02- Days, International Conference on Recent Trends in Engineering Science, Humanities and Management
Sri S Ramasamy Naidu Memorial College, Sattur, Tamil Nadu, India (RTESHM-17)
02nd-03rd February 2017, www.conferenceworld.in ISBN: 978-93-86171-18-4

A STUDY ON SEGMENTATION TECHNIQUES FOR WEB MINING

R. Amuthavalli[1] and P. Balamurugan[2,]

MPhil Research Scholar[1], Assistant Professor[2]

Department of Computer Science,Government Arts College, Coimbatore, (India)

ABSTRACT

Web mining is the process of mining information from the large set of data. Today the web has become the largestinformation source for people. Web pages are smallest and undividable units, but a web page as a whole may not be appropriate to represent a single topic. This contains various contents such as navigation, decoration, interaction and contact information, which are not related to the topic of the web-page called noise.The segmentation of informative and non-informative segments such as noises is important pre-processing steps in data mining. The main purpose of this paper is to review and discuss the major research work that has been done in this area and identifying the challenges and issues.

Keywords: Web Mining, Web Page Segmentation And Noises

I. INTRODUCTION

The Internet is an extraordinary resource that full of the information, which is strongly needed by many users and applications for different purposes. However this resource is composed of web pages that contain different types of information and extra data which affect the process of extract the intended information from the web. The difficulties of extracting the relevance content from the web come from unstructured formatting of the files where the intended data resides.

Most of the web documents contain extra information which is not relevant to the main article, but used for visual layout concerns among the desired contents. The redundant information in web document, like advertisement banner, copyright information and navigation menus assumed as noisy data for many applications. These noisy data considered harmful to many applications that related to the web, since it affects and reduces the performance of their process.There are many applications work to collect and manipulate the relevant data on web pages, for example web mining applications, which include classification and clustering of web pages, detecting of replicated web pages and information retrieval applications.The process of detecting and removing noisy data out of the web page in most of the approaches composed of two phases: (1) web page segmentation and (2) Removing noisy.

ISBN: 978-93-86171-18-4

II. WEB PAGE SEGMENTATION - A SHORT REVIEW

Web page segmentation and data cleaning are essential step in structure web data extraction. Identifying the web page main content not important region can greatly improve the performance of the extraction process.**Roberto Penerai et. al [8]** proposed a **TPS (Tag Path Sequence)** algorithm searches for position in the TPS where it is possible to split in two regions. This algorithm is very effective in identifying the main content block of several major websites. The TPS consist of a sequence of symbol, each one representing a different tag path.

To address the problem of resource limitation of small screen devices, a unique methodology of web page segmentation with tag path clustering is proposed by **Aruljothi et.al[3],** that reduces the memory space demand of the small hand-held devices. For segmenting web pages, both reappearance key patterns detection technique and page layout information are used to provide better segmentation accuracy.

Sandeep Kauret.al. [10]proposed DOM based page segmentation technique. Initially a XML web page is converted into DOM tree. DOM based page segmentation which converts the pages into blocks and regions. Performance of Web content extraction is analyzed based on complexity and efficiency on proposed algorithm.

S. Debnath [5]proposed two new algorithms, ContentExtractor, and Feature Extractor. The algorithms identify primary content blocks by i) looking for blocks that do not occur a large number of times across webpages and ii) looking for blocks with desired features respectively. The identify the primary content blocks with high precision and recall, reduce the storage requirement for search engines, result in smaller indexes and thereby faster search times, and better user satisfaction.

Ruihua Song et.al. [9] proposed how to find a model to automatically assign importance values to blocks in web page. That defines the block importance estimation as a learning problem. VIPS algorithm used to partition a web page into Semantic blocks with a hierarchy structure. Two types of features extracted for each block namely spatial feature (such as position and size), content feature (such as no of images and links). A Learning algorithm SVM and Neural Network is applied to train various block importance models based on the features.

X. Liu .et.al. [8]proposed a novel web page segmentation algorithm based on finding the Gomory-Hu tree in a planar graph. The algorithm firstly distills vision and structure information from a web page to construct a weighted undirected graph, whose vertices are the leaf nodes of the DOM tree and the edges represent the visible position relationship between vertices. Then it partitions the graph with the Gomory-Hu tree based clustering algorithm.

