An Efficient Approach for Web Data Extraction

Thanda Htwe, Nan Sai Moon Kham
University of Computer Studies, Yangon, Myanmar
tdhtwe80@gmail.com, moonkham.ucsy@gmail.com

Abstract

Most of the Web page typically contains clutter unlike conventional data or text. It usually has such noise data as navigation panels, copyright and privacy notices, and advertisement. These noise data can seriously harm for Web miners by extracting whole document rather than the informative content and also retrieve non-relevant results. So, eliminating these noise patterns is great important. In this paper, we propose an effective technique to detect and remove various noise patterns from Web document to enhance Web mining. Our system first builds DOM tree structure for an incoming Web page and then split it into sub-trees to detect noise data. We also apply back propagation neural network algorithm to classify various noise patterns, data patterns and mixture patterns in current Web page. The classification result of neural network is used for eliminating various noise patterns. The proposed technique is evaluated on several commercial Web sites and News Web sites to show the performance and improvement of our approach.

1. Introduction

Lots of information is available on the Internet. It has billions of web pages occupying significant portion of it and the number is still growing rapidly. Web pages are structured to include not only informative content but also advertisements, static content like navigation panels, copyright notice etc. These noise data in the Web page significantly affect on the performance of the Web data mining. It is also important to distinguish valuable information from noisy data within a single Web page.

When we process Web documents, the main content is surrounded by noise in the retrieved data. Therefore, without removing such data, the efficiency of feature extraction and finally text classification is certainly degraded. Web noise can be classified into global noises and local noise by Yi and Liu [1] [4]. Global noises include mirror sites, legal/illegal duplicated Web pages, old versioned Web page with advertising segments, unnecessary images, or navigation links, etc.

Many studies on information extraction (or information retrieval) also try to discover informative content from a set of Web documents [11]. Extraction of "useful and relevant" content from web pages has many applications, including cell phone and PDA browsing, speech rendering for the visually impaired, and text summarization. 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. However, it is relatively little work has been done on eliminating noisy data from Web pages in the past. Hence, we mainly focus on efficiently and automatically detecting and removing noisy data from Web pages to extract only relevant information.

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. Web pages are often cluttered with distracting features around the body of an article that distract the user from the actual content they're interested in. These "features" may include pop-up advertisement, flashy banner advertisements, search and filtering panel, unnecessary images, or links scattered around the screen. 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.

Therefore, we propose the mechanisms in this paper to eliminate various noises in Web pages to reduce irrelevant and redundancy data. We apply back propagation neural network algorithm to classify various noise patterns. And then we remove these noises in current page for content extraction. The utilization of a neural network in the detection of instance of noise pattern would be the flexibility that the network would provide. A neural network would be capable of analyzing the data from the network, even if the data is incomplete or distorted. A neural network might be trained to recognize

known suspicious events with a high degree of accuracy.

The rest of this paper is organized as follows. Section 2 introduces some related works. Then, in Section 3, we illustrate the representation of noisy data in a page and present our proposed system and the paper concludes with the conclusion and future work in Section 4.

2. Related Work

There are some existing approaches to discover informative content blocks. Most of them have focused on detecting main content blocks in Web pages. Although cleaning noisy data is an important task, relatively little work has been done in this field.

2.1 Informative Content Extraction

An approach that allows for fully-automated extraction of content based on distilling linguistic and structural features from text blocks in HTML News Pages is proposed in [8]. In this approach, content extraction is applicable to any type of news pages using thresholds learned by the Particle Swarm Optimizer. However, human effort is required to label documents for classification. Hung-Yu Kao and Shian- Hua Lin was proposed an approach to discover informative contents from a set of tabular documents of a Web site by dynamically select the entropy threshold in [6]. The system first partitioned a page into several content blocks according to HTML tag <TABLE> in a Web page. The system is not applicable general Web pages which is consisted using tag <DIV>. Song et. al [9] investigated mechanisms for segmenting Web pages by assigning importance values to blocks. The next proposals [12] for content extraction take into account visual cues of Web page rendering, e.g., the distance of conceptual blocks from the screen's center, the alignment of page segments, and so forth. The systems make use of rendering engines in Web browsers, which translate directly between HTML elements and their positioning within the browsing window.

