International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2013): 4.438

A Review on Identifying the Main Content From Web Pages

Madhura R. Kaddu[1], Dr. R. B. Kulkarni [2]

[1, 2] Department of Computer Scienece and Engineering, Walchand Institute of Technology, Solapur University, Solapur,India

Abstract: ***A web page is a web document in which huge amount of information is available and because of rapid growth of World Wide Web there is a great advantage to anyone, the user can easily access the web pages from any place through the internet. In the web page contains noisy information like menus, footers, unnecessary links, logos, etc and the main content. Most of the users are interested in only main content .But the main problem with the extraction process is to greater performance impact on web summarization; question answering system, information retrieval application because of the web page is collection of noisy and main content. So we propose an extraction process for identifying main content from web pages. In the extraction process consist of an automatic extraction techniques and hand crafted rules. In the automatic extraction techniques process the first step is to the web page is segmented into web page block and the second step is to differentiate main content from irrelevant or noisy content. In the hand crafted rule process extracts the main content from web pages by using rules which are already generated.***

Keywords: DOM Tree, content extraction, Web mining, machine learning method,Web page Segmentation.

1. Introduction

In the recent years there is tremendous growth of the World Wide Web for navigation into the web pages by numerous users through the web browser. The web page is the greatest source of information which is useful for the end user to search and many more reasons. The web developer creates web pages which are collection of informative content blocks such as relevant data to the web page and other blocks such as navigational bars, menus, footers, copyright notices, advertisements, category information, etc. which is called uninformative block. These blocks are not related to the main block, so it is the greatest challenge for extracting the main content from web pages. When an end user browsing particular web page, then end user most of the time focuses on relevant content and ignores noisy content. When examining web pages, human can easily distinguish informative content from the other uninformative content, but computer software is not more intelligent than human so it can't identify the main content from the noisy content. So it's become the challenge for web mining, data mining application and other application which contain web document as a data source.

The every web page on the browser you will see as an HTML file. In HTML file consist of HTML tags and content between these tags that displays on the web browser as a web page. To eliminate noisy data from web pages use a regular expression pattern in Uzun et al [1], but using this pattern is not a reliable extraction method from web domains. In the extraction process the content is extracted by using Document Object Model (DOM) method. For DOM methods the content is extracted by using web page segmentation. The web page segmentation is a task which divides the structure into smaller segment. Different methods were used for content extraction from web pages this we will see in the section 2. In fig 1 violet color block shows non-informative content and red color block shows informative content.

The rest of this paper is organized as follows. In section 2 includes related work. In section 3 discusses the proposed system. The section 4 gives the conclusion. The proposed system as extraction process increase efficiency, accuracy for extraction of the informative content from web pages.

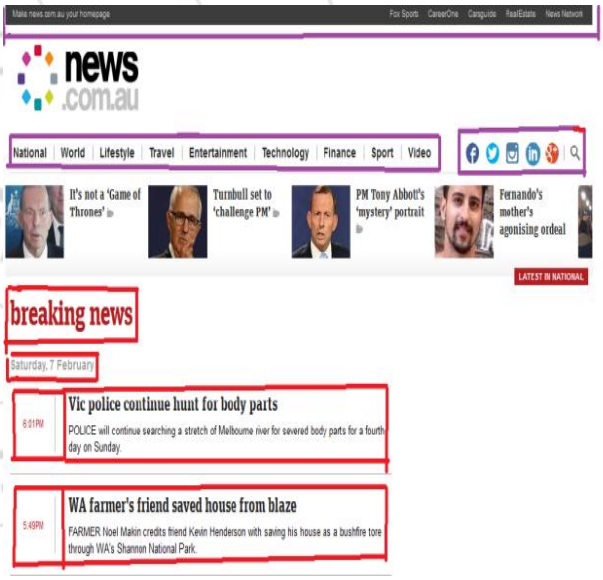

Figure 1: Typical example of web page with differentiate the main and noisy content

2. Related Work

A crawler was developed in which obtain news between 1998 and 2008 of three Turkish daily popular newspapers by Uzun,E. in [1]. After completion of the crawl, shows that the use of redundant or noisy contents such as advertisements, banners, navigation panels, logos and comments on web pages has been increased day by day. So for relevant content require removing this noisy content from web pages, for this purpose they searched a regular expression patterns, but this pattern is not a reliable extraction technique for different web domains. Because of variations in HTML tag naming and

hierarchy it is very difficult to prepare regular expression patterns for extracting the main content from web pages.