III. WEB PAGE SEGMENTATION TECHNIQUES

In this sectiondiscuss the recent developments in web page segmentation and search optimization techniques. Some commonly used approaches are,

02nd-03rd February 2017, www.conferenceworld.in ISBN: 978-93-86171-18-4

- **DOM Tree Based Approach**
- **Vision Based Approach**
- **Text Based Approach**
- **Hybrid Approach**
- **Graph Theoretic Based approach**

DOM Tree Based Approach: The Document Object Model(DOM) is an application programming interface(API) for valid HTML and well-formed XML documents. It defines the logical structure of documents and the way a document is accessed and manipulated. In the DOM-based segmentation approach, an HTML document is represented as DOM tree, which provides a useful structure for a web page, but often not accurate enough to identify different semantic blocks in a web page.

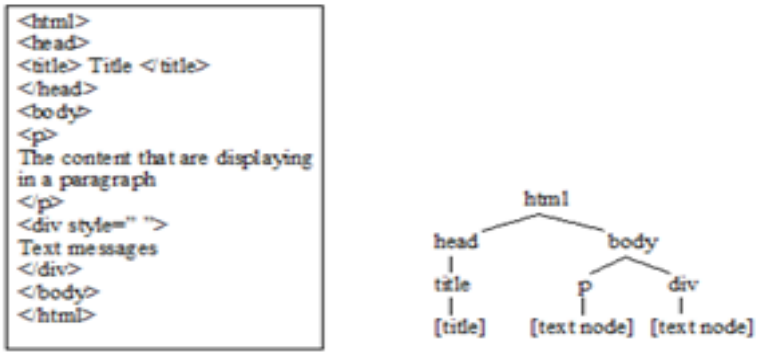

Fig 1: DOM Tree Segmentation from HTML source code.

Vision Based Approach: Visual approaches segment the web page from the browser-side perspective as it is rendered. They used to partition the page into separators, such as lines, whitespaces and images, content and build a content structure out of this information. Also it requires external resourcessuch as CSS files and images in order to work correctly.

Text Based Approaches: Text content based segmentation method[12] to extract content text from diverse webpages byusing the HTML document's tag ratios. Which describe how to compute tag ratios on a line-by-line basis and then cluster the resulting histogram into content and non-content area.These work is not on structured extraction, but instead, on indexing and clustering of web sites.

Example: ***Snippet of a webpage news article.***

1.

2.

ISBN: 978-93-86171-18-4

3. James Smith

4. OKLAHOMA CITY - Police were told that. . .

5. . . . The Oklahoman reported Sunday.

 Jones had. . .

6.

The Tag Ratios for these six lines are computed as follows.

1. Text=0, Tags=1, TR=0

2. Text=0, Tags=1, TR=0

3. Text=11, Tags=2, TR=5.5

4. Text=37, Tags=0, TR=37

5. Text=41, Tags=2, TR=20.5

6. Text=0, Tags=2, TR=0

Hybrid Approach[7]: It uses Concept-Based Mining Model and Hierarchical Agglomerative Clustering (HAC) as a document clustering algorithm along with link based algorithm to cluster the web documents considering both the content of web page as well as and the links of a web page in order to use as much information as possible for the clustering. Fig. 2 shows the architecture that uses the Hybrid Approach (HAC algorithm and Link based algorithm) in order to cluster the documents focusing on both the contents of the web page as well as hyperlinks in the pages.

