Elimination noisy information on web page

Aye Pyae Sone, Nwe Nwe
UNIVERSITY OF COMPUTER STUDIES, HPA-AN
ayepyaesone5@gmail.com

Abstract

Web page typically contains many information blocks. They are navigation panels, copyright and privacy notices and advertisements. These blocks are useful for business purposes. These blocks are called as the noisy blocks which can harm web data mining. And so, eliminating these noises is of great importance. The noisy blocks usually share some common contents and presentation styles. The main contents of web page are different in the common presentation styles. Based on this observation, a site style tree (SST) is presented in this system to capture the common presentation styles and actual contents. An information based algorithm is used to determine which parts of the SST represent noises and which parts represent the main contents of the site. Experimental results show that eliminating noisy information on web pages will be effective for web data mining. The system shows how much noisy information blocks can be removed from web pages depending upon file size. The users can choose desired web page and this system will eliminate unnecessary noise by using noise detection and web page cleaning algorithm.

Keywords:
noise detection, noise elimination, web mining.

1. Introduction

Internet is a rapid expansion for World Wide Web (WWW) which is 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. But, information on web pages is accompanied with a large amount of noise such as navigation panels, copyright buttons, privacy notices, decoration pictures, and advertisements.

Although such information are useful for human viewers and web site owners, which are seriously harm information gathering and web data mining. Web noise can be partitioned into two categories according to their granularities:

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

Local noise: 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.

This paper emphasizes to detect local noise in web pages to improve the performance of web mining. For example, clustering, classification and information retrieval, etc. Local noise can be web mining to produce poor results. In the figure 1, only section 3 is required for web mining. The rest of the page contains many advertisements, navigation links, magazine subscription forms, privacy statements, etc. If the researchers perform clustering on a set of product pages like this page, such items are irrelevant and should be removed.

Figure 1 Example web page with noise

2. Related Work

This paper is presented to detect and eliminate noise to improve web data mining. A highly effective technique is presented to clean web pages with the purpose of preventing in web mining poor results. A frequency based data mining algorithm is presented to detect templates and views those templates as noise. This

partitioning technique is simple and useful for a set of web pages [1]. Some learning mechanisms recognize banner ads, redundant and irrelevant links of Web pages [3], [4]. Web page cleaning is also related to feature selection in traditional machine learning [5]. For the web cleaning application, the system uses the local template detection algorithm in [2] to detect templates.

3. System overview

This paper involves five main stages.

3.1 Changing HTML to DOM: In this step, HTML pages are changed corresponds to a document object model (DOM) tree where tags are internal nodes and the detailed texts, images, or hyperlinks are the leaf nodes.

```
< BODY bgcolor = WHITE>
  <TABLE width= 800 height= 200>
   …..
  </TABLE>
  <IMG src = "image.gif" width= 800>
  <TABLE bgcolor = RED>
  …..
  </TABLE>
</BODY>
```

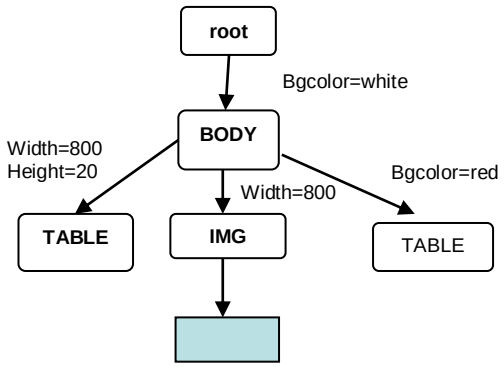

Figure 2 DOM tree

3.2 Merging DOM to SST: DOM trees are not sufficient in the cleaning technique, and so SST is presented which purpose is to compress presentation styles of a set of related Web pages. In the figure 3, except for the four tags (P, IMG, P and A) at the bottom level, all the tags in *d1* are the same in *d2*. Thus, *d1* and *d2* can be compressed. The right TABLE node in the style node figure has two different presentation styles. And so, these are common presentation style.

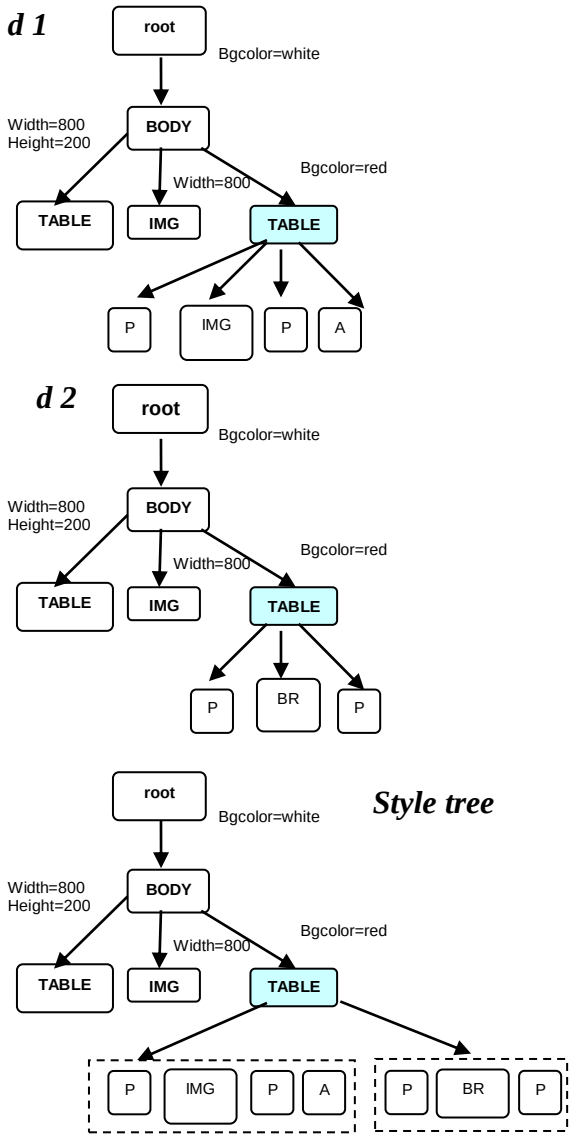

Figure 3 DOM trees and site style tree

3.3 Noise detection: The nodes in SST can be known that which are noises, and which are the main contents by defining the node importance. Simplified SST can be got by using noise detection algorithm.

