International Journal of Engineering Research ISSN:2319-6890)(online),2347-5013(print)
Volume No.5 Issue: Special 4, pp: 790-991 20 May 2016

LANGUAGE INDEPENDENT CONTENT EXTRACTION from WEB PAGES

Chandramma R[1],Ravindranath R C[1]Raviteja B[2],Ravikumar Y B[2], Venkatesh S[2],Yashwanth M[2]
Department of Computer Science and Engineering,
Vivekananda Institute of Technology, Bangalore-74
rchandramma.vkit@gmail.com ,Ravindranth.vkit.1985@gmail.com , raviteja6006@gmail.com
ravikumaryb.518@gmail.com, venkicse49@gmail.com ,yashwanthranju@gmail.com

***Abstract*: The rapid development of the internet and web publishing techniques create numerous information sources published as HTML pages on World Wide Web (WWW). However, there is lot of redundant and irrelevant information also on web pages like Navigation panels, Table of content (TOC), advertisements, copyright statements etc. There are various technologies & researches which are focusing on the extraction of relevant information from large web data storage. But still there is requirement of availability of automatic annotation of this extracted information into a systematic way so to be processed further for various purposes and for language independent also. In this system we present a simple, robust, accurate and language-independent solution for extracting the main content of an HTML formatted Web page and for removing additional content such as navigation menus, functional and design elements, and commercial advertisements. Accurate and efficient content extraction from Web pages is largely needed when searching or mining Web content. So in this system we use a new approach for content extraction called word to leaf ratio and density of links.**

Keywords: Content extraction, Entropy, Document object Model, template, Content Structure Tree, Web page segmentation, Clustering, Anchor text.

I. INTRODUCTON

Now a days, web is growing rapidly with huge amount of information is available in heterogeneous formats, such as, web pages, web archives, news wires, technical documents etc. Extracting high quality content efficiently from these web pages is crucial for many web applications such as, information retrieval, information extraction, topic tracking, text categorization and summarization. Many researchers have studied the problem of extracting content from web by means of different scientific tools in a broad range of application domain. This is a time consuming process & due to human effort it leads to inaccuracy up to a particular extent. There is a need technique for web page which should help a sample web page provide retrieved relevant information as per userrequirements.These techniques deals to locate the specific web page by interacting with web sources and extract the content stored in it. For example, if the source is an HTML web page, the extracted information consists of elements in the page as well as the full-text of the page itself. This requires pre-processing the extracted content, discovering the knowledge by converting it into a convenient structured form and storing it for later usage.

In this trending technology most of the system provides standard language i.e. English as any output, But our system provides language-independent output.The results are comparable or better than state-of-the-art methods that are computationally more complex, when evaluated on a standard dataset. When building a system for searching or mining Web content, a first task is extracting the main content and removing extraneous data such as navigation menus, functional and design elements, and commercial advertisements. Also when showing Web pages on small screens (e.g., of mobile phones) or sending text to screen readers that translate the text to a more appropriate format (e.g., text-to-speech for visually impaired people), the content extraction operation is very valuable. Content extraction (CE) is defined as the process of determining those parts of an HTML document that represent the main textual content [5].Because different Web pages often have a different layout and a variety of configurations are possible, the task is at first sight not trivial. Recently a number of solutions have been proposed. The problem, however, is to find a solution that is generic (i.e., portable to many types of Web pages), accurate (i.e., find all important content in a precise way) and efficient (often a large number of Web pages are processed).

The remainder of the paper is organized as follows. Section II discusses existing methodology. In section Vdescribes how we evaluate this method. Section VI, we present our method for content extraction and gives results and a comparison to existing methods. We conclude in section VII where we also give some hints for future research.

II. EXISTING METHODOLOGY

L-Extractor algorithm combines block-partitioning algorithm (VIPS-Vision based Page Segmentation algorithm) with support vector machine to identify content blocks in a web page. Content extraction and feature extraction was performed on different websites and showed that both algorithms performed better than INFODISCOVERER method in nearly all cases.The approach proposed by Yogesh, Vivek and Dipali was not to divide that contains other tags such as TABLE tag [1].

Moreover they performed experiment only on website containing Hindi pages.

(a) Wang proposed a method which is based on fundamental information of web pages. In first step, method extracts information from each web page and thereby combining that information to get site information. In second step, entropy estimation method is applied to discover actual information which is required by the user [2]

(b) Aanshi and Veenu proposed an approach for extracting informative content from web pages and a new approach for content extraction from web pages using word to leaf ratio and density of links [3].

