EXTRACTING HEAD COMPONENT FROM WEB NEWS PAGES

R Ashok Kumar Research Scholar, Rayalaseema University Kurnool
Dr Y Rama Devi, Professor, CBIT,Hyderabad

ABSTRACT

With the increased use of the Internet, there are huge data available on world wide web. Every Web document consists of useful information. There are many types of web documents. One of the type is a web news page or web news Document. There are n number of approaches exist to extract web content from web document. Existed approaches use an HTML source page of the web document, uses Document Object Model-DOM etc. Among These approaches some are automatic and some are semi-automatic. Many methods available Extracting web content from web news pages. One of the popular method uses Text Tag Ratio. This method extracts content successfully, but leaves noise. This paper mainly focuses to extract the Head components from web news pages. This approach uses existing Text Tag Ratio used in CETR and adds the novel approach to eliminate noise so as to extract only Head Components. As our Proposed system is able to extract 80%-85% user relevant information.

Index Terms: Web data extraction, Web mining, Web content extraction, Tag-Ratio, HTML,Document Object Model, tag-ratios, tag density.

I. INTRODUCTION

From the inception of the Usage of Internet, WWW has experienced remarkable growth. The content of the www is accessed via Web Browser. Huge information on number of domains is available on the internet. The Internet development brought prosperity in many fields such as information retrieval, knowledge communication, etc, and information become overload. To extract the relevant and useful information their different methods developed using information retrieval and data mining.Recent web document containing different kinds of information. Every we page in website contains the information required and besides document also contains noisy contents such as headers advertisements, footers, copyright information decorations, etc. The noisy contents may affect the performance of the user searching for the information.To avoid this noisy content and extracting main contents from web document has become the important process for the users required useful information. The approach is developed to extract the main content from web documents during crawling is required. This approach needs to separate the web document from noisy contents.The content extracted, can be used for processing of knowledge and classification of knowledge, further provides data and information source for stakeholders or other enterprises.Traditional extraction of information methods can be categorized mainly, semi-structured,text extraction and structured web information extraction. There are numerous approaches existing on web content extraction. Different approaches have been proposed for web information extraction algorithm and can be divided into three categories namely HTML source page based, Vision-Based, Document object model DOM-based and density based approaches. Recent research shown best results using the HTML based information extraction when it is applied on web news pages. Text Tag Ratio is used in finding the content. To remove the noise in the extracted content are proposed in this paper. Further this paper proposed the method to extract Head components from the web news pages which is the useful information to the user.

There have been several approaches existing to filter the main content of a webpage. Carey in CETR uses first time Tag Ratios in Web content extraction. This paper proposes a new methodology to extract head components from the

news web pages. From the content extracted using Tag Ratios, this new method uses further filtering of list tags, which has the head components.This new approach presented in this paper uses the further filtration process to find more relevant content.The presented method applied on the web news pages to extract head component. This new approach relies on the use of HTML tags in HTML of news web pages. The approach presented in this paper uses the Tag Ratios to extract the content. This paper concentrates to find the density of the content first and then head components. This new approach presented in this paper uses the noise removal methods to find relevant to information for head content extraction. This paper is structured as follows: Section II states the relevant work in this area. Section III explains the Presented new approach design methodology. Section IV discusses the evaluation measures used to test the method. Section V describes the experimental results for new approach presented in this paper.

II. RELATED WORK

Many algorithms have been proposed to extract web content from web pages. Content extraction via Tag Ratios(CETR) proposed by Weninger et al. [3] and developed Content extraction via Tag Ratios(CETR) and extracted web content from web pages using the text tag ratio measure on HTML source page. In Boiler Plate Detection Using Shallow Text Features [6] Christian Kohlschütter et al. developed method toExtract the content and classified the content into long text and short text. in HTML web content using Paragraph Tags Carey et al. [2] presented a method byUsing a Paragraph Extractor to extract main content in Web News Page. In A Vision Based Approach for Deep Web Data Extraction Wei Liu et al [7] developed Vision based approach which uses the visual information of the webpage to extract web content with web programming language independent approach. In Vision Based Page Segmentation algorithm,Deng Cai et al [8] used visual information of the web page to extract web content.

