Optimized Content Extraction from web pages using Composite Approaches

Sheba Gaikwad[1], G. Naveen Sundar[2]

[1]*Computer Science Department, Karunya University, India*
[2]*Computer Science Department, Karunya University, India*

Abstract- The information available today on web is tremendous and comes with greater challenges. Content extraction identifies the main content and removes the clutter from web pages. The main problem in extracting the content from the web page is the newer architecture of web pages and the diversity in the structure of web pages. Optimized content extraction from HTML documents using collective approaches proposes a hybrid model that operates on Document Object Model (DOM) tree of the corresponding HTML document to extract the content accurately. It combines approaches and techniques like statistical features extraction, formatting characteristic. Content type identification is used along with collective approach to overcome problem of dealing with versatile web pages, and yielding to achieve more accuracy in extracting the contents.

Keywords: Data mining, Information Extraction, Content extraction, HTML, Open source intelligence, Information filtering.

I. INTRODUCTION

The amount of information available on web is increasing day by day and has reached its ultimate from point in any time of history. This store of information is of great use for researchers and general public. The maximal use of this information can be done only if applicable tools are developed to extract and handle this large information. The problem is that the development of tools and techniques to extract and process this massive and diverse information has greater challenges as the web pages contains clutter (such as ads, unnecessary images and extraneous links) around the body of an article, which distracts a user from actual content [6,8]. This mix of unwanted noise and clutter with the real content in a web page complicates the task of automatic information (content) extraction and processing. The term "content extraction" also called as "Information Filtering" means extraction of useful noise-free information from the web pages.

The content extraction is progressively applied in various fields and applications. Content Extraction is beneficial for visually impaired and blind by identifying the real content within a web page and then increase the font size of the portions of the web page containing contents for better visualization or directly transforming the contents of the web page to speech. The content extraction is used in fields of Natural Language Processing (NLP) and information Retrieval (IR). Where these models derive accurate results based on relevance of contents and the reduction of "standard word error rate"[4]. Most of the NLP based IR applications necessitate dedicated extractors for each of the web domain [4] .The generalized content extractors are sufficient and less laborious than hand tailored extractors but is often found less accurate [4]. The field of Open Source Intelligence extracts the information from the web and automatically processes it to gain knowledge uses content extraction.

The expansion of the information on the web and increasing number of web pages introduces the need for new tools and techniques for content extraction in accurate way. A generalized traditional web page contains a title banner, list of links in right or left or both for site navigation and advertisements, a footer containing copyright statements, disclaimers or even sometimes navigational links [9]. The recent web pages tend to have more cleaner architecture using various layers for visual presentation , real content and interaction [10] having abandoned the use of old structural tags and adopted an architecture that makes use of the style sheets and div or span tags[9]. This change in architecture eases the development process but complicates the extraction process hereby reducing the effectiveness of old content extraction systems. The old content extraction systems operate on any varied web pages without considering the type of content the web page represents, which in return yield less accuracy in the content extraction models. This paper proposes methodology to improve accuracy by identifying the content type.

II. STATE OF ART

The content extraction is well versed research area in which a number of the methods have already been developed. The previous models used in content extraction mostly rely on removing clutters, disabling javascript, removing images, etc. Examples include WPAR, NoDoSE [12], XWRAP [13], etc. All of these traditional approaches were based on human aided rule based techniques or regular expressions applying on certain common web page designs. These approaches maintained the blacklist of advertisers for the elimination of advertisements which was an obvious disadvantage. The major drawback was that

these approaches were dependent on manual effort to create rules for each website and also update it as frequently as website updates.

Another approach namely discovering informative content blocks [5] proposes methods to automatically discover the intra-page redundancy and extract informative contents of a page. It concludes informative contents from a set of tabular documents (or web pages) of a web site. The model takes web page and partitions it into several content blocks called CB1, CB2, CBn based on HTML tag <TABLE> in a web page. The statistics and entropy of each feature is calculated, these features are the meaninful keywords except the stop words. The benefits of this approach are that it solves the problem of intra-page expected. However, the precision rate may become lower.

