Cleaning Web Pages for Relevant Text Extraction and Text Categorization

Amit Chauhan, Himanshu Uniyal, Dr. Bhasker Pant  
Department of IT,  
Graphic Era University,  
Dehradun, India

Abstract

Noise on the web pages leads to trouble mining the main content of web. Web pages typically contain a large amount of information like advertisements, navigation bar and copyright notices which are not part of the main information of pages called noise. Typically we apply data mining techniques such as classification and clustering after cleaning noise from the web pages. Eliminating noise blocks from the web pages will improve the accuracy and efficiency of web content mining.

This paper proposes a method “web information extraction and classification framework”. In this method we remove noisy data from the web pages and then retrieve the most relevant information from the same web page and finally the classification is performed by using Support Vector Machine which gives accurate and efficient result.

Keywords- VIPS, Web Page Cleaning, Support Vector Machine, Document Classification.

1. Introduction

Web mining is a multidisciplinary field including data mining, machine learning, natural language processing, statistics, databases, information retrieval, multimedia etc. Web mining consists of: Web usage mining, Web structure mining and Web content mining.

“Mining, extraction and integration of useful data, information and knowledge from web page content is called web content mining.” [1]

Nowadays WWW is growing rapidly and also containing high amount of data which is often accompanied by a large amount of noise. Now there are several billions of HTML files, images and other multimedia files available on Internet and they are still growing at the rapid rate. Web Content Mining uses the ideas and principles of data mining and knowledge discovery to screen more specific data.

The use of the Web as a provider of information is unfortunately more complex than working with unchangeable databases. Because the web is dynamic in nature and it contains large number of data, there is a need for new approaches that are not depending on accessing the complete data on the outset. So focusing on that the information retrieval is the major issue and interest for lots of researchers. From the content of web pages, mining of web contents aims to extort constructive information or knowledge.

The main problem encountered while mining the web are the noises. The html web page contains lots of data, but the information that the user wants is called main information (main content on the web page) and the rest are noise which includes advertisements, navigation bar, copyright notices, comments and reviews. These blocks of information (noise) are essential ingredients of web pages because of the user friendly access and also for the commercial purposes.

Existing work on web content mining [3, 4, 5, and 6] states that, after eliminating noise blocks from web pages and applying data mining techniques like clustering and classification will improve the accuracy and efficiency of mining result. Thereby having an algorithm to extracts only main content could help better quality on web page indexing. Almost all algorithms have been proposed are tag dependent means they could only look for primary content among specific tags such as <TABLE > or <DIV >.
In our paper, we first describe related studies in Section 2. Then, in Section 3, we illustrate the representation of noisy data in a page and present our proposed system “WEB INFORMATION RETRIEVAL & CLASSIFICATION FRAMEWORK”, to detect, eliminate multiple noise patterns for extracting main content information and classify these extracted main content information into predefined classes. Section 4 describes the experiments of our approaches and finally, the paper shows the conclusion and future work in Section 5.

2. Review of Related Researches

The majority approach focus on removing noises from a web page for efficient web mining. Many researchers have undergone infinite researches for removing noisy data from the web pages. Among them, only few have been successful researches that have been able to perform effective web page cleaning. Tag tree or Document Object Mode (DOM) provides each web page a tree structure, show the content and also provides the presentation of the page. In [13] Lin & Ho, proposed a method, to discover informative content blocks from web documents. According to HTML tag, <TABLE>, a web page is parsed into separate blocks.