A small set of shallow text features was analyzed by Kohlschutter C in [2] for classifying the individual text elements in a Web page from different sources. The web page consists of footer, navigational bars, logos and advertisements such elements called as boilerplate text which is not related to the main content. In this paper they analyze most popular features used for boilerplate text detection. These features are the counting number of word, measuring density of link and density of text. The number of words means words are white-space delimited character sequences which at least contain one letter or digit. The link density means ratio of the number of tokens within an A tag divided by the total number of tokens in the block. The text density means it counts the number of tokens in a particular text block b divided by the number of lines covered after word-wrapping the text at a fixed column width.

An approach was developed to correctly identify the content portion of web-pages by Gibson J., Wellner B., & Lubar S in [3]. Most approaches uses hand-crafted tool for content extraction, but it require more time and efforts to implement so it is very tedious job. So they describe a sequence labeling problem to identify the relevant content from web pages. Firstly divides the web page into a sequence of blocks by using the Boundary Detection Method. Then each block is gets labeled as Content or NotContent. For this purpose they employed three different statistical sequence labeling models to learn how to label sequences of blocks so as to identify the content. The 3 statistical sequence labeling models are Conditional Random Fields (CRF), Maximum Entropy Classifiers (MaxEnt), and Maximum Entropy Markov Models (MEMM). From this Conditional Random Fields (CRF) can perfectly identify the content portions of web pages.

A novel approach was proposed by Ma L. [4] in that identifies web page templates and extracts the unstructured data. To extract only the body of the page and eliminating the template increases the retrieval precision also by removing the noisy data such as unnecessary links, etc improves the searching capability by reducing the irrelevant results. In this approach web page is divides into smaller units by using <TR>, <TD> tags. Then the "IMatch" duplicate detection algorithm used for detection of duplicate web pages in the collection. After duplicate are removed the deliminator tag or its respective end tag such as table tag or image map tag is encountered, then we store the identified text chunk in the text chunk map and also store the calculated document frequency for that chunk for further use. Their threshold value is calculated, and then the remaining text chunks are the extracted texts that are passed to the indexer to be indexed and remove the template.

There are several approaches of automatic extraction techniques for web pages segmentation. The automatic extraction technique approaches are location -based segmentation, vision-based segmentation, and segmentation based on DOM. The string manipulation functions are written in the hand crafted rule process which was previously used for extracting the content from the web pages. But for generating the hand crafted rules require much time also it is critical job for extraction because of creating rules require to knowledge about how the web page is designed, so mostly focus on automatic extraction technique for extracting the content from web pages. The above approaches are given below.

2.1 Location -Based Segmentation

The segmentation based on location considers the "location" as a feature for example left menu is mostly link section which is on left side, in the footer section contains author information which is placed at the bottom side, the title of the web page which is on top side so such type of location are used . In Kovacevic, M. [10] an M-Tree was constructed for extracting the visual information from an HTML source based on location. In the M-tree construction, initially by using HTML parser parse the given web page and it extracts two elements which are tags element and data element. To build the tree these two elements are given and by using predefined rules build the tree. After this by using the rendering module coordinates the objects for further use and for classification the Naive Bayes Classifier is used which is the simplest instance of a probabilistic classifier and it is defined as:

$$p(c|d) = \sum_i W_i * p_i(c|d)$$

Where the pattern ***d*** belongs to class ***c*** (posterior probability), weight W_i is i-th classifier.

