

Visual Webpage Content Segmentation and Retrieval based on n-Grams

Kumud Jaglan

UIET, Kurukshetra University, Kurukshetra, India

Dr. Kulvinder Singh

UIET, Kurukshetra University, Kurukshetra, India

Vipul Jaglan

NIFTEM, Kundli

Abstract

Web documents are often viewed as complicated objects which frequently contain multiple entities every of which may represent a separate unit. Though, most processing requests applications for the web and web content because of the smallest indivisible components and knowledge Extraction from Web Pages has continually trusted comprehensive human involvement within the sort of hand crafted extraction algorithms or scripts using usual expressions. Preceding works usually flout the underlying content segments that are composed of un-important knowledge like net ads and knowledge moot to the users. This paper resolve these subjects, we tend to endorsed n-gram established website segmentation algorithmic program that used the density for segmenting the webpage lacking hoping on the DOM tree for the segmentation method.

Keywords: Web page classification& segmentation, Vision based page segmentation, Web information extraction

I. INTRODUCTION

Web documents can be viewed as complex objects which often contain multiple entities each of which can represent a standalone unit. However, most information processing applications developed for the web, consider web pages as the smallest undividable units. This fact is best illustrated by web information retrieval engines whose results are presented in the form of links to web documents rather than to the exact regions within these documents that directly match a user's query [1]. Various approaches that have been adopted in carrying out the segmentation task have been identified and their advantages and disadvantages associated with each. We describe web page segmentation methods & compare them from theoretical point of view fixed-length technique is that no semantic info is taken into consideration within. DOM provides every online page with a fine-grained structure. VIPS discards content analysis and produce blocks based on the visual cues of web pages. In general, passages can be categorized into three classes: discourse, semantic, and window [2]. Discourse passages rely on the logical structure of the documents marked by punctuation, such as sentences, paragraphs and sections. Semantic passages are obtained by partitioning a document into topics or sub-topics according to its semantic structure. A third type of passages, fixed-length passages or windows, are defined to contain fixed number of words. While directly adopting these passage definitions for partitioning web pages is feasible, there exist some new characteristics in web pages which can be utilized. We describe each of them below:

1) Two-Dimension Logical Structure

Different from plain-text documents, web pages have a 2-D view and a more sophisticated internal content structure. Each region of a web page could have relationships with regions from up to four directions and contain or be contained in some other regions. A content structure in semantic level exists for most pages and can be used to enhance retrieval.

2) Visual Layout Presentation

To facilitate browsing and attract attention, web pages usually contain much visual information in the tags and properties in HTML. Typical visual hints include lines, blank areas, colors, pictures, fonts, etc. Visual cues are very helpful to detect the semantic regions in web pages.

Web page segmentation has been done to address a problem in different fields including mobile web, web page phishing, duplicate detection, information retrieval, information classification, information extraction, evaluating visual quality (aesthetics), web page classification / clustering, caching, publishing dynamic content, semantic annotation, web accessibility etc. [3].

A. Segmentation Methods

In this section, we describe web page segmentation methods & compare them from theoretical point of view. And also shows natural conformity between these segmentation methods and traditional passage retrieval methods.

1) Fixed-length Page Segmentation (Fixed PS)

In traditional text retrieval, fixed-length passages, or windows, are used to overcome the issue of length normalization. A fixed length passage contains fixed range of continuous words. An overlapped window approach during which the primary window in one document starts at the primary prevalence of a question term, and consequent windows half-overlap preceding ones [6].

For internet documents, fixed-length page segmentation is just like traditional window approach except that every HTML tags and attributes are removed. The length of window is that the solely parameter and is recommended to be two hundred or 250 from past expertise..

2) DOM-based Page Segmentation (Dom PS)

DOM provides every online page with a fine-grained structure, that illustrates not solely the content however conjointly the presentation of the page. In general, similar to discourse passages, the blocks made by DOM-based strategies tend to partition pages supported their predefined syntactic structure, i.e., the hypertext markup language tags.

There are some approaches that take into consideration the problem of page segmentation, however there's no consistent way to do it and, to the most effective of our knowledge, few works are done on applying DOM primarily based page segmentation strategies on internet info retrieval. Some straightforward experiments are performed where sub-trees labeled with<TITLE>, <P>, <H1>~<H3> and <META> are treated as blocks, but the results are not encouraging. The reasons may lie in the following three aspects.

