International Journal on Recent and Innovation Trends in Computing and Communication ISSN: 2321-8169
Volume: 3 Issue: 11 6163 - 6165

Web Content Extraction Techniques: A survey

Kinnari Ajmera[#1], Khushali Deulkar[#2]
[#]Dept. of Computer Engineering, D.J. Sanghvi College of Engineering,
Mumbai University
Mumbai, India
[1]*kinnari94@gmail.com*
[2] *khushali.deulkar@djsce.ac.in*

***Abstract*—**As technology grows everyday and the amount of research done in various fields rises exponentially the amount of this information being published on the World Wide Web rises in a similar fashion. Along with the rise in useful information being published on the world wide web the amount of excess irrelevant information termed as 'noise' is also published in the form of (advertisement, links, scrollers, etc.). Thus now-a-days systems are being developed for data pre-processing and cleaning for real-time applications. Also these systems help other analyzing systems such as social network mining, web mining, data mining, etc to analyze the data in real time or even special tasks such as false advertisement detection, demand forecasting, and comment extraction on product and service reviews. For web content extraction task, researchers have proposed many different methods, such as wrapper-based method, DOM tree rule-based method, machine learning-based method and so on. This paper presents a comparative study of 4 recently proposed methods for web content extraction. These methods have used the traditional DOM tree rule-based method as the base and worked on using other tools to express better results.

Keywords*—Web Content extraction; DOM Tree; Class attribute; Gaussian smoothing; Webpage layout analysis; Subject detection and node density;*

*****

I. INTRODUCTION

Web provides us with huge amount of information that is displayed on websites and web pages in the form blogs, reviews, articles etc. With the advent of search engines like Google, a large amount of data is copied and extracted daily from the web. But the user is often not satisfied with the result since a lot of "junk" data is also copied with the important information. Gibson, Punera and Tornkins[2] estimated those additional and usually uninteresting contents to make up around 40 to 50% of most web pages on the Internet.This junk data often misguides the user.Junk data includes heavy noise and cluttering with distracted features.Thus web content extraction is used to extract the relevant and important material from the web pages by eliminating unwanted material.Figure below shows a webpage with irrelevant data content like advertisements.

Figure 1:Webpage with irrelevant data

Web content extraction can be classified into three broad categories namely wrapper-based method, DOM tree rule-based method and machine learning-based method. In wrapper based methods,wrappers are manually written by specially designed language and are automatically inducted using training examples.DOM tree rule based method relies on inherent structural features of HTML documents for accomplishing data extraction. It not only uses the HTML structure information but also use the text in the each node of the DOM tree. This method turn the webpage into a parsing tree (DOM tree), a representation that reflects its HTML tag hierarchy. For the machine learning-based method, NLP techniques like filtering, POS tagging, etc are often used. The traning processes make use of machine learning algorithms like SVM, HMM and CRF.

The Extracting context to improve accuracy for HTML content extraction has been proposed by Gupta S, et al. [5]. They have initialized to work towards dynamically detecting the context of the website, in terms of its content genre. They have present a new technique, based on incrementally clustering websites using search engine snippets, to associate a newly requested website with a particular "genre", and then employ settings previously determined to be appropriate for that genre, with dramatically improved content extraction results overall. Gottron T. [6] has developed new idea of combining content extraction heuristics. Content Extraction (CE) is the task to identify and extract the main content. Reis D C, et al. [7] has presented Automatic web news extraction using tree edit distance. It is domain-oriented approach to Web data extraction and discuss its application to automatically extracting news from Websites. Yi L, Liu B, et al. [8] proposed eliminating noisy information in web pages for data mining.

This paper primarily evaluates techniques for web content extraction using DOM tree method.

