International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-2 Issue-1, March 2013

Elimination of Noisy Information from Web Pages

Alpa K. Oza, Shailendra Mishra

Abstract- A Web page typically contains many information blocks. Besides, the content blocks, it usually has such blocks as navigation panels, copyright and privacy notices, and advertisements. These blocks that are not the main content blocks of the page, we call them as noisy blocks. We show that the information contained in these noisy blocks can seriously harm Web data mining. Thus eliminating these noises is of great importance. In our work we focus on identifying and removing local noises in web pages to improve the performance of mining. A simple idea for detection and removal of noises a new DOM tree structure is proposed. The result shows the remarkable increase in F score and accuracy is obtained.

Keywords- Noise elimination, DOM tree, Web page cleaning.

I. INTRODUCTION

The rapid expansion of the Internet has made the WWW a popular place for disseminating and collecting information. Data mining on the Web thus becomes an important task for discovering useful knowledge or information from the Web [6][9]. However, useful information on the Web is often accompanied by a large amount of noise such as banner advertisements, navigation bars, copyright notices, etc. Although such information items are functionally useful for human viewers and necessary for the Web site owners, they often hamper automated information gathering and Web data mining, e.g., Web page clustering, classification, information retrieval and information extraction. Web noises can be grouped into two categories according to their granularities:

Global noises: These are noises on the Web with large granularity, which are usually no smaller than individual pages. Global noises include mirror sites, legal/illegal duplicated Web pages, old versioned Web pages to be deleted, etc.

Local (intra-page) noises: These are noisy regions/items within a Web page. Local noises are usually incoherent with the main contents of the Web page. Such noises include banner advertisements, navigational guides, decoration pictures, etc.

In this work, we focus on detecting and eliminating local noises in Web pages to improve the performance of Web mining, e.g., Web page clustering and classification. This work is motivated by a practical application. A commercial company asked us to build a classifier for a number of products. They want to download product description and review pages from the Web and then use the classifier to classify the pages into different categories.

Revised Manuscript Received on 30 March 2013.
* Correspondence Author
Alpa K. Oza, Information Technology, Parul Institute of Engineering and Technology, Gujarat Technological University, Gujarat, India.
Shailendra Mishra, Computer Science Engineering, Parul Institute of Technology, Gujarat Technological University, Gujarat, India.

In this paper, we will show that local noises in Web pages can seriously harm the accuracy of data mining. Thus cleaning the Web pages before mining becomes critical for improving the data mining results. When examining a Web page, humans can easily distinguish the main content from navigational text, advertisements, related articles and other text portions. A number of approaches have been introduced to automate this distinction, using a combination of heuristic segmentation and features. Figure 1 gives a sample web page from BBC News with main contents, advertisements, navigation links etc.

Elimination of noisy and irrelevant contents from web pages has many applications, including web page classification, clustering, web featuring, proper indexing of search engines, efficient focused crawlers, cell phones and PDA browsing, speech rendering for the visually impaired, improving the quality of search results and text summarization. Thus cleaning web pages for web data extraction becomes crucial for improving the performance of information retrieval. We investigate to remove various noise patterns in web pages.

Figure 1: a sample web page from BBC News with main contents, advertisements, navigation links etc.

II. RELATED WORKS

The original idea of this work has been evolved while developing an innovative approach for effective optimal feature subset selection for web page categorization. The subsistence of local noise is an issue that accompanies the growing need to extract relevant blocks from a web page. Web content mining face huge problems due to the presence of the local noise. There have been a number of studies that analyze an HTML page visually in order to extract the target information from the pages and most of them have focused on detecting main content blocks in web pages but less work have been evolved on detecting and removing noisy information from a web page.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

A new tree structure, called Style Tree, is proposed in [1] to capture the actual contents and the common layouts of the pages in a web site. An information (or entropy) based measure is used to evaluate the importance of each element node in the style tree, which in turn helps to eliminate noises.

The most popular features used for boilerplate detection [3] on two corpora. They show that a combination of just two features - number of words and link density - leads to a simple classification model that achieves competitive accuracy. The features have a strong correspondence to stochastic text models introduced in the field of Quantitative Linguistics.

A new content extraction method is thus proposed [5], which can discover web page content according to the number of punctuations and the ratio of non-hyperlink character number to character number that hyperlinks contain. It can eliminate noise and extract main content blocks from web page effectively. In this paper [6] they propose a novel idea for finding near duplicates of an input web page, from a huge repository. This approach explores the semantic structure, content and context, of a web page rather than the content only approach. They present a three-stage algorithm which receives an input record and a threshold value and returns an optimal set of near duplicates. By introducing a new technique known as Minimum Weight Overlapping (MWO) based on the threshold value and finally we get an optimal number of near duplicate records.