Filtering techniques for content extraction [4] system employs multiple extensible techniques for content extraction. In this technique the web page is converted into DOM tree and the set of filtering techniques with different levels of granularity are applied. These set of filters removes the scripts, styles and many other elements. Other filters used are advertisement removers, link list remover the empty table remover, and the removed link Retainer. The drawbacks are that it does not find the content but eliminates non-content. Also, the system needs to improve latency and scalability while serving many clients on proxy.

The DOM-structure block analysis [7] algorithm introduced separates true content in a web document from hyperlinked-clutter such as text advertisements and long links of syndicated references to other resources. The drawback of this approach is that it's less efficient when used in front pages of portals or other such web documents where an entire page is filled with short summarisations and their hyperlinks without giving prominence to any one body of content.

In the statistical model for content extraction [8] model the DOM tree is constructed for extracting the content from web page. Then the features like the quantity of text and the quantity of hypertext present at different nodes is analysed to determine the usefulness of each node in tree .The deviation and normalized deviation is calculated. The benefit of this approach is that it has High accuracy gain. The drawback is that as formatting Characteristics are not considered, it fails to extract the content when quantity of content at any part of document is low.

The methodology of content extraction from html documents [1] extracts the contents for PDA and other device. It takes each web page decompose it, determine the relationship among content and summarize it. The steps for Content Extraction from HTML documents are as follows. Content extraction by tag ratios [9] evaluates number of tags per line on HTML. Tag Ratios (TRs) are the basis by which CETR analyses a webpage in preparation for clustering. The benefits of this approach are that it is a viable and robust content extraction algorithm. It performs well even on non-news bodies and across multiple languages. It achieves better content extraction performance than existing methods works well across varying web domains, languages and styles. The drawbacks are that in some webpages wherein the HTML mark-up is written in a single line CETR would be forced to either return all text or no text. CETR does not perform well on portal home pages. In CETR the recall is high and precision is low. The webpages which do not have advertisements or menus, such as computer science professors' homepages, do not achieve high extraction accuracy. The content code blurring [3] approach aim's to locate regions in a document which contain mainly content and little code. This approach is specialized for wiki style pages.

III. OPTIMIZED CONTENT EXTRACTION USING COLLECTIVE APPROACHES

The proposed model is combination of collective approaches for effective and optimized content extraction. This approach is called collective because it mainly operates on two models of content extraction one based on statistical features and other on Formatting characteristics and collectively applies this model after identifying the content type to yield more accuracy in the extraction.

To evaluate the statistical features of the web page the structural analysis of that page is done and calculations of the quantity of text associated with different positions in the structure. For structural analysis the web page is converted into DOM (Document object model) tree, which is a cross platform and language independent convention for representing and interacting with objects in HTML , XHTML and XML documents. DOM is implemented by almost all of the HTML parsers as it provides a standard practice for accessing and manipulating HTML documents.

The model operates on the DOM tree representation of web page to calculate different statistical features of that web page, each of them namely Deviation (D), Link Density (L) , Normalized Deviation (N) and Normalized Link density (NL) [14]. The Deviation helps to identify amount of text present at each node, it is calculated as given in Eq(1).

$$D(i) = ø(i) - Avgø(T) \qquad (1)$$

The deviation at each node from the arithmetic mean represents how the node contributes towards the information being rendered to the user. The higher the deviation, the more information is rendered through that

node. The normalized deviation(N) [14] close to interval [0,1] for each node is estimated as given in Eq (2).

$$N(i) = \frac{D(i) - Min(D(T))}{Max(D(T)) - Min(D(T))} \quad (2)$$

The Link Density(L) helps for estimating if the node contributes towards traversing or is present there for information, it can be estimated using Eq (3), the link density can be further normalized by using Eq (4).

$$L(i) = \frac{l(i)}{\emptyset(i)} \quad (3)$$

$$NL(i) = \frac{L(i)}{Max(L(T))} \quad (4)$$

These statistical features of each node in the DOM tree determines the significance of that node towards contribution of information of that web page. These values are normalized, so important content should be retained. The calculation is based on the fact that the nodes associated with the content have higher values for the quantity of the text and lower values for quantity of hypertext Once the statistically useful nodes are identified, other nodes similar to useful nodes based on formatting characteristics and their position in the page are identified. All of the nodes classified as useful and nodes similar to useful nodes are considered to be the nodes containing real contents.

