Deep Web Data Extraction by Using Vision Approach for Multi-Region

Shweta Dhall[1], Parikshit Singla[2]

[1] *Student of Department of Computer Science and Engineering, DVIET, Karnal , Haryana*

[2]*Assist. Professor Department of Computer Science and Engineering , DVIET, Karnal ,Haryana*

***Abstract*— Web Information Extraction (WIE) is entirely dependent on comprehensive human involvement in the form of hand crafted algorithms used for extraction. Furthermore the experienced user is demanded to explicitly enumerate every single relation that he has attention for extraction. Even though data extraction from web has come to be increasingly automated, discovering all probable hobbies relations for the data extraction from each web retrieval arrangement is tremendously problematic for colossal and vibrant periods as the web. Even though WIE has consented a lot of attention by researchers above the years though, most of the works are established on scrutinizing the HTML Web pages. Web documents can be believed as convoluted objects that frequently encompass several entities every single of that can embody a standalone unit. Though, most data processing requests industrialized for the web, ponder web pages as the smallest undividable units. Preceding works flout the underlying content as segments can be composed of un-important data such as web ads, to resolve these subjects we counseled an n-gram established web page segmentation algorithm. That utilized the density for segmenting the webpage lacking relying on the DOM tree for the segmentation process.**

***Index Terms*—Document Object Model, Vision Based Page Segmentation, Web Information Extraction, Web Page Segmentation.**

I. INTRODUCTION

Web documents can be considered as complex objects which often contain multiple entities each of which can represent a standalone unit. However, many of information processing applications developed for the web , contemplate web pages as the smallest undividable units. A better retrieval performance can be achieved by considering the page not as an undividable unit but as having an underlying semantic structure with topically relevant segments.

1.1 Web Information Extraction

Web Information Extraction (WIE) has conventionally relied on comprehensive human involvement in the form of hand-crafted extraction laws or hand-tagged training examples. Data Extraction task is described by its input webpage or website and its extraction target. The input can be unstructured documents as free text that are composed in usual speech or the semi-structured documents which are pervasive on the Web, such as tables or itemized and enumerated lists. A wrapper was primarily described as a constituent in a data integration arrangement that aims at bestowing a solitary uniform query interface to admission several data sources. In a data integration arrangement, a wrapper is usually a plan that “wraps” a data basis such that the data integration arrangement can admission that data basis lacking changing its core query responding mechanism. In the case whereas the data basis is a Web server, a wrapper have to query the Web server to amass the emerging pages via HTTP protocols, present data extraction to remove the contents in the HTML documents, and in the end incorporate alongside supplementary data sources. Amid the three procedures, data extraction has consented most attentions and a little use wrappers to denote extractor programs. Therefore, the words extractors and wrappers can be utilized interchangeably.

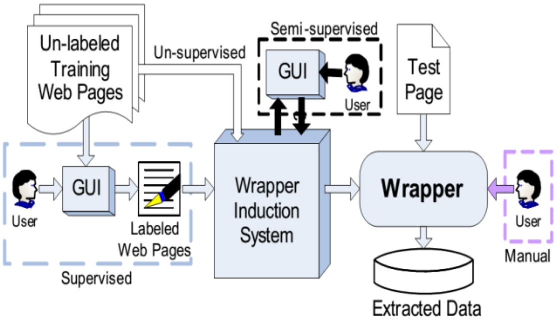


Fig. 1.1 A general view of Web Information Extraction systems [1]

1.2 Webpage Segmentation

Web page segmentation is generously partitioning web pages into several blocks. Segmentation methods can be categorized in three classes: discourse, semantic, and window. Discourse methods rely on the logical construction of the documents marked by punctuation, like sentences, paragraphs and sections. Semantic methods are obtained by partitioning a document into cases or sub-topics according to its semantic structure. A third kind of methods, fixed-length methods or windows, are described to encompass fixed number of words. As undeviatingly adopting these method definitions for partitioning of web pages is viable, there continue a little new characteristics in web pages that can be utilized. We delineate every single of them below:

1. **Two-Dimension Logical Structure –** web pages have a 2-D view and a more confined internal content structure. Each part of a web page could have relationships with other regions and contain or be contained in some other regions.

2. **Visual Layout Presentation –** To facilitate browsing, web pages usually contain much visual information in form of tags and properties in HTML. Typical visual hints can includes lines, blank areas, colors, pictures, fonts and many more. Visual cues are very helpful in detecting semantic regions in web pages.

1.3 Segmentation Methods

In this section, we describe the Major web page segmentation methods.

1. Fixed-length Page Segmentation (Fixed PS)

In established text retrieval, fixed-length methods, or windows, are utilized to vanquish the difficulty of length normalization. A fixed length method encompasses fixed number of constant words. An overlapped window way in that the early window in one document starts at the early occurrence of a query word, and consecutive windows half-overlap preceding ones [4].