(c) Li proposed a novel algorithm for extraction based on new tree CST (Content Structure Tree). CST is a good source for examining structure and content of web pages [4].

III. APPLICATIONS OF CONTENT EXTRACTION

-Online Newspapers , -E-commerce sites
- Research papers ,-Mobile phones

IV. IMPLEMENTATION PLATFORM

Software Requirement
- Intel Pentium 4 or higher ,- JDK 1.6
-Any latest Application server like JBoss, Glassfish
- Eclipse ,- Back End: - MySQL, Oracle 11g.
- Browser: -Internet Explorer 7, Mozilla Firefox.

V. PROPOSED APPROACH

Our approach combines word to leaf ratio (WLR) with link attributes of nodes for content extraction. In previous techniques characters were used instead of words but it purposelessly gives importance to long words. So instead of characters words are used. Leaves are examined in this ratio as these are the only nodes that consist of textual information. In some approach, they have not considered the idea that a block containing more number of links is less informative than the block containing lesser links. So, adding text link and anchor text ratios to word to leaf ratio gives a new approach which is more efficient.

Our method performs following steps:

1. Construct DOM (Document Object Model) tree of web page.
2. Remove noisy nodes which contain title, script, and head, Meta, style, no script, link, select, #comment and the nodes which have no words and are not visible.
3. Compute word to leaf ratio (WLR) as in Eq. 1.

$$WLR(n) = tw(n) / l(n) \quad (1)$$

Where tw (n) = number of words in the node
nl (n) = number of leaves in the sub tree of node n

4. Obtain initial node set which contains higher density of text. Initial node set I is defined as those nodes which satisfy the condition in Eq. 2.

$$WLR(n) \geq \sqrt{\max}_{WLR} \times WLR(root) \quad (2)$$

5. Compute text link ratio (TLR) i.e. the ratio of the length of the text and the number of links of node n in I.
6. Compute link text ratio (TLTR) i.e. the ratio of the length of the text and the length of the link text of node n in I.
7. Calculate weight of each node W (n) as in Eq. 3

$$W(n) = (TLR(n) + TLTR(n)) / a \quad (3)$$

Where a is normalizing factor.

8. Relative position of node n is defined in Eq. 4, Eq. 5, Eq. 6 and Eq.7.

Where min_{id},$max_{id}$are the minimum and maximum values of identifiers in I, min_{WLR}, $max_{WLR}$are the minimum and maximum WLR values from step 3. Relevance includes those nodes which have higher density of text. If two nodes have same R (n) then the node with lower identifier is selected.

$$R(n) = WLR(n) \times \left[\max w(n), \sum_{n_l \in Children(n)} R(n_l) \right] \quad (4)$$

$$\text{Where } w(n) = \begin{cases} r_{pos}(n) \times r_{WLR}(n) & \text{if } n \in I \\ 0 & \text{if } n \notin I \end{cases} \quad (5)$$

$$\text{Where } r_{pos}(n) = 1 - (id(n) - min_{id} / max_{id} - min_{id}) \quad (6)$$

$$r_{WLR}(n) = (WLR(n) - min_{WLR}) / (max_{WLR} - min_{WLR}) \quad (7)$$

9. Select the node which have highest values of W (n) and R (n) in proportion as, 0.7 × R (n) + 0.3 × W(n).
10. Selected node has the required textual information.

VI. ARCHITECTURE OF CONTENT EXTRACTION

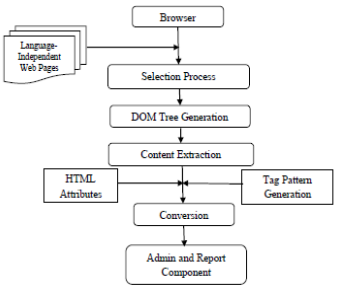

Fig 1: Block diagram of content extraction

The flow of the system architecture is divided into following components.

- User can easily generate the wrapper rules by using the many general purpose languages with the help of event handling.
- Data extraction is easily done by generating wrapper, but pattern generation and wrapper generation is separately done in existing systems.