3) Vision-based Page Segmentation (VIPS)

People view a web page through a web browser and get a 2-D presentation which provides many visual cues to help distinguish different parts of the page, such as lines, blanks, images, colors, etc. For the sake of easy browsing and understanding, a closely packed block within the web page is much likely about a single semantic.

We have previously proposed a vision-based page segmentation method called VIPS. Similar to semantic passages, the blocks obtained by VIPS are based on the semantic structure of web pages. Traditional semantic passages are obtained based on content analysis which is very slow, difficult and inaccurate [7]. VIPS discards content analysis and produce blocks based on the visual cues of web pages. This method simulates how a user understands web layout structure based on his or her visual perception.

4) Hybrid Approaches

Although VIPS can distinguish multiple topics in web pages, it does not consider the document length normalization problem. As can be seen from this figure, the distribution of block length is very diverse. More than 40% of the blocks are only less than 10 words, and 10% blocks are larger than 500 words. Thus the varying length problem still exists even if we perform retrieval on block level. Since fixed-length windows show great consistence on dealing with the varying length problem, The Hybrid Page Segmentation which tries to take advantage of both visual information and fixed length.

II. RELATED WORK

Many web applications uses web page semantic structure and contents. Hattori et al. [1] provided a segmentation method by calculating the distance between content elements of HTML tags structure. Different web page segmentation methods were introduced based on visual and non-visual characteristics. In web information accessing, some researchers use database technique for building wrappers for the information extraction. For building wrappers, web documents are divided into parts. Diao et al. [2] provided segments of web page based on query processing using different HTML tags. Lin et al. [3] used only table tag for content blocks. Cai et al. [4] given an algorithm for extracting web based semantic structure. These semantic structures are in hierarchal nature which corresponds to a block. This algorithm made use of page layout feature. Alcic et al. [5] provided distance measures for content units on the basis of web page properties and DOM structure of web page. Nguyen et al. [6] proposed a method for segmenting a Web page into its semantic parts. There method segmented the page into blocks and then classifies the blocks. Bhardwaj, A. et al. [7], drafted that quick progress of the internet and web publishing methods craft countless data origins published as HTML pages on Web. Though, there was lot of redundant and irrelevant data additionally on web pages. Such data makes varied web excavating tasks such as web page scuttling, web page association, link established ranking, case distillation complex. This paper debated assorted ways for removing informative content from web pages and a new way for content extraction from web pages and density of links. Srivastava, S. et al.[8], presented that in the globe of data knowledge the adjustments happens rapidly. As the new technologies always adjusts the globe of data representation, the result was to find out relevant pieces of data cluttering alongside distracted features(like advertisements, links, scrollers etc.) in the finished web page. Data or functional content Kumud Jaglan1 IJECs Volume 4 Issue 6 June, 2015 Page No.12970-12973 Page 12971 extraction from the web pages (structured or semi structured) becomes a critical subject for web users and web miners. So the data extraction from the web page carries a huge importance. A mystifying mystery for data extraction is to depict the noisy area and its removal. They examine the DOM tree segmentation alongside class attribute established approach. The class attribute can be utilized alongside all HTML agents inside the 'BODY' serving of the document. It was utilized to craft disparate classes of an agent, whereas every single class can have its own properties. To assess the arrangement presentation countless examinations completed on disparate business, news, and entertainment websites. Chee Sheen Chan et al. [9], proposed an ontology based web page segmentation algorithm for extracting web images with its associated contextual information according to its semantic characteristics like picture annotation, clustering of pictures, inference of picture semantic content and picture indexing. Sanoja, A. et al. [10], depicted a web page segmentation framework. It wass a hybrid way inspired by automated

document processing methods and visual-based content segmentation techniques. A web page was associated alongside three structures: the DOM tree, the content construction and the logical structure. The DOM tree embodies the HTML agents of a page, the content construction organizes page objects according to content's groups and geometry and in the end the logical construction is the consequence of mapping content construction on the basis of the human-perceptible meaning that conforms the blocks.

III. PROPOSED WORK

Web Information Extraction (WIE) has always depended on extensive human involvement in the form of hand crafted extraction algorithms or hand crafted training examples. Furthermore the experienced user is needed to explicitly specify each relation which he has interest for extraction. Although information extraction from web has become increasingly automated, finding all possible interests relations for the information extraction from any web retrieval system is extremely problematic for large and dynamic platforms as the web. To enable users to issue various queries on heterogeneous sources, Web information extraction systems must move away architectures that require relationships to be specified before questioning time for those to discover all the possible relationships text.

