A Novel Approach for Web Content Extraction Using DOM Tree

Bhavdeep Mehta[#1], Meera Narvekar[*2]

[#] *Department of Computer Engineering-DJSCOE, Mumbai University*
J.V.P.D. Scheme, Bhaktivedanta Swami Marg, Vile Parle (W), Mumbai, India

[1]mehtabhavdeep@gmail.com

[2] narvekar.meera@gmail.com

Abstract— **The World Wide Web plays an important role while searching for information in the data network. Users are constantly exposed to an ever-growing flood of information. The approach will helps in searching for the exact user relevant content from multiple search engines. This helps in making the search more efficient and reliable. The proposed framework will extracts the relevant result records from multiple search engines by filtering out the noisy and redundant records. Finally the unique set of records is displayed in a common framework's search result page. The extraction is performed using the concepts of Document Object Model (DOM) tree. The paper comprises of a concept of threshold and data filters to detect and remove irrelevant data from the web page. The data filters will also be used to further improve the similarity check of data records. Also, visual cues from the underlying browser rendering engine is made use to locate and extract the relevant data region from the deep web by the keyword matching technique.**

Keywords— **DOM tree, Information Extraction, Content extraction techniques etc.**

I. Introduction

The World Wide Web (WWW) is flooded with a lot of information. Today, the Web can be partitioned into the Surface Web reached by common crawler-based techniques, and the rapidly growing Deep Web or Hidden Web, which consist of structured data hidden behind search forms. There was a need for a mechanism to extract the relevant data and separate it from the useless and irrelevant data. Thus, a mechanism is created for the extraction of information dynamically.

Data mining on the Web becomes an important task for discovering useful knowledge or information from the Web. As more and more information becomes available on the Web, the tools for collecting, organizing, and sharing Web content are becoming increasingly sophisticated. However, useful information on the Web is often accompanied by a large amount of noise such as banner advertisements, navigation bars, copyright notices etc. Therefore, it is essential to identify main content of a web page and automatically isolate it from noisy content for any further analysis. There is a need for a mechanism to extract the relevant data and separate it from the useless and irrelevant data in order to give accurate user queried result. The representation for web content extraction in which user uses the IR system to access the web pages with the help of web browser is shown in Figure 1.

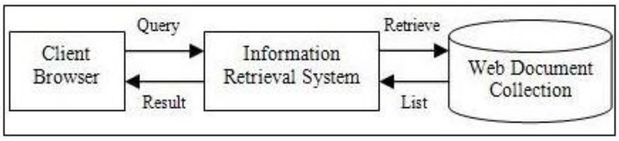

Figure 1: Basic representation for web content extraction

The aim is to develop a system that will be used to identify and extract the contents of the web pages of different structure and format. Along with that the system aims towards increasing speed and accuracy of content extraction. System also aims to extract the information dynamically. Here, the system will not perform only extraction process but also it will mine the data. The objective of the project work is to design and improve the application that will give the result more correctly and accurately.