2.2 Web Noise Cleaning

In [7] proposed a redundant information elimination approach in the Web documents from the same URL path. Three filtering methods, tag based filtering, redundant words filtering and redundant phrase filtering, are used in their system. A redundant word/phrase filtering method is used for single or multiple tokenizations. The approach mentioned in [10] builds site style tree in simplistic manner, which is generalized DOM tree presentation of related pages. Noisy elements in the tree are detected based on entropy calculation over set of features. They identified the common presentation style and content and then compressed them into site style tree. To construct the site style tree, the system needs to learn the whole web site to detect the common presentation style and content. The similar problem is addressed in [11] which based on entropy calculations over set of features, but without using site style tree.

Some Web sites are structured with dynamic Web pages and their content and presentation style are not common. It is difficult to detect different noise patterns for those Web sites by using above technique. These techniques are less successful in identifying noise patterns which vary from expected patterns. This paper, therefore, proposes an effective technique to eliminate multiple noise patterns in Web page for information extraction no need to learn the whole Web sites. Our solution employs multiple extensible techniques that incorporate the advantages of the previous work on content extraction.

3. Our Approach

In this section, we illustrate the two mechanisms in DOM-based analysis to extract informative content of Web document. We firstly transform a current web page into DOM tree structure and split it into sub-trees. And then we apply back propagation neural network algorithm to classify three classes, noise, data and mixture (data and noise region). Lastly, we remove the various noise patterns in Web page and show extracted main content data. The architecture of our proposed approach is shown in Figure 1.

3.1 DOM Tree

This system firstly constructs DOM structure of web pages of many web sites, Commercial web sites and News web sites and split them into sub-trees to use them as input nodes of neural network for classifying three classes. DOM trees remain highly editable and can easily be reconstructed back into a complete webpage. In the DOM, documents have a logical structure which is very much like a tree; to be more precise, it is like a "forest" or "grove", which can contain more than one tree. However, the DOM does not specify that documents must be *implemented* as a tree or a grove, nor does it specify how the relationships among objects be implemented. The DOM is a logical model that may