Although WIE has received a lot of attention by researchers over the years however, most of the works are based on examining the HTML or the DOM tree of the Web pages. Web documents can be viewed as complex objects which often contain multiple entities each of which can represent a standalone unit. However, most information processing applications developed for the web, consider web pages as the smallest undividable units. Also the websites contain a mixture of objects with no unifying structure underneath, with differences in the authoring style and content much greater than in traditional collections of text documents the difficulty of information extraction task can be even more complex when multiple arrangements of attributes exist and typo-graphic errors occur in the input web documents.

Previously Vision based segmentation algorithms such as VIPS (Vision-based Page Segmentation) algorithm exists to extract the semantic structure from web pages. These semantic structures are hierarchical structures, these hierarchical structures represent corresponds to a block in the web page. In VIPS each node is assigned a Degree to indicate visual perception of the block. The Vision-based Page Segmentation algorithm makes use of page layout feature however VIPS ignores the underlying content as segments can be composed of un-important information such as web ads, to solve these issues we proposed a n-gram based web page segmentation algorithm. That utilized the n-grams for segmenting the webpage without relying on the DOM tree for the segmentation process.

A. Problem Formulation

Previously Vision based segmentation algorithms such as VIPS (Vision-based Page Segmentation) algorithm exists to extract the semantic structure from web pages. These semantic structures are hierarchical structures, these hierarchical structures represent corresponds to a block in the web page. In VIPS each node is assigned a Degree to indicate visual perception of the block. The Vision-based Page Segmentation algorithm makes use of page layout feature however VIPS ignores the underlying content as segments can be composed of un-important information such as web ads, to solve these issues we proposed a n-gram based web page segmentation algorithm. That utilized the n-grams for segmenting the webpage without relying on the DOM tree for the segmentation process.

B. Research Objectives

Previously Vision based segmentation algorithms such as VIPS (Vision-based Page Segmentation) algorithm exists to extract the semantic structure from web pages. These semantic structures are hierarchical structures, these hierarchical structures represent corresponds to a block in the web page. In VIPS each node is assigned a Degree to indicate visual perception of the block. The Vision-based Page Segmentation algorithm makes use of page layout feature however VIPS ignores the underlying content as segments can be composed of un-important information such as web ads.

- 1) To propose novel language-independent Tree based web segmentation approach that can be used for partitioning web pages.
- 2) To extract the web segments relying on the DOM tree for the segmentation process of WebPages. This approach will utilize the words frequency and their probability for web data extraction and item extraction.
- 3) To construct and populate the visual tree structure representing visual regions from web pages using HTML structures and to improve the performance of vision approach by making it.
- 4) To keep only relevant information inside the tree and removing the meaningless content.
- 5) To evaluate the process using suitable tools and methodologies.

IV. FLOW CHART

A. Flowchart of VisionIE

An html source is taken as input and with the help of parser tree DOM tree of webpage is constructed. On the basis of style information preprocesses and removes noise from the Source utilizing. Now visit on all valid nodes specified in the allowed tags.

Construct the Density of the nodes in the tree utilizing n-grams. Select Remaining Nodes and Segment Nodes into density regions based on n-grams. Group the equivalent regions into one. Repeat the process for remaining Segments and Calculate Segment area. And extract the Segment text