2.2 Vision-Based Segmentation

For web page segmentation different features are used for example lines, different font, colors, shapes & size. In Cai, D. [6] the automatic top-down approach was developed to detect content structure from web pages. The visual blocks are created by dividing the web page and for each visual block Degree of Coherence (DoC) is measured based on properties for extraction. In this study proposed a VIPS algorithm for segmentation in which combining the visual cues and the DOM structure of the web page. In the vision-based page segmentation algorithm initial task is to construct the DOM structure as shown in fig 2 and visual information like color, font size, lines and position are obtained from a web browser. In block extraction process visual block extraction algorithm is used then the visual block extraction process is started to extract visual blocks of the current level from the DOM tree based on visual cues and. Then in the Visual separators detection process by assigning weights for separators based on some properties, Visual separators among these blocks are identified. Then in the Content Structure Construction process starts from the separators with the lowest weight and the blocks beside these separators are merged to form new virtual blocks, this process iterates till separators with maximum weights are met.

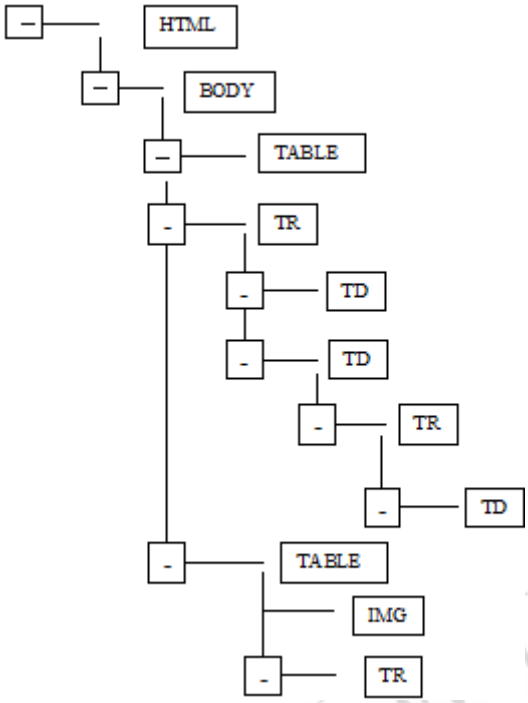

Figure 2: DOM Tree of typical web page

2.3 DOM-Based Segmentation

In DOM-based studies use DOM-level features to extract content from web pages by using trained classifiers, such as in [11] they consider the problem of automatically segmenting web pages in a principled manner. In this the weighted graph is defined where nodes are the DOM tree nodes and the edge-weights defines the cost of placing the end points in same or different segments and this was developed by Chakrabarti, D in [11]. By using this formulation produce two types, first one is based on correlation clustering and second type is based on energy-minimizing cuts in graphs. The energy minimizing formulation is better than the correlation clustering; the quality of segmentation depends on the edge weights of the graph. Then apply our segmentation algorithms in [11] as a pre-processing step to the duplicate webpage detection problem.

Another trained classifier is Function-based Object Model (FOM) that attempts to understand the authors' intention by identifying Object function instead of semantic understanding. The FOM model for website understanding is developed and these FOM models are Basic FOM based on the basic functional properties of Object and Specific FOM based on the category of Object by Chen, J. [12]. In the Basic FOM, Object can be classified into Basic Object (BO) and Composite Object (CO). The Basic objects (BO) have function, Property and a Composite Object (CO) is a set of Objects (BO or CO) that perform some certain functions together. The Specific FOM is represented by its category which reflects the authors' intention directly, it consists object like Information Object, Navigation Object. To automatically detect the functional properties and category of Object is presented for FOM generation; the general rules and specific rules based on FOM are combined for practical adaptation. An application example of the proposed model is a system for web content adaptation over Wireless Application Protocol (WAP).

The new approach is described to segment HTML pages and building on methods from Quantitative Linguistics and strategies borrowed from the area of Computer Vision by Kohlschutter, C. in [13]. The notion of text-density is used as a measure to identify the individual text segments of a web page also it reduce the problem to solving a 1D-partitioning task. They present the Block Fusion algorithm for identifying segments using the text density metric.