 ISSN: 2321-8169
6163 - 6165

Literature Survey

I. Web Document Information Extraction using Class Attribute Approach:

In this approach, the HTML web form is first converted to XHTML format i.e it is cleaned. In this cleaning process,a corresponding closing tag is provided with each opening tag. The information is then segmented using the DOM tree with the help of DOM Inspector tool. A DOM defines a w3c standard for accessing HTML documents and making changes in them. Everything in a HTML DOM document is a node. It consists of attribute nodes, element nodes and text nodes. The HTML DOM views an HTML document as a node tree with all nodes in a tree having a relationship with another node in the tree. The data in all nodes can be accessed through the tree and can be modified.Thus once the data is segmented using the DOM tree, class attribute approach is applied to the data.A class can be included in the body of the HTML code.Each class can have its own attributes. Thus when a child node is traversed within a class, only the data between the opening and closing tag is printed.

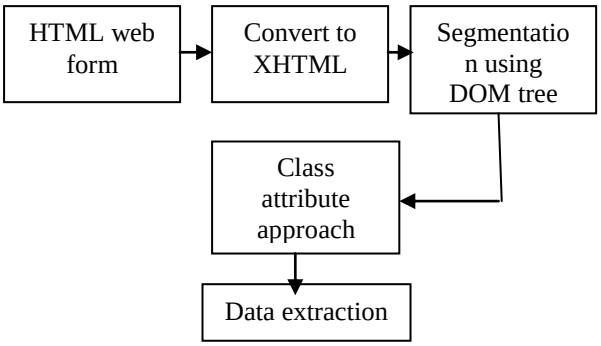

Figure 2:Steps of proposed method[1]

II. Content Extraction from web pages based on Gaussian Smoothing

In this method of data extraction, the web page is first visualized as a picture which has the main content as high density texts and additional content as low density text. The web page is differentiated on these parameters and its Primary Content is stored as useful information. For this the raw web pages are first passed through an HTML parser, which corrects the HTML markup and creates a DOM tree. From the DOM tree the node list is extracted and is made to undergo JS tag filter as a post-processing step to avoid the long continuous codes involved in JS tag being wrongly extracted as Primary Content. The tag nodes are then represented with tag tokens. Then pre-processing is carried out on the text-tag ratio sequence using the Gaussian Smoothing algorithm and then the SRS (Smoothed Ratio Sequence) is obtained. Then a ratio threshold is defined and all tokens with a higher ratio then the threshold is defined as Primary Content. Figure below shows the proposed approach.[3]

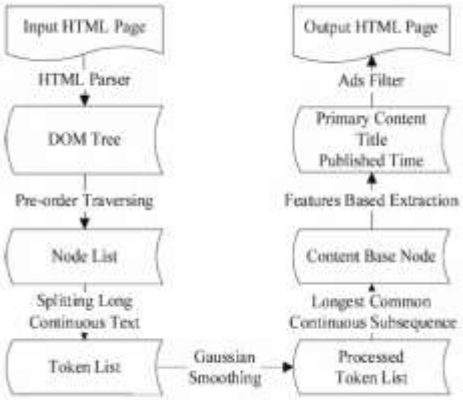

Figure 4. The flow chart of GSCE algorithm

III. Web Content Extraction based on Subject Detection and Node Density

Nowadays it has become important to pre-process data available on the web before analyzing it.Most techniques are semi-automatic and thus require manual help.But in this approach,web data is processed automatically.It is done using subject detection and node density.

Subject is the topic of the web page being considered.Usually the subject of e commerece websites is the product name. This algorithm is used to identify the subject node which consists of tag name, the keywords in meta tag and title tag, and some properties in cascading style sheet (CSS) including font properties . The subject node can be found by finding the node that has the maximum total weight. <h1> and <h2> is used in many websites as a subject butfor some web sites use other tags as a subject. Firstly,the HTML source code is obtained from the webpages.It is then parsed using DOM tree. The total weight for each tag in candidate tags is calculated using the following equation[4]

$$W_i = (\alpha \times N_i) + (\beta \times T_i) + (\gamma \times K_i) + (\delta \times C_i)$$

where ,

W_i = the total weight of the ith candidate tag

N_i =the weight of the tag name of the ith candidate tag.