The content type identification[11] plays a vital role in yielding more accuracy in content extraction. This proposed model identifies the content of each page before applying the collective models. Once the content type is identified then the collaborative models are applied with specific thresholds to yield more accurate content extraction. In this each web page is classified and categorized to apply the respective content extraction technique. This pre-classification helps to yield more accuracy along dealing with versatile web pages.

IV CONCLUSION

This paper discusses different content extraction techniques, featuring their approach each with advantages and disadvantages. It also introduces a new Optimized content extraction using collaborative approaches. This new collaborative approach yields more accuracy than other previous approaches. It determines the type of content the web page is representing which in turn applies the specific extraction method for each different type of web page. This approach increases accuracy along with handling the different variety of web pages.

REFERENCES

[1] R. Alam, A.F.R. Rahman, H. Alam , R. Hartono, "Content extraction from HTML Documents", in : 1st Int. Workshop on Web Document Analysis (WDA2001), 2001.

[2] S. Chakrabarti, Mining the Web : Discoverin Knowledge from Hypertext Data , Morgan Kaufmann Publishers, 2003.

[3] T. Gottron, "Content code blurring: A new approach to content extraction", Proceedings of the 2008 19th International Conference on Database and Expert Systems Application, IEEE Computer Society Press, Washington, DC, USA, 2008, pp.29–33.

[4] S. Gupta, G. Kaiser, D. Neistadt, P. Grimm, "DOM-based content extraction of HTML documents", Proceedings of the 12th International Conference on World Wide Web, WWW '03, ACM, New York, NY, USA, 2003, pp. 207–214.

[5] S. -H. Lin, 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 '02, ACM, New York, NY, USA, 2002, pp. 588–593.

[7] C. Mantratzis, M. Orgun, S. Cassidy, " Separating XHTML content from navigation clutter using DOM-structure block analysis", in: Proceedings of the Sixteenth ACM Conference on Hypertext and Hypermedia, HYPERTEXT '05, ACM, New York, NY, USA, 2005, pp. 145–147.

[8] P.A.R. Qureshi, N. Memon, U.K. Wiil, "Statistical model for content extraction", European Intelligence and Security Informatics Conference (EISIC), IEEE Computer Society Press, Athens, Greece, September 2011, pp. 129–134.

[9] T. Weninger, W.H. Hsu, J. Han, "CETR: content extraction via tag ratios", Proceedings of the 19th International Conference on World Wide Web, WWW '10, ACM, New York, NY, USA, 2010, pp. 971–980.

[10] T.V. Raman, Toward 2W , beyond web 2.0, Commun. ACM 52 (February 2009) 52–59.

[11] Suhit Gupta, Gail Kaiser, Salvatore Stolfo , "Extracting context to improve accuracy for HTML content extraction" , Proceedings of WWW '05 Special interest tracks and posters of the 14th

international conference on World Wide Web, ACM, Pages 1114-1115.

[12] B. Adelberg , NoDoSE – a tool for semi - automatically extracting structured and semistructured data from text documents, SIGMOD Rec. 27 (June 1998) 283 – 294.

[13] L. Liu, C. Pu, W. Han, XWRAP: An XML-enabled wrapper construction system for web information sources, in: Proceedings of the 16th International Conference on Data Engineering, IEEE Computer Society Press, Washington, DC, USA, 2000, p. 611.

[14] Pir Abdul Rasool Qureshi , Nasrullah Memon, Hybrid model of Content Extraction, Journal of Computer and System Sciences , Volume 78, Issue 4, July 2012, Pages 1248–12