II. RELATED WORK

The content extraction from HTML documents was first introduced to meet the demand for user requirement. Gibson D, et al. [1] has proposed the Volume and Evolution of Web Page Templates. According to Gibson et al. [1].About 40%-50% of the data on the Web are noises. In addition to noises, the heterogeneity of pages and demands for automation and efficiency make it difficult to extract contents from pages. If Web pages were written according to a common template, it could easily extract content simply by writing a regular expression. However, this quickly becomes impractical when dealing with hundreds of Web pages that are generated from different templates. Sun F, et al. [2] has proposed system based on DOM Based Content Extraction Text Density (CETD). A method for extracting the contents from web pages by Text Density, based on the observation that the content text is usually lengthy and simply formatted, while noise is usually highly formatted and contain less text with brief sentences. Observing that noise contains many more hyperlinks than meaningful content, they extended Text Density to Composite Text Density by adding statistical information about hyperlinks. In order to extract the content completely, author proposed the DensitySum technique instead of Data Smoothing. HTML Tidy was not sufficiently robust since it may sometimes cause some pages to not be properly parsed. Due to this it is not feasible to remove the unwanted content while searching for the required information. Gupta S, et al. [3] has proposed DOM-based content extraction of HTML documents. They have developed a framework that employs an easily extensible set of techniques that incorporate advantages of previous work on content extraction. Their key insight is to work with the Document Object Model tree, rather than with raw HTML markup. They have implemented their approach in a publicly available Web proxy to extract content from HTML web pages. Issues related to both latency and scalability was there in the current version. Adelberg B. [4] has proposed, Nodose - A tool for semi-automatically extracting semi - structured data from text documents. This research work explains NoDoSe, the Northwestern Document Structure Extractor, which is an interactive tool. Semi automatically approach which determines the structure of the document from where the content needs to be extracted but this approach was not fully automated approach.

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. System is restricted to Web news clipping system and not applicable for different application domains, especially those in which the schema of the data on the pages are complex. 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. Their ongoing research has spawned several CE heuristics of different quality. However, so far only the Crunch framework combines several heuristics to improve its overall CE performance. The CombinE system is designed to test, evaluate and optimize combinations of CE heuristics. Its aim is to develop CE systems which yield better and more reliable extracts of the main content of a web document. These heuristics might not perform outstanding on their own, so they might still improve the extraction performance in combination with other CE filters. 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 Web sites. Their approach is based on a highly efficient tree structure analysis but the research work was to extract the news based contents only. Yi L, Liu B, et al. [8] proposed eliminating noisy information in web pages for data mining. They proposed a noise elimination technique based on some observation: In a given Web site, noisy blocks usually share some common contents and presentation styles, while the main content blocks of the pages are often diverse in their actual contents and/or presentation styles. Based on this observation, they have proposed a tree structure, called Style Tree, to capture the common presentation styles and the actual contents of the pages in a given Web site. By sampling the pages of the site, a Style Tree can be built for the site, which we call the Site Style Tree (SST). The SST is employed to detect and eliminate noises in any Web page of the site by mapping this page to the SST. They proposed technique is evaluated with two data mining tasks, Web page clustering and classification. In their experiments, they have taken randomly 500 Web page sample from each given Web site to build its SST. So it is performed on single machine.

Discovering informative content blocks from web documents has been proposed by Lin S and Ho J. [9]. Researchers had proposed a new approach to discover informative contents from a set of Web pages. They proposed a system InfoDiscoverer, first segment a page into several content blocks according to HTML tag <TABLE> in a Web page. Based on the number of occurrence of the features in the set of pages, it calculates entropy value of each feature. This approach was to search informative content by dividing the page into several blocks of the contents. So it is better to eliminate the noisy contents from the web pages but experiments were evaluated for Chinese pages published by news web sites. Shuang Lin, et al. [10] proposed density based approach in content extraction. The proposed system is responsible to extract contents from web pages which are used to obtain page contents that are crucial to many web mining applications. However, previous density-based approaches were not effective in order to retrieve the pages that contain short relevant contents and long noisy data. It means that web page contains very less user relevant content and it is more surrounded with the noisy content. To overcome this problem, in this research work, researchers had proposed a content extraction approach for obtaining content from news web pages that combines a segmentation and a density-based approach. A tool BlockExtractor was developed for the content extraction. BlockExtractor tool identifies contents in three different steps.

I) Tool looks for all Block- Level Elements (BLE) & Inline Elements (IE) blocks, which are required to divide pages into blocks.

II) BlockExtractor calculates the densities of each BLE & IE block.

III) Tool eliminates all repeated BLE & IE blocks that have appeared in other pages from the same site.