A limitation of this work is that it is restricted to tabular (with <TABLE> tags) web pages. A style tree structure is proposed in [4] to capture the layout and contents of pages in an example web site. In [5] Bar-Yossef Z. et al. proposed a template detection algorithm based on the number of links an HTML element has and it partitions each web page into several page-lets, which are units with well defined topic or functionality. Templates are then detected by identifying duplicate page-lets, but it does not work well when detecting near duplicate page-lets. A Vision-based Page Segmentation (VIPS) algorithm is proposed in [2], that segments web pages using DOM tree with a combination of human visual cues, including tag cue, color cue, size cue, and others. The VIPS algorithm has been applied to information retrieval, information extraction, and learning block importance on a single html web page [14]. In order to deal with Web page noise and to lift up web mining, a feature weighting system has been proposed by Lan Yi et al. [15]. For capturing the general structure and comparable blocks in a group of Web pages, the technique has introduced a compacted structure tree. Then a measure which was based on information was applied to assess the importance of every node in the compacted structure tree. On the structure of the tree and its nodal significance values, a weight is assigned to each word characteristic in its content block. The resulting weights were applied in Web mining. Using two Web mining techniques, which are Web page clustering and Web page classification, the given technique was assessed. Thus the research concluded that weighting methods was highly successful in developing the mining outcomes. Extracting Article Text from the Web with Maximum Subsequence Segmentation on www is done by Jeff Pasternak [11].

In [3], C. Li et al. proposed a method to extract informative block from a web page based on the analysis of both the layouts and the semantic information of the web pages. They need to identify blocks occurring in a web collection based on the Vision-based Page Segmentation algorithm. In [7] Swe Nyein proposed a system based on the CST tree generated by the DOM tree and also could extract the relevant documents from the web pages using cosine similarity measure.

After retrieving the web page’s main information, this main information is further processed to data mining technique like Classification. In [10] Thorsten Joachims et al proposed an approach “Text categorization with support vector machines: learning with many relevant features” which was done to categorize the information into predefined categories.

3. Methods and Approach

In this section we are describing our proposed method “Web Information Extraction and Classification Framework”. Firstly we take 750 web pages each for three different categories namely cancer, tuberculosis and Asthma as our input data.

Our framework is divided into two modules, module 1 which deals with web page cleaning and extraction of main text which is pure text and is input for module 2 in which classification are performed on the pure text as classification techniques cannot be directly applied on the web pages. The output resulting from module 2 gives the model for predicting the categories of the web pages. For example, we have taken a web page which may or may not belong to the above mentioned categories. This experimental web page is processed through our framework to create the model which is matched with our trained model which shows that the experimental web page belongs to which predefined category.
Module 1: Web Page Cleaning and Main Text Extraction

An Introduction to VIPS Algorithm:

In this paper, we use VIPS (Vision-based Page Segmentation) algorithm to extract the content structure for a web page. The algorithm makes full use of page layout features and then tries to segment the web page at the semantic level (e.g. figure 2). Each node in the extracted Dom tree will correspond to a block of coherent content in the original web page [2].
Removing Noise and Relevant Text Extraction

Step 1: Block Extraction

In this step we first segments each web page into different blocks, and then extracts the contents and other features of blocks and save them in database records.

Step 2: Block Importance Retrieval

In Block Importance Retrieval step we computes each block importance which is based on the similarity of the block contents, for example its position in the web page, and the percentage of contained link texts.

Step 3: Cleaned Text Files Generation

This is our final step of retrieving the main content of web page. This is done by grouping the block records with high importance degrees in each web page and save these files into database for web content mining. The above three steps are taken from paper [8].
An Example of retrieved Article from input web page mentioned in figure2:

The scientists, funded by Cancer Research UK, found that women with the fault, and who had the estrogen receptor-positive form of breast cancer, were more likely to die from the disease - 62 out of 100 women who carry the fault will be alive ten years after being diagnosed compared with 73 of 100 who do not carry the fault. The research also found that women with the fault were more likely to develop a second cancer. In the study, 24 per cent of women with the fault developed a second breast cancer compared to seven per cent without the fault.

b) Applying Classification Technique

For classification, we use Support Vector Machine (SVM) because Text categorization problems are linearly separable, which has been shown to perform very well for text classification by many researchers. Which provide consistent improvement in accuracy over other Conventional learning methods like Naive Bayes classifier, Rocchio algorithm, K-nearest Neighbors and Decision tree classifier (e.g. [10 and 12]). We used the LIBSVM package [10] for all our experiments.

