# A New Improved Approach for Elimination of Noisy Content From Web Pages

Priyanka Chaudhari, Prof. Seema Ladhe

*PG Scholar, Computer Engineering, MGMCET Kamothe, Navi Mumbai (India)*
*Professor, Computer Engineering, MGMCET Kamothe, Navi Mumbai (India)*

***Abstract: Web pages typically contain redundant information like banner ads, navigation bars, copy right and privacy notice, advertisements etc. Existing schemes proposed by various authors are not able to extract similar blocks from multiple web sites. To the best of our knowledge none of the authors had proposed the scheme which extracts the similar contents from multiple websites. We have proposed an algorithm, Similar Content Extraction Algorithm (SCEA), which will extract the similar content from multiple websites. In this algorithm, we design Document Object Model (DOM) using HTML parser which helps to segments the document into blocks. SCEA based on page segmentation algorithm which segments the web page into homogeneous blocks. Partitioning the pages into homogeneous blocks helps to check the similarity between two blocks. Our algorithm checks the similarity between two blocks using similarity Features Vector which compares similarity features with set threshold value. SCEA aims to provide more efficient results to the customers on their particular search.***

***Keywords: DOM, SCEA, HTML.***

## I. INTRODUCTION

The rapid expansion of the internet has made web a popular place for disseminating and collecting information. Apart from the useful information on the web, it usually has such information as navigation panels, copyright notices, banner ads, etc . Although these information item are useful for human viewers and necessary for the Web site owners, they can seriously harm automated information collection and Web data mining, e.g. Web page clustering, Web page classification, and information retrieval. So how to extract the main content blocks become very important. Web pages contain Div block, Table block or other HTML blocks. Existing schemes proposed by various authors are not able to extract similar blocks from multiple web sites. So to extract the similar contents from the multiple web sites , Document Object Model (DOM) using HTML parser is design which helps to segments the documents into blocks. Here the page segmentation algorithm is also used to segments the web page into homogeneous blocks, which helps to check the similarities between the two blocks. These similarities between two blocks are checks using the similarity features vector. Features Vector compares the similarity features with the threshold value.

## II. SYSTEM MODEL

The rapid expansion of internet has made web popular now a days. Currently, the world wide web is the largest source of information. So many times same information is available on many different web pages which are differ only by innovation of web. Search engines crawl the World Wide Web to collect web pages. It required to eliminate the noisy contents from web pages and extracting the main contents.

Present system focus on detecting and eliminating local noises in Web pages to improve the performance of Web mining, e.g., Web page clustering and classification. This work is motivated by a practical application [6]. *Web page cleaning* process is present currently to clean the web page and for extracting the main contents from the web pages.

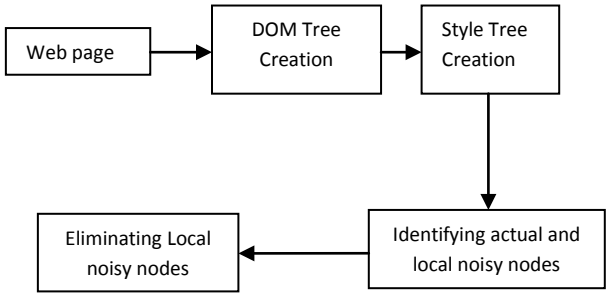


Fig 2.1. Web Cleaning Process

Web Page Cleaning:

In a typical commercial Web site, Web pages tend to follow some fixed layouts or presentation styles as most pages are generated automatically. Those parts of a page whose layouts and actual contents (i.e., texts, images, links, etc) also appear in many other pages in the site are more likely to be noises, and those parts of a page whose layouts or actual contents are quite different from other pages are usually the main contents of the page [6].

This Web page cleaning process is able to eliminate the local noises from the web pages. But this existing system is not able to extract the similar contents from the number of web pages. Extraction of similar contents from web pages is required when user wants to search some information for example News article which is present on number of web pages.So to improve the search result of user it is important to design a proposed system which is able to extract the similar contents from web pages ind improves the search result.

