2011 3rd International Conference on Information and Financial Engineering
IPEDR vol.12 (2011) © (2011) IACSIT Press, Singapore

Extracting Data Region in Web Page by Removing Noise using DOM and Neural Network

Thanda Htwe, Nan Saing Moon Kham

University of Computer Studies, Yangon
tdhtwe80@gmail.com, moonkhamucsy@gmail.com

Abstract. Today, with the explosive growth of information sources available on the World Wide Web, it has become increasingly difficult to identify the relevant pieces of information because Web pages are often cluttered with distracting features around the body of an article that distract Web user from the actual content they are interested in. An important problem for information extraction from the web is the identification and removal of noise. In this paper, we investigate 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. In our task, a possible application of Neural Networks 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 data used in training and testing is collected from several Web sites. The classification result of back propagation neural network is used for eliminating various noise patterns from Web page. In order to evaluate our proposed system, we perform the experiment on several Web pages of different News Web sites and Commercial Web sites as data set. Experiments indicate our method is applicable to extract informative content from Web pages of these Web sites.

Keywords: Noise Removing, Classification, Content Extraction.

1. Introduction

The amount of information that is currently available on the Internet is HTML Web pages. Most Web pages are cluttered with guides and menus attempting to improve the user's efficiency, but may end up distracting from the actual content of interest. With the exponentially growing amount of information available on the Internet, an effective technique for users to discern the useful information from the unnecessary information is urgently required. However, there is a lot of redundant and irrelevant information on the Internet [1], such as contents of mirror sites or identical pages with different URLs. We call this kind of redundancy global noise. We only focus on the local noise within a single web page. In that web page, it is also important to distinguish valuable information from noisy content that may mislead users' attention.

1.1 Web Noise

As we all know Web pages are constructed not only main contents information like product information in shopping domain, job information in a job domain but also advertisements bar, static content like navigation panels, copyright sections, etc. In many Web pages, the main content information exists in the middle block and the rest of page contains advertisements, navigation links, and privacy statements as noisy data. However, these noisy data formed in various patterns in different Web sites. When we extract only relevant information, such items are irrelevant and should be removed. Informative content for Commercial sites include product information such as various kinds of food, furniture and human accessories. News sites contain news information such as business, sport, health, leisure travel and politics.

Most approaches to removing clutter or making content more readable involve changing font size or removing HTML and data components such as images, which takes away from a webpage's inherent look and feel. The proposed system employs an easily extensible set of techniques (DOM and Neural Network model), for multiple noise patterns removing and content extraction from HTML web pages and is practically usable by the end user. In many previous studies from [11], [12], [13], they used a neural network with the capability of detecting normal or attack connections and attack type classification systems.

The remainder of the paper proceeds as follows. We first present related work in Section 2, and Section 3 gives our proposed method for informative content blocks and noise blocks detection. Section 4 shows several experiments to evaluate the effectiveness of our proposed method. Finally, in Section 5 we provide conclusions.

2. Related Work

There is a large body of related work in informative content identification and extraction that attempts to solve similar problems using various techniques. While many algorithms for content extraction already exist, few working implementations can be applied in a general manner. Several methods have been explored to extract informative content from a web page using vision-based and common layout template.

2.1 Informative Content Extraction Techniques

Information extraction systems try to extract useful information from either structured or semi-structured documents. In [5] and [6], a tree structure is introduced to capture the common presentation style of web pages and entropy of its elements is computed to determine which element should be removed. They extract keywords from each block to compute its entropy, and blocks with small entropy are identified and removed. Another intuitive way of page segmentation is based on the layout of webpage.

2.2 Web Page Segmentation Techniques

Several methods have been explored to segment a web page into regions or blocks. In the DOM-based segmentation approach, an HTML document is represented as a DOM tree. Another intuitive way of page segmentation is based on the layout of webpage. In this way, a web page is generally separated into 5 regions: top, down, left, right and center [3]. Another application of block importance is on web page classification [5][6]. The drawback of this method is that such a kind of layout template cannot be fit into all web pages.

Deng Cai [4] have introduced a vision-based page segmentation (VIPS) algorithm. This algorithm segments a Web-page based on its visual characteristics, identifying horizontal spaces and vertical spaces delimiting blocks much as a human being would visually identify semantic blocks in a Web-page. Another work that is closely related by [2] is the VIPS (Vision-based Page Segmentation) algorithm. In VIPS, a tree structure is used to model the page. Each node corresponds to a block in a page, and has a value to indicate the Degree of Coherence (DoC). The DOM tree is analyzed from root to leaves and the DOM nodes are divided based on their spatial layout and several other visual cues, such as color. In our work, we choose to break the DOM nodes down to the deepest level, i.e., collecting only the naturally undividable DOM nodes. (The details are explained in the following sections.)

None of these approaches perform limited analysis of web pages themselves and in some cases information is lost in the analysis process. By parsing a webpage into a DOM tree, we have found that one not only gets better results but has more control over the exact pieces of information that can be manipulated while extracting content.