The approach is developed to correctly identify the content portion of web pages by Gibson, J. in [13]. For this purpose each block is gets labeled as Content or NotContent by 3 statistical sequence labeling models. The first model is *Conditional Random Fields (CRF)* defined as:

$$p(y|x) = \frac{1}{Z_x} \exp\left(\sum_{i=1}^{n} \sum_{k} \lambda_k f_k(y_i, y_{i-1}, x, i)\right)$$

Where Z_x is a normalization term overall possible label sequences, feature functions f_k are arbitrary functions over the current and previous labels y_i, y_{i-1} , the entire observation sequence ***x*** and the current position, *i*.

The second model is *Maximum Entropy Classifiers (MaxEnt)* defined as:

$$p(y = y_i|x) = \frac{\exp\left(\sum_k \lambda_k f_k(y_i, x)\right)}{\sum_{y_j} \exp\left(\sum_k \lambda_k, f_k(y_j, x)\right)}$$

Where *y* is a random variable over Classifications such as (Content or NotContent) or the possible outcomes, *x* describes the observed data, the functions f_k are features over the observed data and a particular outcome, the λ_k is the weights.

The third model is *Maximum Entropy Markov Models (MEMM)* defined as:

$$p(y|x) = \sum_i p(y_i|y_{i-1}, \rho_x)$$

Where $p(y_i|y_{i-1}, x) = \frac{\exp\left(\sum \lambda_k f_k(y_i, y_{i-1}, x, i)\right)}{\sum \exp\left(\sum \lambda_k f_k(y_j, y_{i-1}, x, i)\right)}$

3. Proposed System

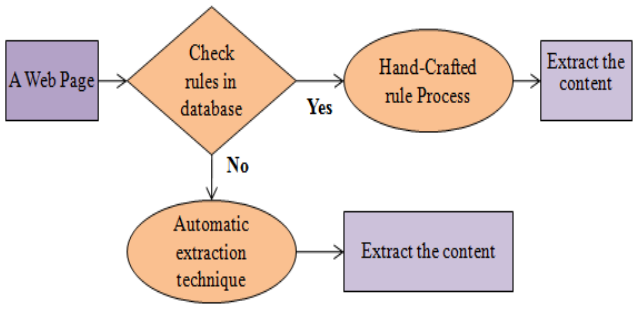

Figure 3: Architecture of proposed system

The architecture of the proposed system is shown in fig 3. The aim of this system is to extract the only main content from web pages and ignoring the noisy content. For this purpose by using extraction process, i.e. automatic extraction techniques and hand crafted rules the user get the appropriate result. In automatic extraction techniques the main content extracts from web pages by using dynamic rule generation. In hand crafted rules process the main content extracts from web pages by using rules which are already generated and stored into the database.

In this section gives step by step how the content is extracted from web pages.

1) Initially a user gives the URL of the web page which he/she wants to extract the content.
2) The rules for a given web page are checked into the database.
3) If the rules for a given web page are present into database, then by using that rules extract the relevant content so this step is called as a hand crafted rules process.
4) If the rules for a given web page are not present into the database, then it must be required to generate the rules for a given web page and stored the generated rules into database for further use. So this step is called as automatic extraction techniques because in this step dynamic rules are generated for a given web page.
5) In automatic extraction techniques process consist of following steps:-

A)DOM Tree Construction

The HTML web page can be defined by using a tree structure called as a DOM tree. The DOM (Document Object model) tree visualizes the web document into a tree hierarchy. The web page is a collection of tags, nested tags. So the DOM tree shows every HTML elements into a tree structure, by using the DOM tree we can traverse the whole document very easily and also missing tags can be easily identified. Every element of DOM tree is nothing but DOM node or block.

