Abstract

World Wide Web as a global service that distributed widely and is a global information service center for news, advertising, consumer information, financial management, education, government, e-commerce and many other intelligence services, so web pages include typically value a lot of worthless content, such as advertising, search, filtering panels, header, footer, navigation links, copyright information, etc. Useful content is original content from the web page that gives to users the information they need. More Recipient and end users search just useful content and need to extract the useful content from Web pages. To extract the useful content of the web page as correct, useful content and worthless content of the web page should be known clearly.

In this paper explained automated methods to extract useful content from Web pages such as Web useful content mining through links, attributes, words and learning models.

Keywords: Web mining, content useful extraction, document object model, Web Cleaning, Web Learning Modeling
Introduction

NASET is a markup language used for develop website, that is simple and efficient. several collection of tags to display the content of web pages. A web browser apprehend these tags and display a web page that readable for a human easily. Developers who want to illustrate their visual content with plenty features, such as Javascript, use a hierarchy called DOM tree. Figure 1 shows three different views from a simple web page [Erdağ-2013].

A human user when reading a web page nearly naturally do some kind of useful information extraction by ignore unhelpful contents[Chaw-2014] [Madhura-2013]. Figure 2, show a simple example using the block tags (such as tags DIV) in web pages. For example, it can determine the useful blocks as those with the least number of links and the unhelpful blocks as those are have a high number of links. Certainly the link and word distributions for each block are two important block are two important selection criteria, but some blocks may contain both useful and unhelpful blocks. When such blocks are encountered, the most specific bock under the parent block should be tested whether it is useful or not [Erdağ-2013].

---

[^1]: Document Object Model
Figure 1. Display of a simple web page in tree view [Erdinç-2013]

Figure 2. an Example of a web page blocks extraction [Erdinç-2013]
In figure 2, the block <DIV id = content> contains both useful and unhelpful blocks. To eliminate the content block extraction problem, it segment blocks into four different states. The four states in figure 2 are as follows [Erdinç-2013]:

unhelpful blocks Extraction:
  o <DIV id = dynamic_box_left>
  o <DIV class = article_tools>
  o <DIV id = fontsize>

the most useful block Extraction:
  o <DIV id = article_body>

the title Extraction:
The H1 tag inside the <DIV id = content-holder class = content -holder class =content _holder > block.

text summary Extraction:
  o <DIV id = article_date>

Although the web useful content extraction is a relatively intuitive for the user, but may be difficult to determine the content of a document automatically. There are several ways to handle this problem under different conditions. [Chaw-2014].

in the second part of This article define useful content extraction, standardization and of web page, DOM tree. the third section is provided conducted techniques to extract web useful content based DOM.

1. web page useful content Extraction

useful Content Extraction is the process to characterize the parts of a web page that contain the main textual content of this document. To identify useful content of unhelpful content in a web page, need to performs some early work including the Web page tags standardization and Web Page segmentation into regions and Semantic blocks [Chaw-2014] [Ajmera-2015].

2.1 Standardization of Web page tags

World Wide Web Consortium (W3C), the standard markup languages like XML have been proposed to reduce errors that lead to the analysis of various Web pages. So the structure worthless tags to process them, should be standardized [shobit-2013]. Cases where is performed in standardization a web page listed below [Gunasundari-2012] [Pranjali-2014].
symbols of "<" and ">" should contain only HTML tags. when used Other spaces should be by "LT &" and & GT "" to be replaced.

All tags must be same, each opening tag has a corresponding closing tag. attributes all tags should be placed inside citations symptoms. All tags should be placed the nesting properly.

2.2 Web page Segmentation into semantic areas

There are several approaches to web page semantic segmentation into districts and blocks. In classification based on DOM, introduced a HTML document as a DOM tree, and HTML document tags are presented form of blocks; Tags such as P , table , UL , H1-H6 and so on. DOM provides a useful general structure for a web page. Another way to segmentation web page, segmentation based on the design of the page. a web page in general separate into five zones: up, down, left, right, center[Chaw-2014]. Another way of classifications, segmentation According to the Vision (VIPS). VIPS² can hold related content together so that different semantic blocks separate from the each other. Visual signs such as font, color, size, etc. is used to identify blocks. Each block is shown in VIPS as a node in a tree [Chaw-2014].

1.3 DOM Tree

DOM model offers a standard for how to access, modify, add or remove HTML elements stored in XML or HTML document [Shobhit-2013]. DOM offers an HTML document as a tree structure so that Commonly called the DOM tree [Chaw-2014][ Yao Zhao-2014]. each DOM tree Related a webpage that tags internal nodes its and texts, images or links are leaf nodes. In general, each node in the DOM tree can provide a visual block. All tags in a web page constituting a tree structure. The nodes that do not contain any text as well as tags that are not associated with content such as <meta> <marquee> <style> <form> <script>, etc. removed then create a DOM tree [Chaw-2014].