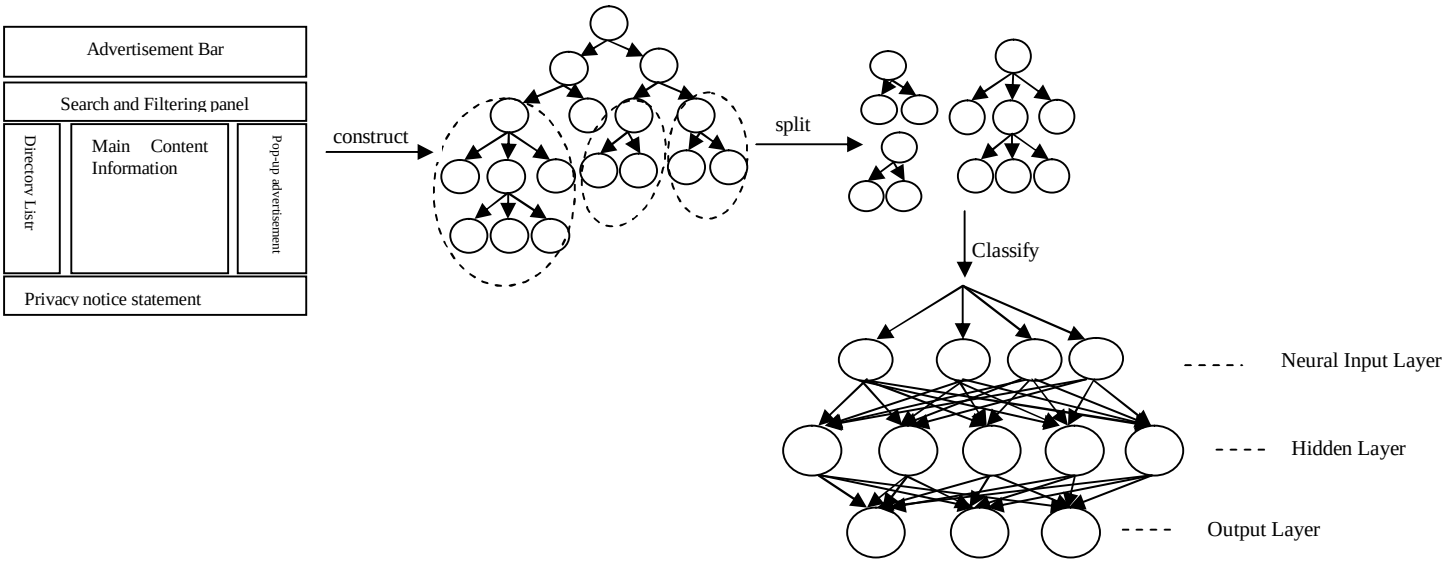

Figure 1. Flow of proposed approach

be implemented in any convenient manner. The DOM tree is hierarchically arranged and can be analyzed in sections or as a whole, providing a wide range of flexibility for our proposed method. By parsing a webpage into a DOM tree, more control can be achieved while eliminating noise data. Moreover, increasing support for the Document Object Model makes our solution widely portable. We use the Xerces HTML DOM [14]. Noise data of Web documents can be categorized into two groups such as global noise and local noise [7]. Global noises are redundant Web pages over the Internet such as mirror sites and legal or illegal duplicated Web pages. Local noises only related intra-page redundancy and exist in the Web page. This paper only focuses on the local noise elimination method. There are at least four different known categories of noise pattern within Web pages of any Web sites including banners with links including search panels, advertisements, navigational panel (directory list) and copy right and privacy notice in each Web site.

Some Web pages are not well formed documents, it is necessary to make well formed document before processing them. We first check the syntax of HTML Web page using online HTML tidy tools [12] before parsing Web page into DOM tree structure. After parsing it, DOM tree structure divided it into several sub trees according to threshold level.

3.2 Artificial Neural Network Model

Artificial 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.

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. The key feature of neural networks is that they learn the input/output relationship through training. There are two types of training used in neural networks: supervised and unsupervised training, of which supervised is the most common.

The model structure is trained with known samples of data. It is a learning scheme for updating a node's weights. In this phase, a pattern detected in the data set is presented to the user. It also corresponds to a particular abstract learning task. Nodes apply on iterative process to the number of inputs to adjust the weights of the network in order to optimally predict the sample data on which the training is performed.

To train the model, randomly selected several Web pages 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. Neural networks process numeric data in a fairly limited range. So, currently enter Web page is parsed into sub-trees according to the threshold level and then converted these sub trees into a standardized numeric representation by using eq (1).

$$x_i = \frac{S_n}{T_n} \qquad \text{eq (1)}$$

where x_i be input nodes at input layer, S_n be the number of occurrence of same leaf nodes in sub-tree and T_n be the total number of leaf nodes in sub-tree.

Here, three classes are described which can be extended to cases with more noise types. An output layer with three neurons output states was used: [1 0 0] for Noise class, [0 1 0] for Data class and [0 0 1] for Mixture (data and noise) class. All the implemented neural networks had sixteen neurons and three output neurons (equal to the number of classes). The number of hidden layers and neurons in each were parameters used for the optimization of the architecture of the neural network. The widely used learning method, back propagation algorithm is used to train for classifying which case is probable based on current Web page. It has been found that the standard sigmoid activation function is suitable for modeling the occurrence of noise pattern using the pattern matching and learning model. During training process in neural network, we can occur one problem. In an over fitted ANN, the error (number of incorrectly classified patterns) on the training set is driven to a very small value, however, when new data is presented, the error is large. One possible solution for the over-fitting problem is to find the suitable number of training epochs by trial and error.

4. Experiments and Results

In this section, we describe experiments on six different commercial Web sites and three different News Web sites as data set. We implemented a three layer MLP (one hidden layer with fourteen neurons) network. One of the objectives of the present study is to evaluate the possibility of achieving the same results with this less complicated neural network structure. In this technique, the available data is divided into three subsets. The first subset is the training set, which is used for training and updating the ANN parameters. The second subset is the validation set. The error on the validation set is monitored during the training process. The third is used for testing. Table 1 shows detailed information about the number of records such as noise, data and mixture for training, validation and testing sets.