B)Feature Extraction

By using different features each node must be classified as informative or uninformative node. In [2] number of word, link density and text density features were used for relevant content detection. But in proposed system different features are used like the Word Frequency, Density in HTML, Link Frequency, Word Frequency within Links, Average Word Frequency in Links, Ratio of Word Frequency in Links to All Words these features are used for extracting the main content from web pages.

C)Apply ML(Machine Learning) Method

The different machine learning methods are used in extraction process, but the decision tree classification method gives best performance. A decision tree [7] has a root node and branch node used for decision making purpose. The decision tree learning algorithm starts from the root node, then split each node recursively based on C4.5 algorithm. The C4.5 algorithm [8] is applied as decision tree classification.

6.By using the rules which are generated in the above ML method extract the content from web pages.

4. Conclusion

The informative content is extracted from web pages and noisy content such as links, footer, header etc. are avoided. The main content is identified by using extraction process, in the extraction process consist of automatic extraction techniques and hand crafted rules. In automatic extraction techniques a web page is converted into a DOM tree and features are extracted. The extracted features are input to the machine learning method like decision tree classification method. The rules are generated and by using these rules main content is extracted. In hand crafted rules, the rules are into database by using these rules the relevant content is extracted from web pages. The proposed system produces effective rules and building more accurate model.

References

[1] Uzun,E.,Yerlikaya,T.,&Kurt,M, "Analyzing of the evolution of web pages by using a domain based web crawler". Engineering, Technologies and Systems – Techsys, 26 28May, pp.151–156. (2011a).
[2] Kohlschutter, C.,Fankhauser,P.,&Nejdl,W, "Boilerplate detection using shallow text features". In Proceedings of the third ACM international conference on Web search and data mining (WSDM'10),pp. 441–450 New York. (2010)
[3] Gibson, J.,Wellner,B.,&Lubar,S, "Adaptive web-page content identification". In WIDM '07: Proceedings of the 9th annual ACM international workshop on Web information and data management, pp. 105–112, New York, NY, USA, ACM. (2007)
[4] Ma, L.,Goharian,N.,Chowdhury,A.,&Chung,M, "Extracting unstructured data from template generated web documents". In CIKM '03 Proceedings of the twelfth international conference on Information and knowledge management, p. 512.ACM Press. (2003)
[5] Cai, D.,Yu,S.,Wen,J.-R.,&Ma,W.-Y, "VIPS: A vision based page segmentation algorithm". Microsoft technical report. MSR-TR-2003-79. (2003).
[6] Cai, D.,Yu,S.,Wen,J.-R.,Ma,W.-Y, "Extracting content structure for web pages based on visual representation". In X.Zhou.Y.Zhang, & M.E.Orlowska (Eds.), APWeb (Vol. 2642 of LNCS, pp.406–417).Springer. (2003)
[7] http://en.wikipedia.org/wiki/Decision_tree_learning
[8] http://en.wikipedia.org/wiki/C4.5_algorithm
[9] Baluja, S. "Browsing on small screens:Recasting web page segmentation into an efficient machine learning framework". In WWW '06: Proceedings of the 15th international conference on World Wide Web pp. 33–42, .New York: NY, USA, ACM. (2006)
[10] Kovacevic, M., Diligenti, M.,Gori, M.,&Milutinovic, V.," Recognition of common areas in a web page using visual information: A possible application in a page classification". In The proceedings of 2002 IEEE international conference on data mining (ICDM'02), Maebashi City, Japan, December.(2002)

[11] Chakrabarti, D., Kumar,R. ,& Punera,K, "A graph-theoretic approach to web page segmentation".In WWW 2008:Proceeding of the 17th international conference on World Wide Web (pp. 377–386). New York: NY, USA, ACM. (2008)
[12] Chen, J., Zhou,B., Shi,J.,Zhang ,H.-J., & Qiu,F., "Function-based object model towards web site adaptation". In The proceedings of the 10th World Wide Web conference (WWW10), Budapest,Hungary. (2001)
[13] Kohlschutter, C., & Nejdl,W.,"A densitometric approach to web page segmentation ".In ACM 17th conference on information and knowledge management (CIKM 2008).