3. The proposed system

The main contribution of this paper is three-fold. First, our system builds DOM tree structure for an incoming Web page and then split it into sub trees to get input data for neural network in second. In the last, we also apply back propagation neural network algorithm for web page's region classification such as data, mixture and noise group. The detail explanation of our system as follows.

3.1 DOM

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 [7]. 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 [8]. 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 [9]. 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.

$X_i = S_n/T_n$ eq.1

Where, S_n be the number of occurrence of same leaf nodes in sub-tree, T_n 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. Then we apply back propagation neural network algorithm to classify three classes, noise, data and mixture (data and noise region). Lastly, we remove the noise class in Web page and show extracted main content data in HTML page.

3.2 Informative Content block Identification using Neural Network

Neural network (ANN) model can be known as a good problem solving method for problems that can't be solved using conventional algorithms. If the input is one it has never seen before, it produces an output similar to the one associated with the closest matching training input pattern [10]. In neural network model architecture, each node at input layers receives input values, processes and passes to the next layer. This process is conducted by weight which is the connection strength between two nodes.

A Neural Network is a structure which is composed of a number of simple elements or nodes called neurons. These elements are always operating in parallel. The function of the Neural Network is determined largely by the connection between the neurons. These neurons are connected by links and each link is adjusted by values called weights. The process of updating the weights is called learning. Neuron input p associated with weight w and there is a scalar bias b. Eq.2 forms an input to the second component which is the transfer function.

$n=wp+b$ eq.2

The output of the neuron is the output of the transfer function. The general equation is

$a = f(wp+b)$ eq. 3

Here f is a transfer function which takes the argument n and produces the output a. The Neural Network will exhibit the desired or interested behaviour by adjusting its parameters. That means, the Neural Network can be trained to a particular job by adjusting the weight or bias parameters or perhaps the network itself will adjust these parameter to achieve some desired results.

The input *p* to the neuron can be expanded to *R*-elements input and each input is multiplied by weight. Their sum is simply (W●P) which is the dot product of the matrix W and the vector P. The argument n which is the input to the neuron transfer function will be:

$$n = w1,1\, p1 + w1,2\, p2 + w1,3\, p3 + \ldots + w1,R\, p\, R + b \qquad \text{eq.4}$$

One of the most commonly used Neural Networks is the multilayer feed-forward network. It falls under the category called "Networks for Classification and Prediction". Our system is built using this specific type of Neural Network. Our classification model consists of two layers in which the neurons are logically arranged. The last layer is the output layer and only one hidden layer.

To train the model, randomly selected several Web pages of different web sites used as a data set. The present study is aimed to solve a multi class problem in which not only noise patterns are distinguished from Web page, but also data and mixture patterns are identified. All the implemented neural networks had fifteen neurons and two output neurons [0, 1] for noise pattern, [1, 0] for data pattern and [1, 1] for mixture pattern. The widely used learning method, back propagation algorithm is used to train. It has been found that the standard sigmoid activation function is suitable on both layers for web page's region classification model.

The classification result of neural network is used for eliminating various noise patterns. Noisy information in web pages may decrease not only depends on the accuracy of classification result but also the threshold level of DOM tree.

4. Experimental Results

Our experiments evaluated by using Commercial and News Web pages collected from different Web sites in Table 1. While development is still ongoing, we will show our initial experiments using different structure of Web pages. The implemented system solved a three class problem. Our approach does not depend on manually specified rules or any domain knowledge or semantic information. The training performance is measured using the mean square error (MSE). As mentioned before, the MSE is the difference between the target and the Neural Network's actual output. So, the best MSE is the closest to 0. If MSE is 0, this indicates Neural Network's output is equal to the target which is the best situation. The learning rate is 0.6 and bias value is 0.4. The number of iteration (epochs) we used is 3000 for the MSE 0.005.

Table. 1 : Web Page's Region Classification Result for Selected Data Set

URL	Maximum Level	Training Patterns	Data	Mixture	Noise
www.Indianexpress.com/	8	17	46.1	0	53.9
www.CBSnews.com/	11-12	22	0	10.3	89.7
www.CNN.com/	22	19	0	7.4	92.6
www.Foxnews.com/	10-14	20	12.5	37.5	50
www.Bing.com/	19-20	16	68.18	0	31.82
www.Productwiki.com/	10	27	0	46.15	53.85
www.Infobanc.com/	20	9	0	29.41	70.59