2. Techniques

Most methods for the analysis of complex visual Web page, is based on learning techniques on analysis of the text density add link density. A model for identifying standard text using the shallow text attributes introduced [Kohlschütter-2010]. Moreover tools and software such as

² Vision-Based Page Segmentation
InstaPaper and Readability library\(^3\) used to extract and store the original main content of the webpage [Emilio-2014].

### 3.1 useful Content extraction using a hybrid model

Erdinç and Hardy [Erdinç-2013] presented a model that automatic law creation instead of manually law. The rules are to discover useful content from simple HTML pages, this model, extracts features based DOM and enjoying these features are used to extract useful content. This model is based on two tag, DIV and TD to determine the boundaries selected as useful signs; can determine rules comprehensive collections automatically. Figure 3 shows workflow of this method that includes the learning process, extraction process, choice of law and create a document be formed based on the low proportionality criterion on web pages. In general, this method involves two steps: In the first phase, induction law by using ML\(^4\) done and secondly, the extracted lows is used for the determination of useful content on web pages. In the process first checked the web page for whether the rules stored in the database, tested or not. If the lows are stored in the database then done test whether the laws are good for web page or not? Otherwise, created a ML method for laws to create a formed document it occurs when the law generates a single result. In the rule induction phase, created the DOM tree and extracted attributes from the DOM tree. The used ML method. At this stage compared several different machine learning methods And selected decision-tree learning and raising under the tree methods, as the most effective and most carefully for data collection. In this stage after extraction features to overcome the obscurity of nested, block tags are used to find the largest useful block, is often useful content and sub-blocks; As a result, R-WF-L- AL\(^5\) characteristic is derived from this method, which is an important feature in the ML. These features are very effective in extracting web content.

---

\(^3\) The reader is referred to the PHP port [www.keyvan.net/2010/08/php-readability](http://www.keyvan.net/2010/08/php-readability).

\(^4\) Machine learning

\(^5\) Ratio of word frequency in link to all words
3.2 Extraction useful content Using to EIFCE\(^6\) algorithm

Chow and Mim [Chaw-2014] provides the method that after receiving a web page by using DOM tree, returns useful block content to the output. Figure 4 shows the structure of the system that includes five steps. These five steps are carried out in two phases: the first phase web page segmentation and the second phase useful content extraction. EIFCE algorithm is presented to extract web useful content. This algorithm first applying the main block extractor algorithm (EVBE Algorithm) To segmentation Website into Semantic blocks. EVBE algorithm extracting web page semantic blocks using DOM tree of Web page and visual properties of each DOM node. EVBE algorithm receives DOM tree and create the block tree, then the main parameters of each block that Contains total amount of the specifications of each text, links, etc. It returns as output. The parameters are effective to find useful blocks. After that the EIFCE algorithm, applying to the algorithm of the block selection function for select a block. Then, applies the useful block find method for determining the useful content block of the selected blocks. Finally, returns useful content block of the web page as a result.

\(^6\) Effective Informative Content Extractor
2.3 useful content Extraction based links

Gunasundari and Karthikeyan [Gunasundari -2012] devised a method that after cleaning the webpage and create a DOM tree, can explore the web page content based on the number of punctuation and the ratio of non-link characters to characters that include links. The contents worthless are removed and useful content blocks are extracted from the web page. The content is diagnosed using four parameters: By comparing these four parameters, text link density (LTD), amount link (LA), link amount density (LAD) and node text length (NTL) With the respective neighboring nodes on the same level, nodes that contain the original content are selected for the next step. This action causes optimized nodes that include content are selected. LTD and NTL important parameters to determine the content and LA and LAD are an indicator to determine the exact content. Considering experimental results and statistical, Parameters values that be consistent with LTD=0.5  LA=10  NTL=100  LAD=0.5, selected to identify main content node. If the
nodes are not in accordance with the amount, text of under this node is unclear. For this purpose, the content must be under main nodes. If optimally nodes are coordinated with LTD=0.1 LA=10 NTL=50 LAD=0.1. Content is extracted by tools such as HTML parser. If the node is not harmonious with the situation, goes back to the previous step to find the optimal nodes of the next level nodes (child nodes of node).

In this step, Only selected nodes that are likely to be useful. But if the web page structure be decentralized relatively, extracted a section or a paragraph of text. Then matched with above results. If the size of text is greater than neighbor nodes that have visited and conditions to extract strict content, will be extraction all text from eligible nodes of the same level.