Browser development, Selection process, HTML attributes viewer, DOM tree generation, Tag pattern generation, Data extraction and Result. Fig. 1 shows that the system architecture. The process first start from the browsing web page, then selection process is identifying the

data rich section which has been selected by user. DOM tree generation is responsible to create the hierarchical structure of the selected web page. The main task of extractor is handled by wrapper/extractor, extraction process depends on pattern generation as per the pattern assign to the web page it form regular expression and extract relevant contents from web page.

A. Browser Development

The web browser or simply browser is application software, mainly used for three purposes such as Presentation of HTML contents, Crawling or Traversing web databases and Fetching relevant information.

B. Selection Process

In selection process input web pages in different rich data section such as Multiple HTML values selection. These processes are carried as follows:

First examine the data rich section of the web pages and retrieved the relevant web page. Second identify the relatively valuable semantic token and attributes of the web pages. Third determine the web page tags for generation of the DOM tag tree. And last selection process responsible for end user whether wants to select text documents, images, scripts from relevant pages.

C. HTML-Attribute Generation

Here we first generate attribute generation, to identify the list of attributes of the tags. Attribute generation allows us to specify the grammar that matches tags elements which have certain attributes defined in the HTML source document.

D. DOM-Tree Generation

DOM-tree generation is the next step to generate DOM tree of selected web page HTML tag tree structure of web page. Most HTML tags work in pairs. Generating a DOM tree from a web page uses its HTML source code. The pattern generation is responsible for the generation of certain specific pattern for extraction of relevant contents from web pages. Mainly it performs following tasks:

- To identify the section and generates tag patterns from token and HTML attributes.
- Rich section is identification and extraction of unique content using patterns generation.

E. Content Extraction

In content extraction we will see, how to extract the contents with the help of pattern provide to extractor. Here we extract the whole contents from web page which is displayed on the browser page of web sites. Pattern generation uses the HTML parser which provides very effective and convenient tags for extracting the relevant data using DOM, CSS and JQuery Methods.

F. Conversion

In Conversion we will see, how the Unicode are converted to ASCII codes for other languages and it is reported to the next and final step for further process.

G. Admin and Report Component

In Admin and Report Component, we will see how the output is displayed on user created website and how it is reported as well.

VII. CONCLUSION

Informative Content Extraction from web pages is very important because web pages are unstructured and its number is growing at a very fast rate. Content Extraction is useful for the human users as they will get the required information in a time efficient manner and also used as a pre-processing stage for systems like robots, indexers, crawlers, etc. that need to extract the main content of a web page to prevent the treatment and processing of noisy, irrelevant and useless information. We have presented traditional approaches for extracting main content from web pages and also a new approach for content extraction from web pages using the concept of word to leaf ratio and link density and a language-independent system also.

VIII. FUTURE SCOPE

Automatic Content Extraction is an emerging field in research area as the amount and type of information added is increasing and changing day by day. We will perform our method on different websites like news, shopping, business and e-commerce and mobile phones and education and compare precision and recall with present methods. We will try to incorporate hypertext information also to the above method and work on event information generation.

ACKNOWLEDGEMENT

I sincerely thank all those who helped me in completing this task.

REFERENCES

i. Yogesh W. Wanjari, Vivek D. Mohod, Dipali B. Gaikwad, "Automatic News Extraction System for Indian Online News Papers", IEEE Knowledge and Data Engg , 2014.
ii. J. Wang and F. H. Lochovsky, "Data Extraction and Label Assignment for Web Databases," Proc.12th Int'l Conf. World Wide Web, 2003.
iii. Aanshi Bhardwaj and VeenuMangat, "A Novel Approach for Content Extraction from Web Pages", IEEE Knowledge and Data Engg, 2014.
iv. Y.Li," A novel method to extract informative blocks from web pages" International Join Conference on Artificial Intelligence, 2001.
v. Chia-Hui Chang and Shao-Chen Lui , "Information Extraction Based on Pattern Discovery," WWW10'01, Hong Kong, ACM, pp. 681-688, 2001.
vi. A.F.R.Rahman, H.Alam and R.Hartono, "Content extraction from HTML documents", International workshop on Web Document Analysis, pp. 7-10, 2001.
vii. H.Kao and J.Ho, "WISDOM: web intrapage informative structure mining based on document object model", IEEE Trans. Knowledge and Data Eng., vol. 17, no. 5, pp. 614-627, 2005.
viii. S.Debnath, P.Mitra, N.Pal and C.Giles, "Automatic identification of informative sections of web pages", IEEE Trans. Knowledge and Data g., vol. 17, no. 9, pp. 1233-1246, 2005.