August-2003

ISSN: 2321-8169
30 – 33

[5] Manjusha Annam and G.P.Sajeev, "Entropy based Informative Content density Approach for Efficient Web Content Extraction," International conference on Advances in Computing, Communications and Informatics, pp. 118-124, Sept. 2016 IEEE.

[6] Sandeep Sirsat and Dr.Vinay Chavan, "Pattern matching for extraction of core contents from web pages," Second International Conference on Web research, pp. 13-18, 2016 IEEE.

[7] Mohammad Yadollahi and Masoud Asadpour, "AWS: Automatic webpage Segmentation," Second International Conference on Web research, pp. 25-30, 2016 IEEE

[8] Abdelghny Orogat and Hamed Hemeda , "Web Page segmentation Using Block function Tree ," 2016 IEEE

[9] Madhura R. Kaddu and Dr. R.B.Kulkarni , "To Extract informative content from online web pages by using Hybrid Approach," International Conference on Electrical ,Electronics, and Optimization Techniques, pp. 972-977, 2016 IEEE.

[10] Kevin Deouza and Zaheed ansari , "A Novel Data Mining Approach For Multi Variant Text classification ," International Conference on Cloud Computing in Emerging Markets , pp. 68-73 ,2016 IEEE

[11] Priyanka Dastanwala and Vibha Patel, "A Review on social Audience Identification on Twitter using Text mining methods ," WiSPNET 2016 Conference, pp. 1917-1920, 2016 IEEE.

[12] Dr. Anna Saro Vijendran and C Deepa, "LBDA : A Novel Framework for extracting content from web pages," International conference on Advanced Computing and Communication systems , 2013 IEEE.

[13] Xuhong Zhang, Jing He and Frank Cobin, "Vision-based Web Page Block Segmentation and informative Block Detection," International Conference on Web Intelligence and Intelligent Agent Technology, pp. 265-269, 2013 IEEE.

[14] G. Poonkuzhali, R. Kishore Kumar, P. Sudhakar, G.V.Uma,and K.Sarukesi, "Relevance Ranking and Evaluation of Search Results through Web Content Mining," Proceedings of the Multi Conference of Engineers and Computer Scientists , vol.1, pp. 456-460, Mar 2012.