# ECETR-Extended Content Extraction via Tag Ratios

**R Ashok Kumar, Y Rama Devi**

***Abstract*: The regular approach for the Common internet user to search the Contents of World Wide Web is through web query interfaces. Enormous use of the Internet to for the desired information around the world, the collection of important information from multiple web pages remains a difficult problem. There are multiple web content extraction systems are proposed to extract desired information from webpages. There are many number of manually constructed, supervised, semi supervised systems are developed in the field of web information extraction. There are many ways to extract the content from web pages are developed, such as document Object trees (DOM), Text Density, Tag Ratio proportion, visual information based algorithms. This paper proposes a novel web content extraction method on web content extraction uses Tag Ratios and added clustering methods. As our Proposed system is able to extract 85%-90% user relevant information.***

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

## I. INTRODUCTION

The Internet the main growing source of information for the modern era. With the increased billions of internet users and innumerable web sites, the data available to an individual is valuable. Large amounts of Data available that may be important to different entities are provided in news articles on the Internet. Now a day's web news pages are repeatedly updated to get consistent data. From web news pages we get more valuable information, updated information, easy to extract from the software agent, or web programming using web apps. To analyse news articles from a paper source, the articles would first have to be read into a computer, making the process of extracting the information within much more cumbersome and time consuming. Thus, automating the extraction of the primary article is necessary to allow further data analysis on any information within a web page. To help analyse the content of web pages, researchers developed ways to extract the information required from a web page. The modern web page today consists of many different links, ads, and navigation data. This additional information may not be relevant to the main content of the website, and may be ignored in many cases. This additional information, such as ads, can also result in misleading or incorrect information. Therefore, determining the relevant main content of a Web page between additional information is a difficult problem. 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 ECETR an extended content extraction via Tag Ratios, which is a new method used to determine the content of the main text within an article on a Web site results most relevant information as possible. ECETR relies on the use of HTML tags in HTML. ECETR uses the Tag Ratios to extract the content. ECETR concentrates the density of the content and extracts the relevant content by eliminating the irrelevant content. ECETR uses the datamining clustering methods to combine relevant to information for meaningful content. This paper is structured as follows: Section II states the relevant work in this area. Details of section III explains the ECETR design methodology. Section IV discusses the evaluation measures used to test the method. Section V describes the experimental results for ECETR.

## II. RELATED WORK

Different algorithms have been proposed earlier for web content extraction. Weninger et al. [3] developed Content extraction via Tag Ratios(CETR) extracted contents from diverse web pages using the tag ratio measure on HTML documents. Christian Kohlschütter et al. [6] presented Boiler Plate Detection Using Shallow Text Features Extracted the content and classified the content into long text and short text. Carey et al. [2] presented in HTML web content using Paragraph Tags Used a Paragraph Extractor to extract main content in Web News Page. Wei Liu et al [7]. developed A Vision Based Approach for Deep Web Data Extraction used Vision based approach to extract web content with web programming language independent approach. Deng Cai et al. [8]in Vision Based Page Segmentation algorithm used visual information of the web page to extract web content. First 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. Kushmerick et al. has developed a way to identify ads on the page and to remove them. Many methods try to use the Document Object Model (DOM) to extract formatted HTML data from Web sites. Much research tends to rely on the work of former researchers. Pinto et al. Extend body text extraction using the document slope curve to

**Revised Manuscript Received on March 20, 2019.**

**First Author name: R Ashok Kumar** , Research Scholor, Computer Science and Engineering, Rayalaseema University, Kurnool, Andhrapradesh,India.

**Second Author name**, **Dr Y Rama Devi**, Professor Department of CSE, CBIT,Gandipet,Hyderabad,Telangana,India.

.
.

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. ECETR ALGORITHM

**Input**: Web pages.

**Output**: Extracted Content.

1) Recognize web pages of which analysis is to be made.
2) Get HTML Source code of that web page.
3) Calculate Text to Tag Ratio for each Line.
4) Identify the Lines which has more Ratio from HTML Source.
5) Apply Smoothening to Get Required Lines from (4)
6) Using DBSCAN algorithm find the content to be merged.
7) Extract Content from the Identified Lines from (6).