## III. PREVIOUS WORK

**R. R. Mehta, P. Mitra, and H. Kamick(2005)** proposed **Extracting Semantic Structure of Web Documents Using Content and Visual Information:**

This system provide a page segmentation algorithm which uses both visual and content information to extract the semantic structure of a web page.The output of the algorithm is a semantic structure tree whose leaves represent segments having unique topic.This algorithm is expected to outperform other existing page segmentation algorithm since it utilizes both content and visual information [5].

**Divya C.** proposed **Mining Contents in Web pages and Ranking of Web Pages Using Cosine Similarity:**

This paper introduces a method for calculating the rank of web page based on content similarity between the web documents and the user query.Here cosine similarity retrived most relevant pages to the user than the jaccard similarity [1].

**Bing Liu, Robert Grossman, Yanhong Zhai** proposed **Mining Data Records in Web Pages:**

It is a technique which is based on two important observations about data records on the web and a string matching algorithm. This system is able to mine both contiguous and noncontiguous [10].

**S. H. Lin and J. M .Ho** proposed **Discovering Informative Content Blocks from Web Documents:**

This approach discovers informative contents from a set of tabular documents or web pages. Here InfoDiscover system first partitions a page into several contents blocks. By analyzing the information measure, this system dynamically selects the entropy-threshold that partition blocks into either informative or redundant [8].

## IV. PROPOSED METHODOLOGY

In this proposed system aim is to extract the similar blocks across different web sites. Developing more challenging technique for retrieval of main contents from web pages is challenging task, but for that we need to remove following drawbacks detected in earlier available main content extraction schemes from websites:

1. Unable to extract similar contents from many websites.
2. Unable to make a jump to the relationship between words.
3. Many techniques that semantically related contents are grouped together in a web page are not always true.
4. Some techniques are applicable to tabular pages only instead of general web pages.
5. Some techniques are unable to divide web page into homogeneous blocks.

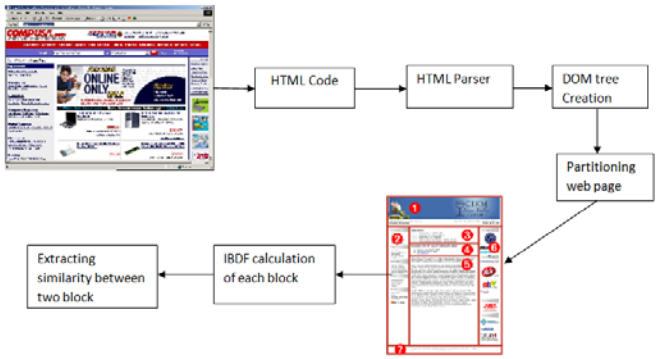


Fig 4.1. Similar Content Extraction Algorithm (SCEA)

To resolve these problems, we propose the advance system called similar content extraction algorithm (SCEA). The overall Proposed System architecture shown above basically consist of four major blocks:

(1) DOM (Document Object Model) creation.

(2) Partitioning of web page.

(3) IBDF calculation of each block.

(4) Checking similarity between two blocks.

DOM (Document object model) tree Creation: System will use open source HTML Parser that builds a DOM tree from a page using its HTML code. HTML documents contain HTML tags and plain text. HTML lets format text, add graphics, create link, input forms, frames and tables, etc. In a DOM tree, tags are internal nodes and the detailed texts, images or hyperlinks are the leaf nodes. Following Figure shows some html segments and its corresponding DOM tree. In the DOM tree, we need to tidy some unnecessary nodes, such as script, style or other customized nodes. HTML Web pages begin from the BODY tag since all the viewable parts are within the scope of BODY.