To extract informative content block from Web documents by removing noise blocks. So, to extract information from these pages, several challenges must be overcome. Another existing informative content extraction system implemented nowadays depends on rule based systems where Web sites with various templates are not applicable. A possible application of Neural Networks [4] is presented for three pattern classification combine with DOM structure to extract content information. The type of Neural Network used to implement our system is feed forward which uses the back propagation learning algorithm. The proposed paper [2] intends to introduce an algorithm which could extract main content that is not necessarily the dominant content and without any learning phase, with one random page and by using visual cues to simulate user page visit and block the page based on it and gains higher precision. Next section demonstrates the proposed algorithm in this paper in two main phases, first block tree construction, and then it finds main block from the block nodes in the computed block tree.

III. PROPOSED WORK

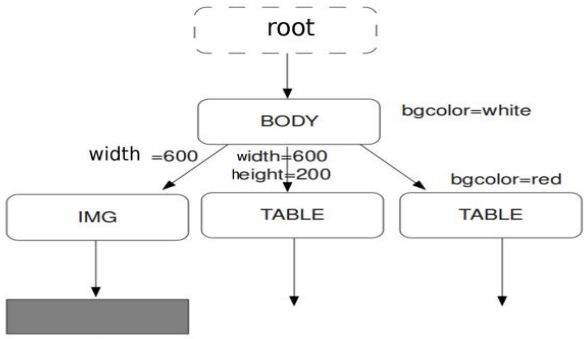
The Document Object Model (DOM) specification is an object-based interface developed by the World Wide Web Consortium (W3C) that builds an XML and HTML document as a tree structure in memory. An application accesses the XML data through the tree in memory, which is a replication of how the data is actually structured. The DOM also allows the user to dynamically traverse and update the XML document. It provides a model for the whole document, not just for a single HTML tag. The Document Object Model represents a document as a tree. DOM trees are highly transformable and can be easily used to reconstruct a complete webpage. DOM tree is a well defined HTML document model. Some HTML tags do not include a closing bracket. For some of these tags, the closing bracket is inferred by the following tag, for example <LI> tag is closed by the following </LI> tag.

In order to analyze a web page, we first check the syntax of HTML document because most HTML Web pages are not well-formed. And then we pass web pages through an HTML parser, which corrects the markup and creates a Document Object Model (DOM) tree. Figure 2, shows the example of a DOM tree of HTML web page. After creating the DOM tree, the system split it into multiple sub-trees according to threshold level. Different Web Sites have different layout and presentation style, therefore the depth of the tree of the Web page is varied according to their presentation style. The system must know the maximum level of DOM tree to choose the good choice of threshold level. Therefore, the system traverses the whole DOM tree to get the maximum depth of DOM. For the training data set, we picked the best suited threshold level up by setting various threshold levels. Then, the system chooses the suitable threshold level for test data set by using these known pair of series. The system estimates the nature of the relationship between the maximum level and threshold level based on linear regression analysis. A regression is a statistical analysis assessing the association between two variables. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. Once we obtained the threshold level, the system determine some nodes of DOM less than the threshold level as noise and remove them before classification process start. After splitting sub-trees, we transform them into numeric representation for input patterns of neural network classification model using eq.1.

$$Xi = Sn/Tn \qquad \text{eq.1}$$

```
<BODY bgcolor =WHITE
   <IMG src="picture.gif" height=200>
   <TABLE width=600 height=200>
   . . .
   </TABLE>
   <TABLE bgcolor=RED>
   . . .
   </TABLE>
</TABLE>
```

(a)



(b)

Figure 2: An example of DOM Tree of HTML web page.

Published By:
Blue Eyes Intelligence Engineering
and Sciences Publication (BEIESP)

Where, Sn is the number of occurrence of same leaf nodes in sub-tree, Tn be the total number of leaf nodes in sub-tree. By parsing a webpage into a DOM tree, more control can be achieved for the proposed system. Lastly, we remove the noise class in Web page and show extracted main content data in HTML page.

A. The Steps Of Content Extraction

Step 1 Standardizing the web page tags [10].
a. Symbols, ”<” and ”>”, should only contain html tags. When used in other place, they should be replaced by ”&lt” and ”&gt” respectively.
b. All tags must be matched, i.e. every starting tag has a corresponding ending tag.
c. Attributes of all tags must be encircled by quotation marks.
d. All tags must be nested correctly. For example, <a>. . .. . is a correct nest,
while <a>. . .. . is incorrect.