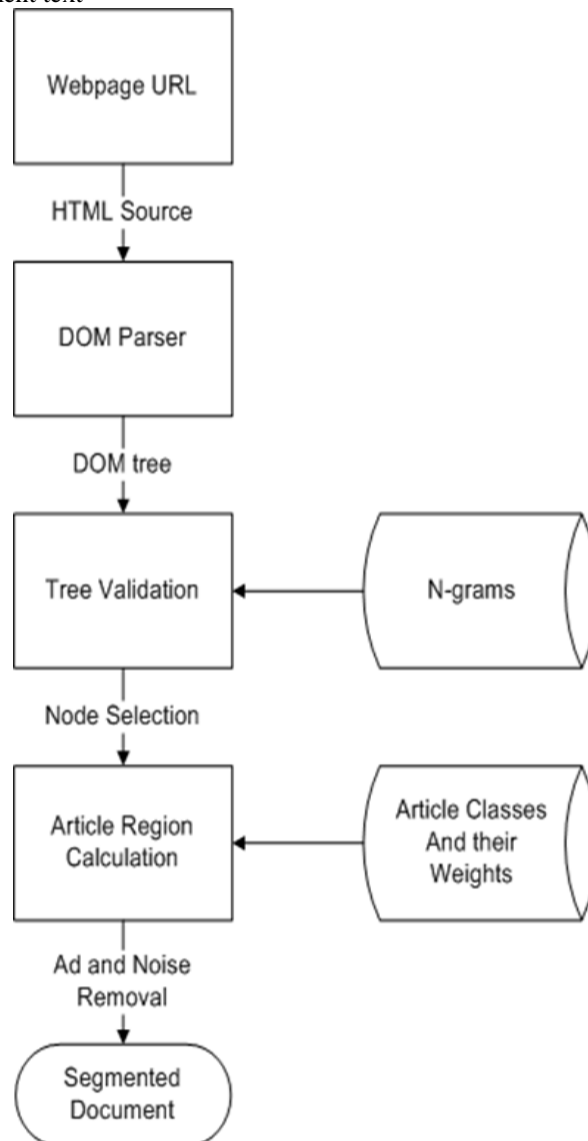


Fig. 1: Flowchart for VisionIE

V. ALGORITHM OF VISIONIE

- Input: URL of the page being segmented
 - Args: VisionIE Initializing Arguments
- Allowed Tags: DIV, DD, TD.....
- 1) Populate the DOM tree by getting Source code of the page, i.e HTML content with CSS
 - 2) Preprocess and Remove Noise from the Source utilizing Style information
 - 3) Visit All Valid Nodes Specified in the Allowed Tags
 - 4) Construct the Density of the nodes in the tree utilizing n-grams. The Density of the Tags using min and max density thresholds, remove nodes outside the min-max region threshold.
 - 5) Select Remaining Nodes and Segment Nodes into density regions based on ngrams
 - 6) Group equivalent Regions by merging based on tag formation and by Finding Segments with allowed maximum inter region distance
 - 7) Select Remaining Segments and Calculate Segment area.
 - 8) for each Segment in Selected Segments do:
 - 9) if Text with Given Density is Found in Segment Area
 - a) Extract the Segment text

- 10) else
- a) Continue to next segment

VI. RESULTS AND ANALYSIS

A. HTML Source Size of Pages

Web search engines and some other sites use Web crawling or spidering software to update their web content or indexes of others sites' web content. Web crawlers can copy all the pages they visit for later processing by a search engine which indexes the downloaded pages so the users can search much more efficiently. All web crawlers allow users to view the HTML or other source code of any of the web pages they visit. For example, a user can view the code used to generate. The Search Engines use document indexing and Parsing to analyze the page content.

Table 4.2:
Original Source Size as the Browser or a Search Engine Sees

URL	HTML Size
1	81820
2	35551
3	27260
4	23859
5	18566
6	71760

Web page parsing breaks apart the components (words) of a document or other form of media for insertion into the forward and inverted indices. The words found are called tokens, and so, in the context of search engine indexing and natural language processing, parsing is more commonly referred to as tokenization. It is also sometimes called word boundary disambiguation, tagging, text segmentation, content analysis, text analysis, text mining, concordance generation, speech segmentation, lexing, or lexical analysis.