For this experiment samples of news web pages were taken and the content extraction experiment was performed on the sampled data. Shine N. Das, et al. [11] had developed a system which was responsible to identify and to remove local noises resides in web pages in order to improve the performance. A three stage algorithm was proposed in which at the first stage feature selection is done. At the second stage a featured DOM tree is created and the noise is marked and pruned in the final phase. Alpa K and Shailendra Mishra [12] have proposed a system for identifying noisy data in web pages to improve the performance of the system. A simple idea for detection and removal of noises a new DOM tree structure is proposed. The result shows the remarkable increase in F score and accuracy is obtained but expansion of semantic and synonyms, which are needs to be, investigate.

III. THE APPROACH

A. Related Concepts - HTML DOM

Basically HTML DOM defines a standard way to access and manipulate HTML documents and it looks the HTML page as a tree like structure known as node tree. Everything in a web page contains HTML tags which all are called as node. The entire document consists of HTML tags and elements, text nodes, comment nodes. Figure 2 shows an example of an HTML document and its node tree.

```
<table>
<tr>
  <td>Amit</td>
  <td>Vishal</td>
  <td>Pratik</td>
</tr>
<tr>
  <td>Samir</td>
  <td>Ketan</td>
  <td>Arjun</td>
</tr>
</table>
```

Graphical representation

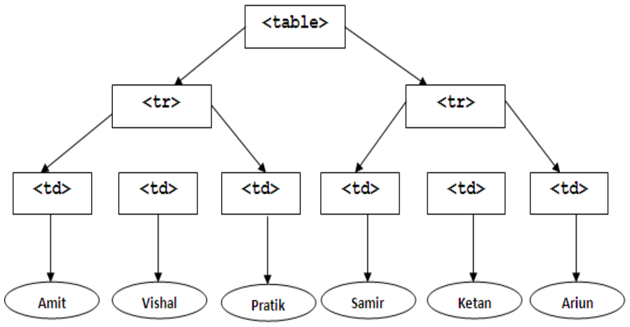

Figure 2: Representation of DOM Tree for HTML code

The Document Object Model (DOM) is a standard for accessing HTML and XML elements from made from the web page. The Document Object Model is a platform and language-independent approach for representing objects in HTML and XML documents. The Document Object Model (DOM) is an application programming interface (API) for valid HTML and XML documents and it is a parser which parses the HTML and XML document in order to form a tree like structure. Objects in the DOM tree can be manipulated by using methods on the objects. DOM tree allows dynamic access of programs and scripts and update the content, structure and style of page. DOM defines the logical structure of documents and the way a document is accessed and manipulated. With the help of Document Object Model a tree structure can be formed for entire web page including relevant and noisy node. With the Document Object Model, programmers can build documents, navigate their structure, and add, modify, or delete elements and content in the tree like structure.

B. *Design and Proposed Work*

In previous techniques content extraction approach that combines a segmentation-like approach and a density-based approach. This system was applicable to extract the contents from the sampled web pages related to news websites only. This approach can't effectively manage the web pages that contain short contents and long noises and it was not developed for the dynamic environment instead they took sampled data and experiment is made on it.

To overcome this problem it was proposed and redesigned the content extraction approach to make it applicable for other kinds of Web pages such as blogs, forums and articles etc. The proposed approach will extract the contents from the web pages dynamically. With this approach user will get maximum queried content and eliminate the unwanted content. Every web applications have different format and different structure that they follow where contents can reside.

For example

1) News channel website: They have fixed format for their content, editor and writer.

2) Blogs: It has different formats. There might be images, questions and answers, views etc.

3) Forums: It has again different format. It follows its own structure. It may be small discussion. Each posts reply is different. It is used for collaborative learning.