Table 1. Number of records used as data set for three subsets

Class	Training Set	Validation Set	Test Set
Noise	200	50	100
Data	50	30	50
Mixture	50	30	50

The final correct classification rate on known test data is 99.8% for Noise class, 98.9% for Data class and 98.2% for Mixture class. However, unseen data (test set) was fed to the neural network, the correct classification rate was less than 75%. These classification result is reported in Figure 2. The classification rate on unknown test data is 83.7% for Noise class, 77.9% for Data class and 65.3% for Mixture class. Noise removing accuracy of this system is depend on the correct classification result of neural network. Therefore, we can eliminate various noise patterns for Web sites (www.productwiki.com, www.digg.com and www.infobanc.com) nearly 80% because these sites are structured with various noise patterns and data patterns seperately. However, the system can eliminate variety of noise patterns with accuracy rate less than 70% for two third of all data set which are structured by mixing noise and data regions. So, noise removing accuracy degrade for these sites.

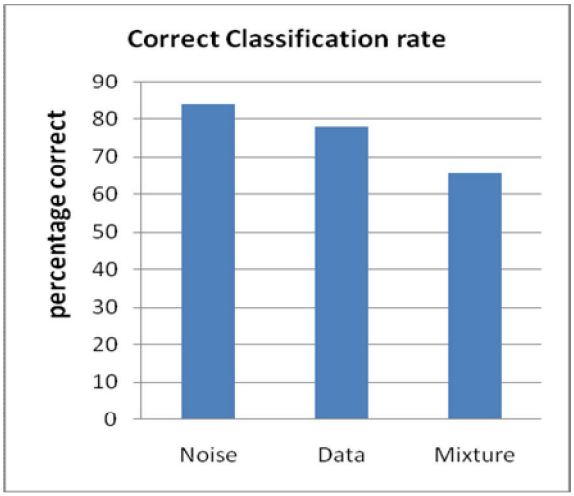

Figure 2. correct classification result

5. Conclusion

The goal of the proposed system is to extract web informative content by cleaning noise data from web pages for the purpose of improving the accuracy and efficiency of web mining. The implemented system solved a three class problem. The most popular back propagation algorithm of neural network was also successfully used for the classification of correct classes. Benefit to us by learning from past experience to deal with new and unexpected situations.

6. References

[1]. Yi, L., B. Liu, and X. Li. *Eliminating Noisy Information in Web Pages for Data Mining*. in *Proceedings of the ACM SIGKDD International*

Conference on Knowledge Discovery & Data Mining (KDD-2003). 2003. Washington, DC, USA.

[2]. Hsu, C. N. and Dung, M. T., "Generating Finite-state Transducers for Semi-structured Data Extraction from the Web," Information Systems, 23(8):521-538, 1998.

[3]. Kushmerick, N., "Wrapper Induction for Information Extraction," Ph.D. Dissertation, Department of Computer Science and Engineering, University of Washington, 1997.

[4]. Yi, L. and B. Liu. *Web page cleaning for Web mining through feature weighting*. in *Proceedings of Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03)*. 2003. Acapulco, Mexico.

[5]. S.-H. Lin and J.-M. Ho. Discovering informative content blocks from web documents. In *KDD*, pages 588-593. ACM, 2002.

[6]. H.-Y. Kao and S.-H. Lin, Mining Web Informative Structures and Contents Based on Entropy Analysis. In *IEEE*, 2004

[7]. B. H. Kang and Y. S. Kim. Noise Elimination from the Web documents by using URL paths and Information Redundancy.

[8]. C.-N Ziegler and M. Skubacz. Content Extraction From New Pages using Particle Swarm Optimization. In *IEEE*, 2007

[9]. R. Song, H. Liu, J.-R, Wen, and W.-Y. Ma. Learning block importance models for Web pages. In WWW, 2004.

[10]. L. Yi, B. Liu, and X. Li. Eliminating noisy information in Web pages for data mining. In KDD, 2003.

[11]. H., -Y. Kao, J. –M. Ho, and M. –S. Chen. Wisdom Web intrapage informative structure mining based on document Object model. *IEEE* Trans KDD, 2005.

[12]. Y. Yang and H.-J. Zhang. HTML page analysis based on visual cues. In *Proceedings of the Sixth International Conference on Document Analysis and Recognition*, pages 859– 864, Washington, DC, USA, 2001. IEEE Computer Society.

[13]. http://infohound.net/tidy

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