For web documents, fixed-length page segmentation [5] is identical to established window way except that all the HTML tags and qualities are removed. The length of window is the merely parameter and is counseled to be 200 or 250 from past experience. Although its simplicity, fixed-length segmentation is extremely robust and competent for enhancing presentation, chiefly for collections alongside long or mixed-length documents. The main shortcoming of the fixed-length method is that no semantic data is seized into report in the segmentation process.

2. DOM-based Page Segmentation (Dom PS)

DOM provides every single web page alongside a fine-grained construction, that illustrates not merely the content but additionally the presentation of the page. In finish, comparable to discourse methods, the blocks produced by DOM-based methods incline to partition pages established on their predefined syntactic construction, i.e., the HTML tags.

There are a little ways that seize into report the setback of page segmentation, but there is no consistent method to do it and, to the best of our vision, insufficient works are completed on requesting DOM established page segmentation methods on web data retrieval. A little easy examinations are gave whereas sub-trees tagged alongside <TITLE>, <P>, <H1>~<H3> and <META> are indulged as blocks, but the aftermath are not encouraging. The reasons could lie in the pursuing three aspects. First, DOM is yet a linear construction, so visually adjacent blocks could be distant from every single supplementary in the construction and departed wrongly. Secondly, tags such as <TABLE> and <P> are utilized not merely for content presentation but additionally for layout structuring. It is consequently tough to attain the appropriate segmentation granularity. Thirdly, in countless cases DOM prefers extra on presentation to content and consequently not precise plenty to discriminate disparate semantic blocks in a web page.

3. Vision-based Page Segmentation (VIPS)

People think a web page across a web browser and become a 2-D presentation that provides countless discernible cues to aid discriminate disparate portions of the page, such as lines, blanks, pictures, colors, etc . For the sake of facile browsing and understanding, a closely packed block inside the web page is far probable concerning a solitary semantic.

We have beforehand counseled a vision-based page segmentation method shouted VIPS. Comparable to semantic methods, the blocks obtained by VIPS are established on the semantic construction of web pages. Instituted semantic methods are obtained established on content scrutiny that is extremely sluggish, tough and inaccurate. VIPS discards content scrutiny and produce blocks established on the discernible cues of web pages. This method simulates how a user understands web layout construction established on his or her discernible perception. The DOM construction and discernible data are utilized iteratively for discernible block extraction, discernible separator detection and content construction. In the end a vision-based content construction can be extracted. As the method is totally top-down and the permitted degree of coherence can be pre-defined, the finished page segmentation procedure is effectual, flexible and extra precise from semantic perspective.

4. Hybrid Approaches

Although VIPS can differentiate multiple topics in web pages, it does not take under consideration the document length normalization problem. The distribution of block length is very diverse. Thus the varying length problem still exists even when we perform retrieval on block level. As fixed-length windows show great consistence on dealing with the varying length problem, Therefore Hybrid Page Segmentation approach tries to take advantage of both visual information and fixed length.

II PROBLEM FORMULATION

Previously Vision established segmentation algorithms such as VIPS (Vision-based Page Segmentation) algorithm exists to remove the semantic construction from web pages. These semantic constructions are hierarchical constructions, these hierarchical constructions embody corresponds to a block in the web page. In VIPS every single node is allocated a Degree to indicate discernible understanding of the block. The Vision-based Page Segmentation algorithm makes use of page layout feature though VIPS ignores the underlying content as segments can be composed of un-important data such as web ads, to resolve these subjects we counseled a n-gram established web page segmentation algorithm. That utilized the n-grams for segmenting the webpage lacking relying on the DOM tree for the segmentation process.

2.1 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.

2.2 Proposed work

2.2.1 Introduction of DOM-Tree

DOM (Document Object Model) is the request plan interface (API) for HTML and XML document. Employing the Document Object Model, programmers can craft documents, add, adjust or delete agents and content. DOM is a set of objects and admission, interface dealing the document object. In the DOM, documents have a logical construction that is extremely far like a tree. Every single document encompasses zero or one doctype nodes, one origin agent node, and zero or extra comments or processing instructions; the origin agent serves as the origin of the agent tree for the document [2]. HTML document contain the label, head, paragraph, hyperlinks and supplementary assorted components. DOM parses the HTML file and generates the inner tree construction of the file. DOM-based page segmentation is usually established on the predefined syntactic construction, that is, HTML tags. HTML tags are dependent.