C. *Proposed Architectural Diagram*

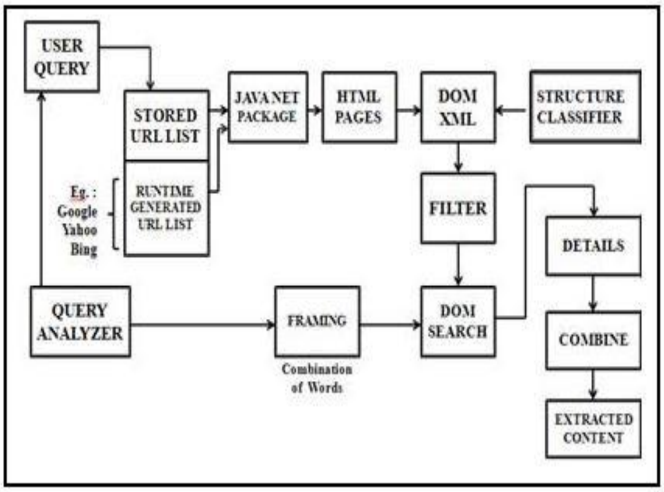

Figure 3: Architectural Design

Architectural design shown in figure 3 it explains two different techniques for content extraction.

1. *Extraction of content using stored URL list:*

This approach includes the extraction of information from the stored URL list. The URL list that can be used frequently will be stored and upon its need it will get fetch from the system itself. For Example, it stores the URL list for news websites, for stock exchange websites etc.

2. *Extraction of content using runtime generated URL list:*

This approach again useful for extraction of web content information. This approach will get used when user don't know the URL from where the information need to get extract.

In this event information will be extracted from the runtime generated lists.

In this approach it will search the contents from the various different search engines like GOOGLE, YAHOO, BING etc. and it will display the best search result URL links from the various search engines in one framework.

- *Query Analyser:*

The query analyser takes the query from the users, parses it and analyses the user query and their order to make sure they comply with the rules of the language grammar. If an error is found in the query submitted by the user, it is rejected and an error code together with an explanation of why the query was rejected is returned to the user. If entered user query is correct then query analyser will decide which URL list need to refer i.e. Extraction of content using stored URL list or Extraction of content using runtime generated URL list that is discussed above.

- *JAVA NET Package:*

The main purpose is to permit the creation of a TCP/IP connection between two applications. The java.net package contains a collection of classes and interfaces that provide the low-level communication. The java.net package provides support for the two common network protocols i.e. TCP / IP and UDP. This package is simply used to connect to the network and to load the html pages.

- *Framing:*

With the help of query analyser framing module is responsible to frame the user entered sentence in to appropriate query using combination of words.

- *XML DOM:*

The XML DOM defines a standard way for accessing and manipulating XML documents. A DOM implementation presents an XML document as a tree structure, or allows client code to build such a structure from scratch. It then gives access to the structure through a set of objects which provided well-known interfaces.

- *Structure Classifier:*

Structure classifier determines the DOM tree structure of entire web page. This classifier includes all the contents of web page in DOM tree including relevant as well as noisy contents.

- *Filter:*

Filters allow the user to "filter out" the unwanted nodes from the DOM tree. Each filter contains a user-written function that looks at a node and determines whether or not it should be filtered out. Filters allow applying a restricted view on the DOM tree of an XML document.

- *DOM Search:*

After applying filter to DOM tree structure which is created from the web page, the unwanted or noisy nodes are eliminated. As per the user query required node will be search from the filtered DOM tree using DOM search block which is present in DOM tree structure.

- *Details and Combine block:*

Sometimes it may happen that there are multiple nodes gives the same words for the user query. So such nodes need to get combine in order to maintain the accuracy of the result.

- *Extracted Contents:*

After combining all the nodes which gives the same result for the user query this block is responsible for extracting the accurate user queried contents from the DOM tree which is developed from the web page.

- *Threshold Concept:*

Threshold concept is very useful in order to search the required word from the node of the DOM tree. If the required word found from the node of the DOM tree and its threshold value is equal or greater than the set threshold value, in that case then the node information is useful else discard the entire node.