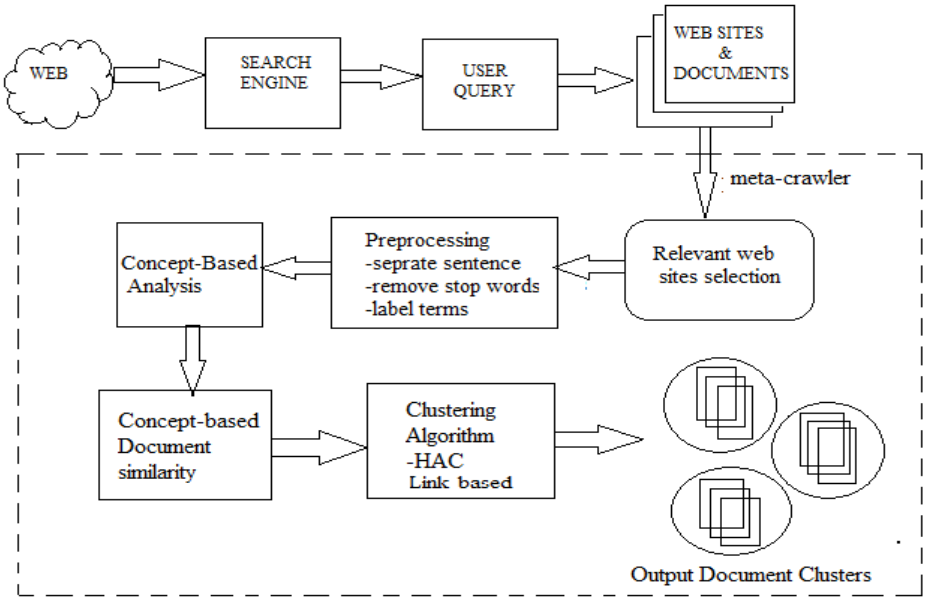

Fig 2: Hybrid Approach System Architecture

ISBN: 978-93-86171-18-4

Graph Theoretic Approach[4]:Segmenting a webpage into visuallyand semantically cohesive pieces. These approaches isbased on formulating an appropriate optimization problem on weighted graphs, where the weights capture if two nodes in the DOM tree should be placed together or apart inthe segmentation.

Fig 3. Example web page for segment a weighted graph

IV. CONCLUSION AND FUTURE WORK

In this study work discussed the various approaches that concern with the problem of web page segmentation. The review explores various approaches and explained the difference between them, also it presents the different algorithm's approach and the application area related to each algorithm. The results of this review showed the importance of web page segmentation as an essential step to various applications that related to it.

REFERENCES

[1] Ananthi J.et.al. , A Survey on Web Content Mining Methods and Applications for Information Extraction from Online Shopping Sites, International Journal of Computer Science and information Technologies. Vol. 5(3), 2014, pp. 4091-4094.

[2] AndresSanoja and S. Gancarski, Block-o-Matic: Web Page Segmentation Tool and its Evaluation, 2009.

[3] S. Aruljothi and S. Sivaranjani, et.al, Web page segmentation for small screen Devices Using Tag Path Clustering Approach, IJCSE, Vol 5, 2013.

[4] Chakrabarthi D.et al., A graph-theoretic approach to webpage Segmentation, In Proceeding of the International conference on World Wide Web, 2008.

[5] Debnath S, Mitra P., and C.L. Giles,Automatic Extraction of Information Blocks from Web pages, In ACM Symposium on Applied Computing, 2005, pp.1722-1726.

[6] H. F. Eldirdiery and A. H. Ahmed, Detecting and removing noisy data using Text Density Approach, International journal of Computer Applications, Vol. 112, 2015.

ISBN: 978-93-86171-18-4

[7] S. Pralhad, G. Gamare1, and A. Patil, Web Document Clustering using Hybrid Approach in data mining, International Journal of Research in Advent Technology,*Vol.3, 2015*.

[8] R. Panerai, F. Carina andDorneles,Automatic Web Page Segmentation and Noise Removal for Structure Extraction using Tag Path sequences, Vol.4, JIDM, 2013, pp. 173-187.

[9] Ruihua Songand H. Liu. et.al, Learning Block Importance Model for Web pages, ACM, 2004.

[10]S. Kaur,and A. Tyagi et.al, Noise Reduction and Content Extraction from Web Pages Using DOM based page segmentation,Vol. 5(6), IJCTA. Dec 2014.

[11] R. Sharma and M. Bhatia,Eliminating the Noise from webpages using Page Replacement algorithm,International Journal of Computer Science and Information Technologies, Vol. 5(3), pp. 2014, 3066-3068.

[12] T. Weninger, and HsuW.H., et al., Content extraction via tag rations, In Proceedings of the International Conference on World Wide Web, 2010.

[13] Y. K. Patel and N. Limbad, Noise Removal from Web Pages for Web Content Mining,IJARIIE-ISSN(O)-2395-4396. Vol-2, 2016.

[14] Liu X et. al., Segmenting Webpage with Gomory-Hu TreeBased Clustering,JournalofSoftware, Vol. 6, No. 12, 2011.