## 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 web pages for which the content to be extract

**Step2.HTML Source Page.**
It this part, we take the HTML Source page for the pages we required content.

**Step3. Calculating the Text to Tag Ratio**
We calculate the Text to Tag Ratio for each line in the source page. We form an array of these Ratio's for further step.

**Step4. Identification of web content.**
In this first step we identify the lines which has no content. We eliminate the Lines which has no content. Now we identify the lines which has the content.

**Step5. Smoothening technique**
Using the smoothening techniques, we get the required to combine and the lines to be eliminated.

**Step6. Merging process**
Using the DBSCAN algorithm we find the content to be merged. By Combining the required Content, we get the Required Content.

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

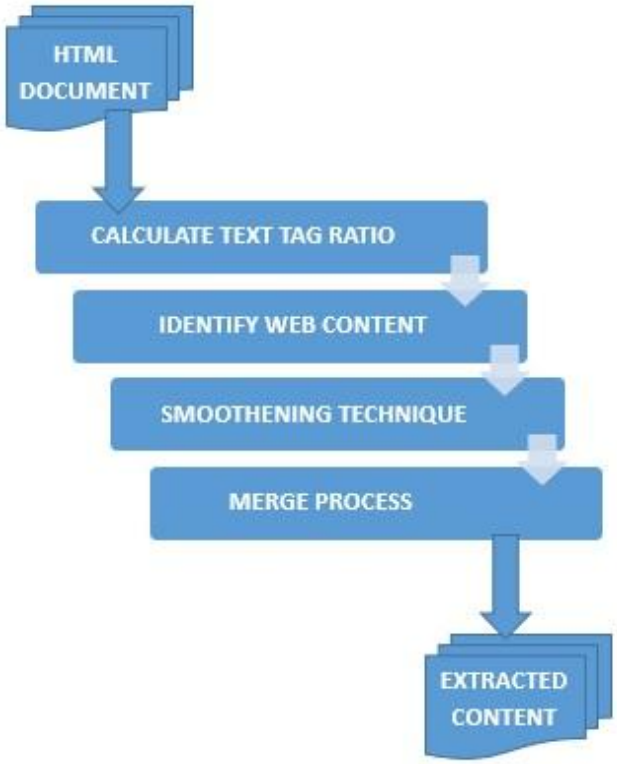


FIGURE 1: SYSTEM ARCHITECTURE

## V. EXPERIMENTAL RESULTS

The hypothesis to be tested whether the ECETR 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. ECETR extracted the content by removing the noise and finally 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.

**TABLE I: COMPARISON BETWEEN CETR AND ECETR ON WEBNEWS SITES.**

| Method | ECETR | | CETR | |
|---|---|---|---|---|
| Metric | Scores | Standard Deviation | Scores | Standard Deviation |
| Precision | 90.70% | 3.30% | 85.70% | 6.04% |
| Recall | 88.37% | 2.63% | 83.00% | 4.20% |
| F1 Score | 97.40% | 2.60% | 88.23% | 5.12% |

The test of differences in each method between Table shows that whereas ECETR may be more accurate in sites that display required properties, the CETR is a more general method. ECETR reports an F1 ratio of 97.40% in the first website. However, CETR achieves an F1 rate of 88.23% in the first website. Testing on the 10 news websites, it found that ECETR shown 40% more accurate than CETR. CETR also does not work in groups that display the features required for ECETR.

## VI. CONCLUSION

This document presents a revised method called ECETR which extends CETR to increase the relevant content extraction from web news pages. Based on previous work CETR with usage of the Tag Ratio method, the ECETR method improves methodology to extract of relevant content extraction. In this paper Along with the Tag Ration Two important methodologies were found to improve the CETR: 1) Usage the methodology for finding the dense of the content of the article, 2) Usage of the DBSCAN the data mining clustering method to increase the relevance in content extractions. The results were proved that ECETR method showed a better overall capacity than CETR in the News websites. Comparing the results using Precision, Recall and F measures we have shown ECETR extracted 40% more relevant content. ECETR limited to the news extraction from the News Web sites. It can further extend to Extract relevant content extraction 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

## AUTHORS PROFILE

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

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