2.2.2 Structure of DOM-Tree

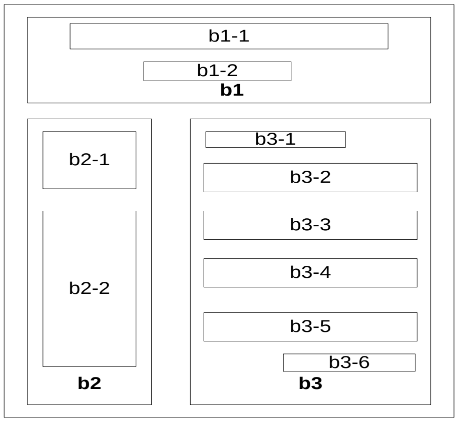


Fig. 2.2 The Presentation Structure

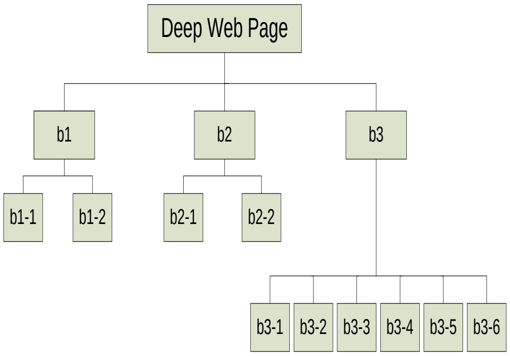


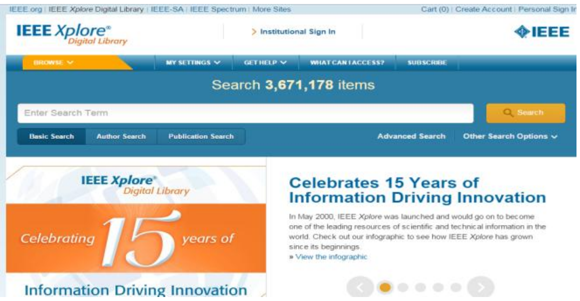
Fig. 2.3 Visual Block Tree

2.3 Algorithm: 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  on basis of ngrams [1]
6.  Group equivalent Regions  by merging based on formation of tag 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
11. End
```

III RESULTS AND ANALYSIS



Celebrates 15 Years of Information Driving Innovation

In May 2000, IEEE Xplore was launched and would go on to become one of the leading resources of scientific and technical information in the world. Check out our infographic to see how IEEE Xplore has grown since its beginnings.

» View the infographic

Fig. 3.1 IEEE Journal Site Taken for vision based page segmentation

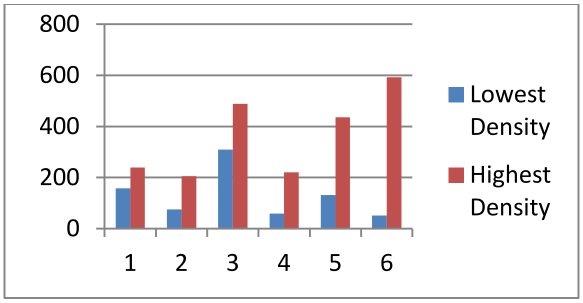


Fig. 3.2 Lowest and Highest Density of Text Segments

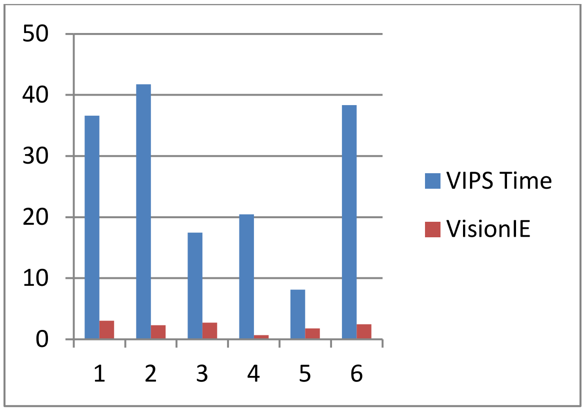


Fig. 3.3 Execution time of VIPS vs. VisionIE

Figure 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.

IV CONCLUSION

Previously Vision established segmentation algorithms such as VIPS (Vision-based Page Segmentation) algorithm exists to remove the semantic construction from web pages. These semantic constructions are hierarchical constructions, these hierarchical constructions embody corresponds to a block in the web page. In VIPS every single node is allocated a Degree to indicate discernible understanding of the block. Though VIPS ignores the underlying content as segments can be composed of un-important data such as web ads, to resolve these subjects, we established web page segmentation algorithm which is VisionIE. That utilized the n-grams for segmenting the webpage lacking relying on the DOM tree for the segmentation process. VisionIE 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. It then Select all the remaining Nodes and Segment Nodes into density regions on basis of ngrams.

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.

ACKNOWLEDGMENT

First of all I would like to thank my supervisors whose continuous support and encouragement have kept me going over the last few years as a Research scholar. My Guide is undoubtedly the person who has the strongest influence on this work. As an advisor, he taught me practices and skills that I will use in my academic career. He helped me to develop my own views and opinions about the research I undertook.

REFERENCES

[1] Kumud, Kulvinder Singh, and Vipul Jaglan. "Visual Webpage Content Segmentation and Retrieval Based on n-Grams." International Journal for Innovative Research in Science and Technology 2, no. 4 (42-49), 2015

[2] 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

[3] Yates, Alexander. "Information extraction from the web: Techniques and applications." PhD diss., University of Washington, 2007.

[4] Cai, Deng, Shipeng Yu, Ji-Rong Wen, and Wei-Ying Ma. "Block-based web search." In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 456-463. ACM, 2004.

[5] Hoffmann, Ralf, and Kirstin Krauss. "A critical evaluation of literature on visual aesthetics for the web." In Proceedings of the 2004 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries, pp. 205-209. South African Institute for Computer Scientists and Information Technologists, 2004.