4. Experiments and Discussion

Experiments contain two parts: web page cleaning and classification on cleaning results of records.

4.1 Experiments on Web Page Cleaning

In our experiment we are presenting the results of our proposed approach for removing the noises form the web pages.

The setup has been implemented in .net framework 3.5 (visual c#) and has been performed on Windows 7 Professional with 2.13 GHz Intel(R) Core(TM) i3 CPU machine with 3 GB RAM. For this particular experiment, we have taken 750 web pages from 2 medical news web sites, including E-cancer (http://ecancer.org/news/), Medical News Today (http://www.medicalnewstoday.com/sections/tuberculosis/) and (http://www.medicalnewstoday.com/sections/asthma-respiratory/). These web pages contain news/articles from 3 categories, including Cancer, Tuberculosis (TB), and Asthma. These pages which contain noisy blocks are then made to pass through our proposed approach for removing noises form the web pages. Finally, we get pure text files cancer.txt, tuberculosis.txt and asthma.txt which are noise free and having all the news/articles related to them. Details of distributions of pages are shown in Table 1.
4.2 Experiments on Web Page Classification

This section presents the classification accuracy and efficiency on the three datasets obtained after the above experiment. Libsvm is used for classifying the data. Classification on the dataset is performed by dividing the documents into training data and testing data. Documents in the training set know their class labels, and are used for training classifiers. Testing documents are used to test the accuracy of the trained classifiers.

Table 1. Details of distributions of pages.

<table>
<thead>
<tr>
<th>Web Sites</th>
<th>Number of Web Pages in collection</th>
<th>Cancer News/Articles</th>
<th>Tuberculosis News/Articles</th>
<th>Asthma News/Articles</th>
<th>Total News/Articles in Site</th>
</tr>
</thead>
</table>

We have taken 80% of the data from each category to train our classifier and 20% of the data is used for testing. We got 73.3333% average accuracy for the three classes. The average accuracy and gamma values are shown in table 2. A counter graph is used to evaluate the performance of our classifier as shown in figure 4.

Table 2. The values of C and gamma with their accuracies.

<table>
<thead>
<tr>
<th>Class</th>
<th>C</th>
<th>Gamma</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cancer</td>
<td>2</td>
<td>2</td>
<td>73.3333%</td>
</tr>
<tr>
<td>Tuberculosis</td>
<td>0.03125</td>
<td>0.0078125</td>
<td>66.668%</td>
</tr>
<tr>
<td>Asthma</td>
<td>2</td>
<td>2</td>
<td>73.3333%</td>
</tr>
</tbody>
</table>
5. Conclusion

The proposed “WEB INFORMATION RETRIEVAL & CLASSIFICATION FRAMEWORK” consists of two modules: web page cleaning and main text extraction and text classification. In module 1 web pages are cleaned using VIPS and the main text is extracted and module 2 deals with the classification on the retrieved text to generate a classifier. Our experiment shows the average accuracy of three classes is 73.3333%. We finally conclude our work by proving that we developed very efficient and reliable web information retrieval & classification framework. Technique we developed is mainly useful for noise removing and classification of web pages. In our work we only take three classes of news that could be extended up to large level of classes and subclasses. As we defined above our classifier is working on multiclass classification so that the classes can be extended.

10. References


[6] Mohsen Asfia, Mir Mohsen Pedram, and Amir Masoud Rahmani, Main Content Extraction from


[8] Jing Li and C.I. Ezeife, Cleaning Web Pages for Effective Web Content Mining. This research was supported by (NSERC) of Canada under an Operating grant (OGP-0194134) and a University of Windsor grant.


1 (http://www.w3.org/DOM/).