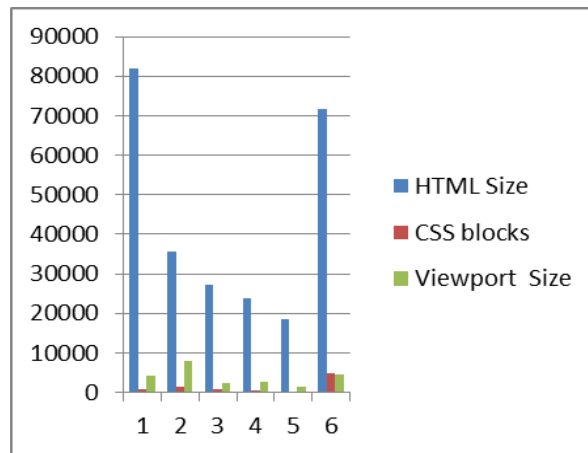


Fig. 2: HTML Source Size of Pages

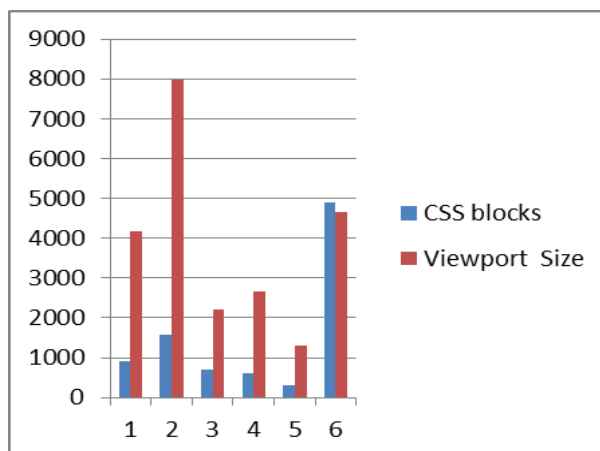


Fig. 3: HTML Viewport Size with respect to CSS Block Size

```
Step 1: Initilize DOM tree...
Visiting all html nodes and select character density greater than the mindensit
threshold.
Calculating initial groups...
Merging groups...
Creating segments...
Calculating distances from max segment...
Printing segments...

-----
Segment: /html/body/div[1]/div/div/div/div[2]/div[2]/div[2]/div
-----

Tag: div
Class: hp-section-side
Id:
Level: 10
Parts: 2
Density: 205
Distance from max: 0.8
Has title on parents: (True, 11)
Full text:
```