2.4 useful content Extraction from web news pages

Guo & Young [Guo-2010] presented a simple and efficient model called ECON Which uses the DOM tree and significant features, extracts web page content news automatically. first Econ finds a piece node of a portion of news content is categorized then finds the roll-backs from node piece to node Summary and news content total is categorized by the node summery. ECON During this process, removing the noise. There is An apparel exploration but important and intelligent to useful content identify. This is a web useful content including punctuation marks greeter than of the contents rest on the same page news. This invention can distinguish between news content and noise.

2.5 useful Content extraction using the class attribute

Shobhit and Aaron [Shobhit-2013] a method is provided that after the web page tags standardization and classification based on tree DOM, using class feature ,HTML elements within the 'BODY' of the document. Class attribute used to create different classes of an element that each class can have its own characteristics. The model first be found intended class then be found all children associated with the class node. the children represents the useful content block. Figure 5 shows workflow of this method. This extraction method is only for document. for image extraction is not useful, required graphical representation.
Extracting useful content using filtering noise

Neha & Saba [Neha-2011] A proposed automatic method for filtering out noise and showing web content useful. Figure 6 shows workflow of this way. In this method, after standardization and create the DOM tree, performing Extract the nested contents and content separators using intelligent technology to identify of users interest content. To extract the nesting contents of the tree DOM, implemented DFS\(^7\) techniques as a recursive. implementation return of DFS technique be checked to any DOM node content. DFS technique can be performed in two stages: The first step, if necessary, review and edit tag nodes, but not deleted. Second step, the tag nodes refer to unnecessary links. DOM nodes Related to advertisements removed by the parser filter. As a result will be preserved content nodes that are the user Intended; To be known as the useful content area of the web page. As soon as identified the main content area, it is necessary to separate useful content from other content. To do this, is required the separators. The task of separates among

\(^7\) depth first search
within the nesting contents is very difficult, To run it, used three intelligent technology: PR\textsuperscript{8} [Laender-2002], SHD\textsuperscript{9} [Laender-2002] [Jon-1998] and STH\textsuperscript{10} [Laender-2002]

![Diagram of content extraction process]

Figure 6: useful content extraction models Structure using noise filter [Neha-2011]

\textbf{8.3 useful content extraction through the leaves and the word}

This method [Aanshi-2013] is The combination of proportion the word to leaves (WLR) [D.Insa-2013] with link Features [Shen-2013] from the node to content extraction. In previous methods, characters were used instead of words but give useless important to words, so instead of characters, are used words. In this respect checked just nodes as nodes that include contextual content. So by adding text links and text anchor for the word to leaf ratios, can be extracted web content efficiently.

\textbf{8.4 useful Content extraction using learning model}

---

\textsuperscript{8} PR: Pattern Repeating
\textsuperscript{9} SHD: Standard Deviation
\textsuperscript{10} STH: Sibling Tag Heuristics
The method [Nethra-2014] A combined approach is proposed to extract the useful content of web pages. Figure 7 shows workflow of this way. A web page HTML be converted DOM tree are then extracted features. The extracted features to produce rules. the rules Using machine learning techniques such as decision tree classification and Naive Bayes classification, are produced. using this rules, will be deleted part of the unhelpful contents of the web page and extracted useful content from Web page.

![Figure 7: structure of web useful content extraction system using learning model [Nethra - 2014]](image)

### 3.8 Useful content extraction by Noise Reduction

In this way [Shalaka-2015] After standardization of the web page and create a tree DOM, web page is divided into two categories dataset. Then using a growth algorithm , remove the unhelpful blocks and identify similar blocks, extract useful blocks across a web page. This method can extract any number useful content of web pages that belong to different domains of websites. Figure 8 shows workflow this way.
Figure 8: workflow of useful content extraction, the Noise Reduction model [Shalaka-2015]

Conclusion

In this article have been collected web useful content extraction based the DOM tree and techniques. Web useful content extraction is the main content of the web page that is more intended the user. To extract Web useful content should be clear useful content of unhelpful content. There are techniques to address the issue which then standardization and classification of web pages on the tree DOM, are diagnosed useful content from unhelpful content and extracted Useful contents. Ways have been studied to extract web useful content: useful Content extraction using a hybrid model, useful content Extraction Using to EIFCE algorithm, useful content Extraction based links, useful content Extraction from web news pages, useful Content extraction using the class attribute, Extracting useful content using noise filtering, Extracting useful content through the leaves and the word, useful Content extraction using learning model and Useful content extraction by Noise Reduction.

References
1. Erdinç Uzun & Hayri Volkan Agun & Tarık Yerlikaya “A hybrid approach for extracting informative content from web pages” Information Processing and Management 49 (2013) 928–944


8. Yao Zhao, Yuanxin Zhao, Rongheng Lin, Hua Zou “Mining Service Tags with Enriched Information from the Internet” 2014 IEEE 10th World Congress on Services


18. K Nethra and J Anitha “web content extraction using hybrid approach” ictact journal on soft computing, january 2014, volume: 04, issue: 02