Early stages of research on web content extraction attempts to extract content were often some kind of human interaction required to identify important features of the Web site, while these methods could be accurate, and were not easily expandable to collect collective data. The other previous methods used several natural language processing methods to help define relationships between web page regions, or use HTML tags to identify multiple areas within text. Identifying the ads on the web page and removing the ad is developed by Kushmerick et al. [12]. Many papers presented the use of the Document Object Model (DOM) to extract formatted HTML data from Web sites. Mantratzis et al. [13] developed an algorithm that recursively searches through a DOM tree to find which HTML tags contained a high density of hyperlinks. Much research tends to rely on the work of former researchers. Pinto et al. Extend body text extraction using the document slope curve to determine the content in front of pages without content in the hope of determining whether a webpage contains content worthy of extraction. Gottron et al.[5] Suggested Content Extractor and Extractor algorithms that compare similarities between blocks on many web pages and classify sections as content in relation to a set of user-defined attributes. Though number of methodologies existing in web content extraction, recent research shows that tag ratio is used in many extracting web content methods. This paper proposes the extraction of main content by eliminating the noise and extracting the useful content. This paper proposed the frame work and the architecture for extracting the web content from News web pages.

III. ALGORITHM

Input: Web pages.

Output: Extracted Head Components.

1) Recognize web pages of which Head Components to be extracted.
2) Get HTML Source code of that web page.
3) Calculate List Tag Ratio for each Line.
4) Identify the Lines which has more List Tag Ratio from HTML Source.
5) Identify the Anchor Tag Lines from (4)
6) Extract Content from the Identified Lines from (5).

IV. SYSTEM ARCHITECTURE AND FRAME WORK

This section explains the step by step process of the proposed web content extraction methodology and figure representing the Architecture

Step1. Selection of Web pages.

We collect the News web pages for which the head components to be extract

Step2.HTML Source Page.

We take the HTML document as Input.

Step3. Calculating the Text Tag Ratio

We calculate the Text Tag Ratio for each line in the HTML source page. We create table of these Ratio's for further step.

Step4. Identification of web content.

By using the algorithm, we find the lines which has content. We delete the Lines which has no content.

Step5. Finding the Lines has List Tags.

From the Lines from the step 4 we find the lines has List Tags.

Step6. Removing Noise

Using the Noise removing algorithm we further eliminate the noise to get required content has Head components.

System Architecture: System architecture is explained in the following figure.

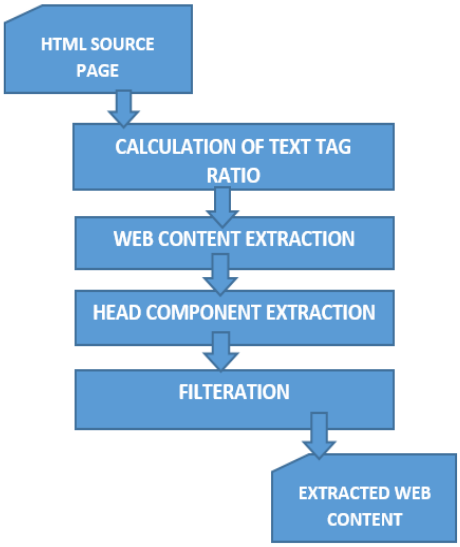

FIGURE1: FRAME WORK TO EXTRACT HEAD COMPONENTS

V. EXPERIMENTAL RESULTS

The hypothesis to be tested whether the proposed method to extract Head Components method would be a more effective extraction algorithm than CETR on News websites that use tag ratios to extract the content of from the Web News Pages. To test hypothesis, 10 Telugu Web News sites selected were local news-based websites such as eenadu, Saakshi, Telangana News and others, with a focus on vital infrastructure events in local areas. Proposed approach extracted the Head Components content by using the list tags and the noise removal method and finally extracted head components and more relevant content then CETR. The use of news article sites instead of other sites, such as blogs, forums or shopping sites, is because news sites tend to have a central article for discussion by site. We can simply down load the HTML source code from these sites which is used to extract the content from it. To compare the content extracted in the algorithm we used Precision, Recall and F1 Score.