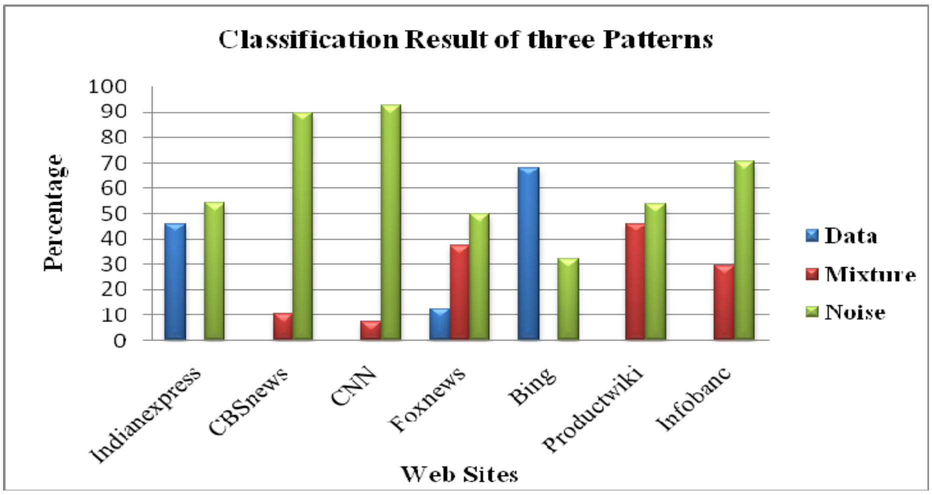

Fig. 1: Percentage of consisted Noise, Mixture and Data for each Web Site

Noise removing accuracy of this system depends on not only the correct classification result of neural network but also detail splitting of sub trees. It is more important to split multiple sub trees under heuristics of the system because Web pages are constructed with complex and different structure for variety of Web sites. We can classify the three regions within web page for selected Web sites with values in percentage as shown in Table 1.We can detect the pure data region for Indianexpress and Bing. Even though the structures of CBSnews, CNN, Productwiki and Infobanc web sites are simple, we cannot determine pure data region because data regions of these sites are surrounded by noise. Moreover, we found the three blocks structure for Foxnews web site. Fig. 1 shows the percentage of noise class and data or mixture class involved in various Web pages of selected Web sites. By combining the DOM tree analysis and back propagation Neural Network algorithm, we can effectively eliminate various noise patterns and extract informative Web data with 100% precision and recall for these web sites except from Indianexpress web site with 95% accuracy rate.

5. Conclusion

To evaluate the performance of the system, we calculated how much sub trees of noise class detect and eliminate for each Web page of selected Web sites is shown in Table 1. We calculated the value percentage for each region (data, mixture and noise) from the number of noise trees are divided by the total number of sub trees that the system collected. The implemented system solved three classes problem and remove noise class. We develop the system not only to classify three classes in Web page but also to remove the noise class based on classification result. The proposed system effectively eliminates all noise groups from the selected seven web sites. Noise removing accuracy of this system not only depends on the correct classification result of neural network but also the threshold level for splitting DOM sub-trees.

6. References

[1] A. Broder, S. Glassman, M. Manasse, and G. Zweig. Syntactic Clustering of the Web. *In Proc. of Sixth World Wide Web Conference*, 1997.

[2] D. Cai, S. Yu, J.-R. Wen and W.-Y. Ma. Extracting Content Structure for Web Pages Based on Visual Representation. *In Proc. of 5th Asia Pacific Web Conference*, 2003.

[3] M. Kovacevic, M. Diligenti, M. Gori and V. Mulutinovic. Recognition of Common Areas in a Web Page Using Visual Information: a possible application in a page classification*. In Proc. of IEEE International Conference on Data Mining (ICDM'02)*, Maebashi City, Japan, December, 2002

[4] D. Cai, S. Yu, J.-R. Wen and W.-Y. Ma. Block based web search. *In Proc. of 27th Annual International ACM SIGIR conference on Research and development in information retrieval,* pp: 456 – 463, 2004

[5] L. Yi and B. Liu. Web Page Cleaning for Web Mining through Feature Weighting. *In Pro.c of Eighteenth International Joint Conference on Artificial Intelligence* (IJCAI-03), Acapulco, Mexico, August, 2003.

[6] L. Yi and B. Liu. Eliminating Noisy Information in Web Pages for Data Mining, *In Proc. of the ACM SIGKDD International Conference on Knowledge Discovery& Data Mining* (KDD-2003), Washington, DC, USA, August, 2003.

[7] S. Gupta, G. Kaiser, D. Neistadt and P. Grimm. DOM Based Content Extraction of HTML Documents. *In Proc. of the WWW2003 Conference*, May 20-24, 2003, Budapest, Hungary.

[8] http://www.bearcave.com

[9] http://www.w3.org/DOM/

[10] http://www.learnartificialneuralnetworks.com

[11] James Cannady. Artificial neural networks for misuse detection. *In Proc. of the 1998 National Information Systems Security Conference* (NISSC'98), Arlington, VA, 1998.

[12] J. Ryan, M. Lin, and R. Miikkulainen. Intrusion Detection with Neural Networks. *AI Approaches to Fraud Detection and Risk Management: Papers from the 1997 AAAI Workshop,* Providence, RI, pp. 72-79, 1997.
Srinivas Mukkamala. Intrusion detection using neural networks and support vector machine. *In Proc. of the IEEE International Honolulu*, HI, 2002.