Fig. 4: Output of Proposed work in Console Showing Statistics with Text Segments

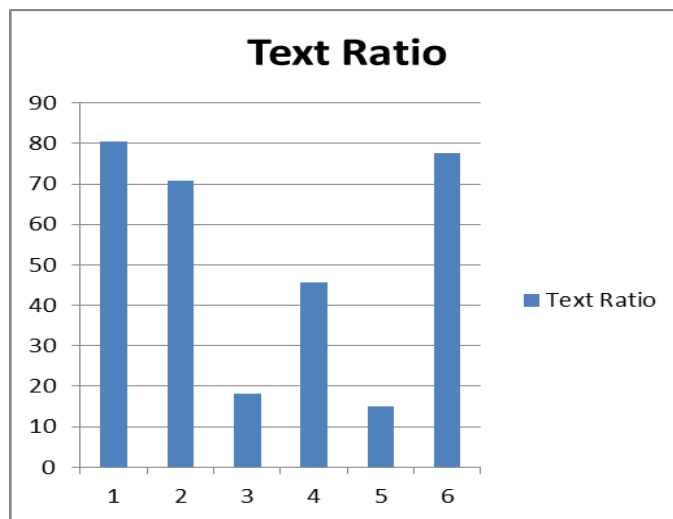


Fig. 5: Extracted Text Ratio Compared with Original HTML Size and Segmented Text

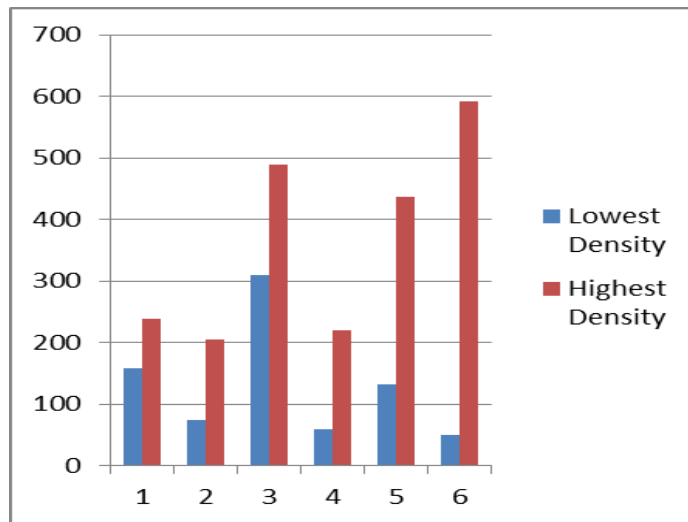


Fig. 6: Lowest and Highest Density of Text Segments

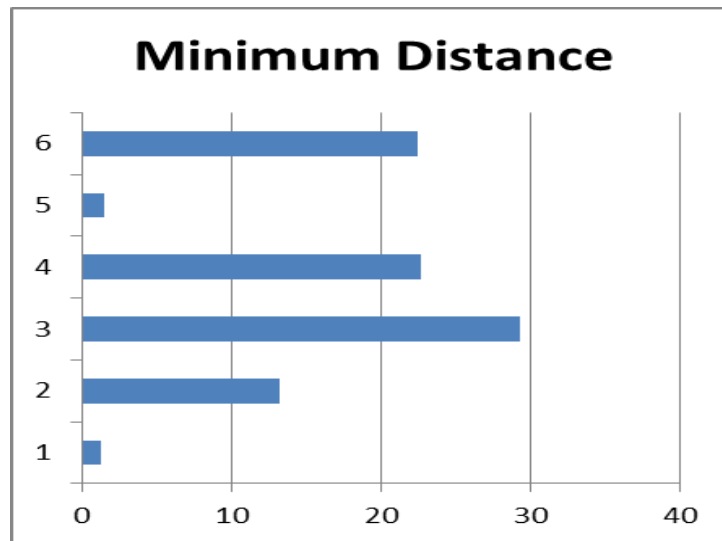


Fig. 7: Minimum Distances from Maximum Segment

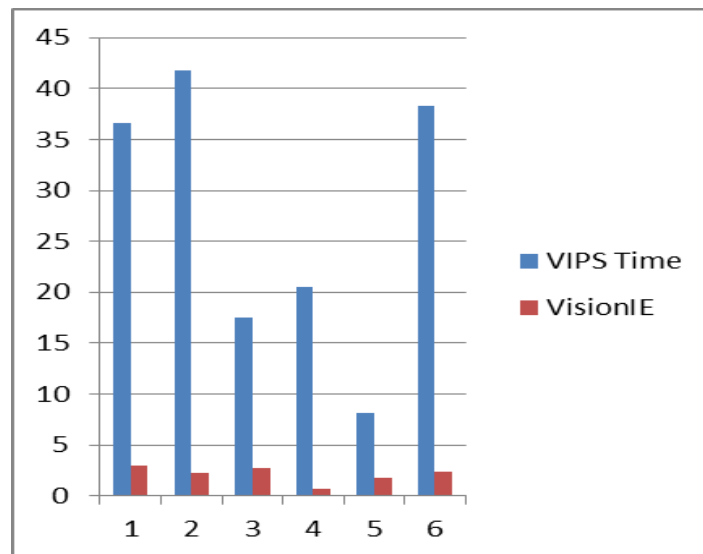


Fig. 8: Execution time of VIPS vs. VisionIE

Fig above shows the Execution time between VIPS and VisionIE in Seconds, The Text extraction using VisionIE takes much less time than that of VIPS algorithm, on average the VisionIE is 12-15 times faster than that of VIPS. VIPS on average takes 168 seconds and VisionIE takes about 12 seconds. making VisionIE an Efficient algorithm than CSS segmenting VIPS.

VII. CONCLUSION & FUTURE SCOPE

A. Conclusion

Web Information Extraction (WIE) has always depended on extensive human involvement in the form of hand crafted extraction algorithms or hand crafted training examples. Furthermore the experienced user is needed to explicitly specify each relation which he has interest for extraction. Although information extraction from web has become increasingly automated, finding all possible interests relations for the information extraction from any web retrieval system is extremely problematic for large and dynamic platforms as the web. To enable users to issue various queries on heterogeneous sources, Web information extraction systems must move away architectures that require relationships to be specified before questioning time for those to discover all the possible relationships text.

Although WIE has received a lot of attention by researchers over the years however, most of the works are based on examining the HTML or the DOM tree of the Web pages. Web documents can be viewed as complex objects which often contain multiple entities each of which can represent a standalone unit. However, most information processing applications developed for the web, consider web pages as the smallest undividable units. Also the websites contain a mixture of objects with no unifying structure underneath, with differences in the authoring style and content much greater than in traditional collections

of text documents the difficulty of information extraction task can be even more complex when multiple arrangements of attributes exist and typo-graphic errors occur in the input web documents.

Previously Vision based segmentation algorithms such as VIPS (Vision-based Page Segmentation) algorithm exists to extract the semantic structure from web pages. These semantic structures are hierarchical structures, these hierarchical structures represent corresponds to a block in the web page. In VIPS each node is assigned a Degree to indicate visual perception of the block. The Vision-based Page Segmentation algorithm makes use of page layout feature however VIPS ignores the underlying content as segments can be composed of un-important information such as web ads, to solve these issues we proposed a n-gram based web page segmentation algorithm. That utilized the n-grams for segmenting the webpage without relying on the DOM tree for the segmentation process.