Step 2 Preprocessing the web page tags.
All tags on the page form a tree structure. Those nodes that do not contain any text should be removed, as well as invalid tags such as<script> <style> <form> <marquee> <meta> etc, which are unrelated to the content. Then the structure tree is built.

Step 3 Judging the location of content
The aim of this process is to select the optimum node containing content. If a node is not satisfied with this condition, the text under this node is not identified. As the news web page is a tree structure, the content must be under a general node.

Step 4 Extracting the content
The content is extracted by tools such as html parser. If the node is not satisfied with the conditions, return the step 3 in order to find the optimal nodes of the next level nodes (the child nodes of the node).

Step 5 Adjusting the extraction results from step 4
In step 3, only the node that most likely contains the content is selected. But if the structure of a web page is relatively decentralized, it is very prone to extract a section or a paragraph of the whole content. As the adjacent nodes on the same level are free of judge, in this step, we must adjust the above result. The text also should be extracted from the adjacent nodes that meet the conditions of the precise content extraction. So all text will be extracted from the qualified nodes on the same level.

IV. EXPERIMENATAL RESULTS

In this section we evaluate the performance of the proposed noise elimination algorithm. Since the purpose of our noise elimination is to improve web mining, we performed a web mining task, automatic web page classification, to evaluate our system. By comparing the classification results before and after cleaning, we show that the proposed technique is better enough to improve the classification results. The methodology followed here consisted of selecting a random set of web pages from selected categories to form different data sets, determining a set of features to represent each data set, preparing a pair of datasets before cleaning and after cleaning.

We show that how the noise misleads data mining algorithms to produce poor results. We use the popular F score measure to evaluate the results before and after cleaning. We also include the accuracy of results for classification. F score measures the performance of a system on a particular class, and it reflects the average effect of both precision and recall during automatic web page classification.

$$precision = \frac{categories\ found\ and\ correct}{total\ categories\ found}$$

$$recall = \frac{categories\ found\ and\ correct}{total\ categories\ correct}$$

$$F\ score = \frac{2 * precision * recall}{precision + recall}$$

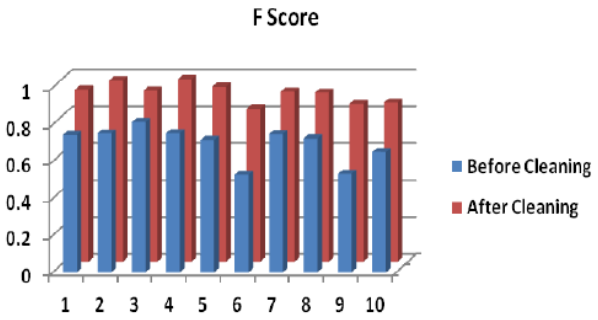


Figure 3: F Score before and after cleaning

V. CONCLUSION AND FUTURE WORK

This paper proposed a novel task for finding local noise in web pages. The proposed technique aims at helping document classification in web content mining based on a new tree structure, featured DOM tree. We could detect and remove local noises with an increased relevancy. We evaluate the performance of our algorithm in terms of F score and accuracy of web page classification and we could achieve an improved result with a large margin than before cleaning. A lot of work needs to be done before the use of search engine interface, such as concept retrieval and the expansion of semantic and synonyms, which are necessary for further investigations.

REFERENCE

1. Lan Yi, Bing Liu and Xiaoli Li, Eliminating Noisy Infromation in Web Pages for Data Mining. ACM- 2003
2. M. Asfia, M. M. Pedram and A. M. Rahmani, Main Content Extraction from Detailed Web Pages. International Journal of Computer Applications- 2010.
3. C. Kohlschutter, P. Fankhauser and W. Nejdl, Boilerplate Detection using Shallow Text Features. WSDM- 2010.
4. T. Htwe, N. S. M. Khan, Extracting Data Region in Web Page by Removing Noise using DOM and Neural Network. ICIFE- 2011.
5. R. Gunaundari and Dr. S. Karthikeyan. A Study of Content Extraction from Web Pages Based on Links. IJDKP-2012.
6. S. N. Das, M. Mathew and P. K. Vijayaraghavan. An Efficient Approach for Finding Near Duplicate Web Pages using Mininum Weight Overlapping Method. IJECE-2011.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)