Partitioning of web page: A Web page is usually contains several pieces of information and it is necessary to partition a Web page into several segments (or information blocks) before organizing the content into hierarchical groups. The page segmentation algorithm relies on the DOM tree representation of the Web page and traverses it in a top-down fashion in order to segment the content of the page, which lies at the leaf nodes. We define a segment as a contiguous set of leaf nodes within a Web page. The algorithm aims to find homogeneous segments, where the presentation of the content within each segment is uniform.

IBDF (Inverse Block Document Frequency) calculation of each block: To eliminates redundant blocks depending upon the inverse block document frequency (IBDF) of a block. The IBDF is inversely proportional to the number of documents in which the block occurs. The blocks that occur in multiple pages are redundant blocks and block which appear in one page is a content block.

Checking similarity between two blocks: To extract content block similarity between two blocks must be find out. For this block feature vectors of two blocks are used. These features are number of images, number of terms etc. If a feature is present in a block then its corresponding entry in the feature vector is one otherwise it is zero. Two blocks are identical if the similarity feature between two bocks is greater than a threshold value.

## V. CONCLUSION

Many researchers have been developed several approaches for extract main content from web pages. Most of the approaches based on the only DOM tree. Here, the proposed system based on the content extraction algorithm which is based on DOM tree , helps in extracting similar blocks across different web pages obtained from different web sites. Also page segmentation algorithm used here for partitioning the web page into number of blocks is partition the web page into the homogeneous blocks. Many times news articles written by global news agencies appear in many news papers. User wants only one of these several copies of articles. These copies of articles differ only in their non- content blocks, so by separating non-content blocks from content blocks these same copies can be identified. So this proposed system helps in extracting these similar blocks present in many web sites.

## VI. FUTURE SCOPE

Here is the system works on extraction of similar contents from multiple web pages. This system improves the search result of the user. In future work we will detect the noisy contents in the web page with the extraction of the similar contents from the web pages.

## VII. ACKNOLEDGEMENT

I am Priyanka Chaudhari and I am greatly thankful to Prof. Seema Ladhe for her guidance. We also thank the college authorities, PG coordinator and Principal Sir for providing the required infrastructure and support. Finally, we would like to extend a heartfelt gratitude to M.E. Staff and friends.

## REFRENCES

[1] Divya C.,"Mining Contents in web pages and Ranking of Web Pages Using Cosne Similarity", International Journal of Science and Research (IJSR) April 2014

[2] Y. Li and J. Yang, "A Novel Method to Extract Informative Blocks from Web Pages", IEEE. DOI 10. 1109/ *JCAI, 2009*.

[3] P. M. Joshi, and S. Liu, " Web Document Text and Images Extraction using DOM Analysis and Natural Language Processing", ACM, DocEng, 2009.

[4] Y. Fu, D. Yang, and S. Tang,"Using XPath to Discover Informative Content Blocks of Web Pages", IEEE. DOI 10.1109/SKG, 2007.

[5] Rupesh R. Mehta, Pabitra Mitra, Harish Karnick, "Extracting Semantic Structure of Web Documents Using Content and Visual Information", *WWW 2005*, May 10–14, 2005, Chiba, Japan,ACM 1-59593-051-5/05/0005.

[6] L. Yi, B. Liu, and X. Li, "Eliminating Noisy Information in Web Pages for Data Mining", in Proc. ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-2003).

[7] D. Cai, S. Yu, J. R. Wen, and W. Y. Ma, "VIPS: a Vision-based Page Segmentation Algorithm", Technical Report, MSR-TR, Nov. 1, 2003.

[8] S. H. Lin and J. M .Ho, "Discovering Informative Content Blocks from Web Documents", in Pro. ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, pp.588-593, July 2002.

[9] S. Gupta, G. Kaiser, D. Neistadt, and P. Grimm,"DOM-based Content Extraction of HTML Documents", *Pro. 12 th International* Conference on WWW, ISBN: 1-58113-680-3, 2003.

[10] Bing Liu, Robert Grossman, Yanhong Zhai, "Mining Data Records in Web Pages", *Conference '00*, Month 1-2, 2000