B. Future Work

The web structure mining analyzes the structure of hyperlinks within the web. The structure mining can be quite useful in determining the correlation between two or more Web pages. This understood connection conveys a useful tool for mapping completion and third party, such as resellers or consumers. This web information retrieval mechanism can help result links to the exact regions within these documents that directly match a user's query instead of to web documents. Yet another improved retrieval performance can be attained by considering the web pages having underlying structure with segments as fragments of the page. This can potentially result in a targeted model of query and search. For matter of fact, many applications can be assisted by the operation of semantically independent units or fragments. Such applications comprise web browsers for cell phones, Smart phones and non-PC terminals including as well as text summarization applications.

Web pages are accessed by queries are submitted to the web databases and returned data is wrapped inside HTML Web pages. Extracting structured data from deep Web pages is a challenging problem due to the underlying structures of web pages. A large number of techniques have been proposed to solve this problem, but all of them have inherent limitations because they are language dependent. VisionIE provide the underlying content as segments can be composed of un-important information such as web ads. In future, a novel language-independent n-gram based approach can be Web-page implemented. An n-gram based web page segmentation algorithm can be implemented for extraction of web segments without relying on the DOM tree for the segmentation process. This approach will primarily utilize the word grams and their probability for web data extraction and segmentation, with data record extraction and data item extraction.

REFERENCES

- [1] Hj G. Hattori, K. Hoashi, K. Matsumoto, and F. Sugaya. Robust web page segmentation for mobile terminal using content distances and page layout information. In WWW, pages 361–370, 2007.
- [2] Y. Diao, H. Lu, S. Chen, and Z. Tian. Toward learning based web query processing. Proceedings of the 26th International Conference on Very Large Data Bases, pages 317–328, San Francisco, CA, USA, 2000.
- [3] S.-H. Lin and J.-M. Ho. Discovering informative Content blocks from web documents. Proceedings of the 8th international conference on Knowledge Discovery and data mining (SIGKDD), 2002.
- [4] D. Cai, S. Yu, and J.-r. Wen. VIPS: a Vision-based Page Segmentation Algorithm. Microsoft Technical Report (MSRTR-2003-79), 2003.
- [5] S. Alcic and S. Conrad. Page segmentation by web content clustering. In Proceedings of the International Conference on Web Intelligence, Mining and Semantics, WIMS '11, pages 24:1–24:9, New York, NY, USA, 2011.
- [6] Nguyen, Cong Kinh, Laurence Likforman-Sulem, J-C. Moissinac, Claudie Faure, and Jérémy Lardon. "Web document analysis based on visual segmentation and page rendering." In Document Analysis Systems (DAS), 2012 10th IAPR International Workshop on, pp. 354-358. IEEE, 2012.
- [7] Bhardwaj, Aanshi, and Veenu Mangat. "A novel approach for content extraction from web pages." Engineering and Computational Sciences (RAECS), 2014 Recent Advances in. IEEE, 2014
- [8] Srivastava, Shobhit, Mohd Haroon, and Abhishek Bajaj. "Web document information extraction using classes attribute approach." Computer and Communication Technology (ICCCT), 2013 4th International Conference on. IEEE, 2013.
- [9] Chan, Chee Sheen, Adel Johar, Jer Lang Hong, and Wei Wei Goh. "Ontological based webpage segmentation." In Fuzzy Systems and Knowledge Discovery (FSKD), 2013 10th International Conference on, pp. 883-887. IEEE, 2013.
- [10] Sanoja, Andres, and Stephane Gancarski. "Block-o-Matic: A web page segmentation framework." In Multimedia Computing and Systems (ICMCS), 2014 International Conference on, pp. 595-600. IEEE, 2014.
- [11] Nikolaos Pappas, Georgios Katsimpras, and Efstathios Stamatatos. "Extracting informative textual parts from web pages containing user-generated content." In Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies, p. 4. ACM, 2012.
- [12] Yaohui Li, Li Xia Wang, Jian Xiong Wang, Jie Yue, and Ming Zhan Zhao. "An approach of web page information extraction." Applied Mechanics and Materials 347 (2013): 2479- 2482.