α, β, γ, and δ are a constant whose value is between 0 and 1 such that $\alpha + \beta + \gamma + \delta = 1$.

T_i = similarity between words in the title tag and words in the candidate tags.

K_i =similarity between words in the meta tag whose value of name attribute is “keywords” and words in the ith candidate tag.

C_i = weight of CSS properties which is calculated from some CSS properties including display, font weight, and font size.

The node that has a density greater than the others is assigned to the subject node which is then passed over to the next process i.e node density.

Node density is used to identify the data rich region for extraction.The content data or main detail of something is extracted from the main node or subject node.The input

provided here is the output from the earlier process. Then, the value of the starting threshold is set as 0. Secondly, the current threshold is calculated by using equation[4]

$$\text{threshold} = \frac{\text{node density - link density}}{\text{link density}}$$

Thirdly, the algorithm checks whether the current threshold is greater than or equal to the previous threshold. If the current threshold value is greater than or equal to the previous threshold, it means that the current node is not the data rich region node. When the data rich region node is found, the content data is extracted from this node.

IV. Web Content Extraction based on Web Page Layout Analysis

This approach is mainly focused on improving on the DOM tree-based method. The web page is first made to undergo DOM tree parsing. This corrects the format errors on a web page. Then the layout of the web page is analyzed using VIPS (Vision based page segmentation). After this segmentation the marginal parts of a web page like the upper most, left most, etc are often found to be noisy elements and thus they are eliminated. After the elimination of the marginal blocks the candidate content block, the two boundary blocks and the beginning and ending blocks of the candidate content block are extracted.

V. Conclusion

In this paper DOM based web content extraction techniques are studied. The class attribute approach uses classes to extract data.All the data between the opening and closing tags is printed. Extraction through HTML classes effinciates the accuracy of the extraction algorithm.It is not applicable for hindi text page and requires graphical representation.The Gaussian smoothing approach uses Gaussian Smoothing Content Extractor (GSCE) to solve the problem.It is robust and efficient but makes extraction errors due to dates and titles that are irrelevant.Subject detection and node density makes use of the subjects of the e-commerece websites to extract relevant information.It does not require any information regarding the web pages except the URL of the webpage but the drawback is that it depends on subject detection method and link density.Finally , the webpage layout analysis makes full use of the layout information of the webpage to guide the whole extraction process.

References

[1] Shobhit Srivastava, Mohd. Haroon, Abhishek Bajaj.Web Document Information Extraction using Class Attribute Approach. 2013 4th International Conference on Computer and Communication Technology (ICCCT)

[2] D. Gibson, K. Punera, and A. Tomkins. The volume and evolution of web page templates.In WWW '05: Special Interest Tracks and Posters of the 14th International Conference on World Wide Web, pages 830 - 839, New York, NY, USA, 2005. ACM Press.

[3] Baohua Liao,Bo Cheng,Chuanchang Liu, Junliang Cheng,Gang Tan.Content Extraction from web pages based on Gaussian Smoothing. Proceedings of IC-BNMT20 10

[4] Warid Petprasit and Saichon Jaiyen.Web content extraction based on subject node and node density.

[5] Gupta S, Kaiser G, Stolfo S. Extracting context to improve accuracy for HTML content extraction. In: Pro ceedings of WWW'05. New York, NY, USA, 2005: 1114-1115.

[6] Gottron T. Combining content extraction heuristics: The CombinE system. In: Proceedings of iiWAS '08. NY, USA, 2008: 591-595.

[7] Reis D C, Golgher P B, Silva A S. Automatic web news extraction using tree edit distance. In: Proceedings of the 13th International Conference.

[8] Yi L, Liu B, Li X. Eliminating noisy information in web pages for data mining. In: Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY,USA, 2003: 296-305.