IV. EXPECTED OUTCOMES

The system should functions more accurately in dynamic environment to search the content from the web pages such as blogs, forums, articles etc. Result will include the maximum relevant extracted content as per the user query.

In order to get accurate result, it is very necessary to process the content of the web page and remove unwanted content. To improve the system performance, Perform appropriate comparison of extracted content. Analyze the extracted result and Improve the result and to enhance the accuracy.

Appropriate graph can be useful in order to display the extracted result. From the graph it's feasible to determine the accuracy of the relevant content and noisy content. As per the assumption averagely system will be able to extract 70%-80% user relevant content by eliminating noisy content from the different structured web pages like blogs, forums, articles etc. in the dynamic environment. False alarm will generate in order to improve the system performance.

Deliverables:

The web applications which will accept the user query and search the relevant web content and eliminate the noisy web content.

V. CONCLUSION

The findings in the paper proposed that a content extraction approach that uses DOM tree structure to represent the data in better format. The system will extract the content dynamically from the different structured web pages such as blogs, forums, articles etc. The approach helps in searching for the exact user relevant content from multiple search engines by filtering out the noisy and redundant records. Finally the unique set of records is displayed in a common framework's search result page. The concept of threshold and data filters to detect and remove irrelevant contents. The data filters will also be used to further improve the similarity check of data records. Also, visual cues from the underlying browser rendering engine is made use to locate and extract the relevant data region from the deep web by the keyword matching technique. To improve the performance, system will perform appropriate comparison of extracted content. Thus the method will successful in retrieving the data records with the help of visual cues and DOM tree properties. As per the assumption averagely system will be able to extract 70%-80% user relevant content by eliminating unwanted content from the different structured web pages like blogs, forums, articles etc. in the dynamic environment.

ACKNOWLEDGMENT

The authors would like to thank Shuang Lin for proposing the idea of designing a fully automatic content extraction system for different types of web applications such as blogs, forums etc. I am thankful to Head, Department of Computer Engineering (DJSCOE) and Research guide Meera Narvekar for her valuable comments and suggestions during the course of research work.

REFERENCES

[1] Gibson D, Punera K, Tomkins A. The volume and evolution of web page templates. In: Proceedings of WWW'05. New York, NY, USA, 2005: 830-839.

[2] Sun F, Song D, Liao L. DOM based content extraction via text density. In: Proceedings of the 34th International ACM SIGIR Conference. Beijing, China, 2011: 245-254.

[3] Gupta S, Kaiser G, Neistadt D, DOM-based content extraction of HTML documents. In: Proceedings of the 12th International Conference on WWW. Budapest, Hungary 2003.

[3] Adelberg B. Nodose—A tool for semi-automatically extracting semistructured data from text documents. In: Proceedings of SIGMOD'98. New York, NY, USA, 1998: 283-294.

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

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

[6] Reis D C, Golgher P B, Silva A S. Automatic web news extraction using tree edit distance. In: Proceedings of the 13th International Conference on World Wide Web. New York, NY, USA, 2004: 502-511.

[7] 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.

[8] Lin S, Ho J. Discovering informative content blocks from web documents. In: Proceedings of SIGKDD'02. New York, NY, USA, 2002: 588-593.

[9] Shuang Lin, Jie Chen, Zhendong Niu Combining a Segmentation-Like Approach and a Density-Based Approach in Content Extraction. School of Computer Science, Beijing Institute of Technology, Beijing 100081, China.

[10] Shine N. Das, Pramod K. Vijayaraghavan, Midhun Mathew - Eliminating Noisy Information in Web Pages using featured DOM tree. In: International Journal of Applied Information Systems (IJAIS) – ISSN : 2249-0868 Foundation of Computer Science FCS, New York, USA.

[12] Alpa K. Oza, Shailendra Mishra. Elimination of Noisy Information from Web Pages. In: International Journal of Recent Technology and Engineering (IJRTE). ISSN: 2277-3878, Volume-2, Issue-1, March 2013.