Method	Head Component		CETR	
Metric	Scores	Standard Deviation	Scores	Standard Deviation
Precision	88.40%	3.20%	84.60 %	6.04%
Recall	78.36%	2.73%	52.00 %	4.20%
F1 Score	92.40%	2.50%	68.20 %	5.12%

TABLE I: SCORE COMPARISON BETWEEN BOTH METHODS CETR AND HEAD COMPONET METHOD ON WEBNEWS SITES.

The test of differences in each method between Table shows that whereas Proposed approach using list tags may be more accurate in sites that display required properties, the CETR is a more general method. Proposed approach reports an F1 ratio of 92.40% in the first website. However, CETR achieves an F1 rate of 68.2% in the first website. Testing on the 10 news websites, it found that presented new method in this paper shown 40% more accurate than CETR. CETR also does not work in groups that display the features required for List Tags, but it works more consistently on multiple types of websites.

VI. CONCLUSION

This paper presents a new approach for extracting head components which extends CETR which is used to extract content from webpages using text tag ratio. This new approach is more useful to extract content from web news pages. Based on previous work CETR with usage of the Tag Ratio method, the method presented in this paperimproves methodology to extract of head components from web news pages. In this paper Along with the Tag

Ration Two important methodologies were found to improve the CETR: 1) Usage the list tags for finding the head components in the web news page, 2) Further Noise removal method to filter the Head components by removing the Noise. The results showed that our approach showed a better overall performance than CETR in the News websites.Comparing the results using Precision, Recall and F measures we have shown our proposed approach extracted 60% more relevant content. Our approach limited to the extract head components from the News Web sites. It can further extend to Extract Head components from all the dynamic web pages.

REFERENCES

[1] S. Gupta, G. Kaiser, P. Grimm, M. Chiang, J. Starren, “Automating Content Extraction of HTML Documents,” in World Wide Web, vol. 8, no. 2, pp. 179-224, June 2005.

[2] H.J Carey,Milos Manic,” HTML Content Extraction Using Paragraphs tags” in IEEE 25th International Symposium on Industrial Electronics (ISIE), june 2016

[3] T. Weninger, W.H. Hsu, "Text Extraction from the Web via Text-to-Tag Ratio," in Database and Expert Systems Application, pp.23-28, Sept. 2008.

[4] T. Weninger, W.H. Hsu, J. Han, “CETR: content extraction via tag ratios,” in Proc. Intl. conf. on World wide web, pp. 971-980, April 2010.

[5] T. Gottron, "Evaluating content extraction on HTML documents," in Proc. Intl. conf. on Internet Technologies and Apps, pp. 123-132. 2007.

[6] C. Kohlschütter, P. Fankhauser, W. Nejdl, “Boilerplate detection using shallow text features,” in Proc. ACM intl. conf. on Web search and data mining, pp. 441-450, 2010.

[7] Liu, W., Meng, X.F., Meng, W.Y.: “ViDE: A Vision-Based Approach for Deep Web Data Extraction”. IEEE Trans. on Knowl. and Data Eng. 22(3), 447-460 (2010)

[8] Cai D, Yu S, Wen JR et al (2003) VIPS: a vision-based page segmentation algorithm. Microsoft Research

[9] D. Song, F. Sun, L. Liao, "A hybrid approach for content extraction with text density and visual importance of DOM nodes,"in Knowledge and Information Systems, vol. 42, no. 1, pp. 75-96, 2015.

[11] F. Sun, D. Song, L. Liao, “DOM based content extraction via text density,” in Proc. Intl. conference on Research and development in Information Retrieval, pp. 245-254, 2011.

[12] N. Kushmerick, "Learning to remove Internet advertisements," in Proc. Conf. on Autonomous Agents, pp. 175-181, 1999.

[13] C. Mantratzis, M. Orgun, S. Cassidy, "Separating XHTML content from navigation clutter using DOM-structure block analysis," in Proc. ACM conf. on Hypertext and hypermedia, pp. 145-147, 2005.

AUTHORS PROFILE

R ASHOK KUMAR is a Research Scholar in Rayalaseema university Kurnool, He is Having 22 years of Teaching experience and 5 years of research experience.

Dr Y RamaDevi Working as Professor in Department of CSE, CBIT,Gandipet. She is having 20 years of teaching experience and 10 years of research experience in the field of Datamining rough sets etc.