In the noise detection algorithm, noises are defined as the following assumption. For an element node (E) in the SST, if all of its descendents and itself have composite importance less than a specified threshold (t), that element node is noise.

3.3.1 Noise detection algorithm:

Input: E: root element node
Output: TRUE if E and its descendents are noise, else FALSE
Method:
(1) Mark noise (E)
    if(MarkNoise (E)== FALSE) then
        return FALSE

else return TRUE

(2) if (E.CompImp<= t) then
Mark E as noisy
return TRUE
else return FALSE

3.3.2 Calculating entropy measure

$$CompImp\ (E) = \begin{cases} 1 & if\ m = 1 \\ 1 - \frac{\sum_{i=1}^{l} H(a_i)}{l} & if\ m > 1 \end{cases} \quad (1)$$

l be the number of features (words, image files, link references, etc). m be the number of pages containing E. a_i is an actual feature of content in E. $H(a_i)$ is the information entropy of a_i.

$$H(a_i) = -\sum_{j=1}^{m} p_{ij} \log_m p_{ij} \quad (2)$$

P_{ij} is the probability that a_i appears in E of page j.

Calculating CompImp is 1 (all the values of CompImp are normalized to between 0 and 1).
The cleaning algorithm needs threshold (t) value to decide noisy and meaningful nodes.

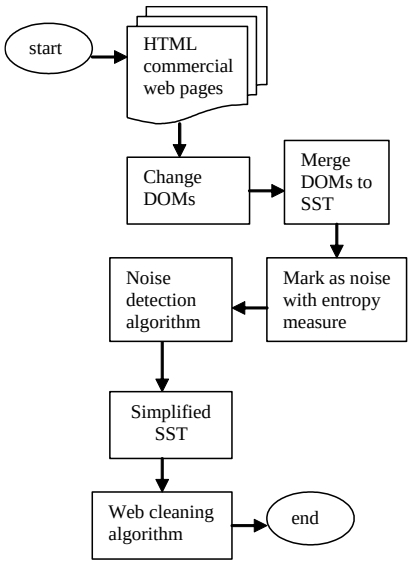

Figure 4 System overview

3.4 Web cleaning : Given a web site, the system first randomly crawls a number of web pages from the web site and builds the SST based on these pages. In many sites, all its pages could not be crawled because they are too large. By calculating the composite importance of each element node in the SST, maximal noisy nodes and maximal meaningful nodes can be found. By the above assumptions, web cleaning algorithm is utilized.

Web cleaning algorithm is as follows:
Input: a set of web pages
Output: a page without noise

Method:
(1): Randomly crawl k pages from the given web sites.
(2): set null SST with virtual root node.
(3): (re) assign each DOM tree to SST.
(4): compare element nodes in SST.
CalcCompImp(E);
MarkNoise(E);
Markmeaningful(E);
(5): display result.

3.5 Output result: Finally, the required output result is got that does not contain unnecessary information blocks (noise).

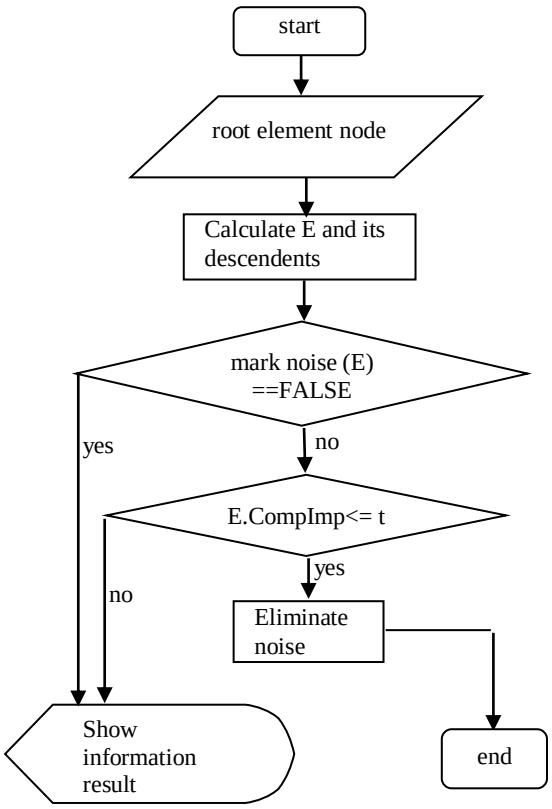

Figure 5 System flow diagram of noise cleaning

The cleaning algorithm needs to decide noisy nodes and meaningful elements. For each website, a small number of pages (20) has been chosen, and then clean them. The result will show percentage of the portion eliminated noise by depending file size of web page.

As shown in figure 6, preprocessing menu contains web page files to eliminate noise. In this

menu, the user can add or remove files and he can see how much noisy information blocks can be removed from web pages depending upon file size. If the user browses the web pages and click “eliminate” button, he will see the web pages without noise. The users can choose desired web page and this system will eliminate unnecessary noise by using noise detection and web page cleaning algorithm. Furthermore, he can see how much size will be reduced on web pages. The user can choose the desired number of web pages on this system. But, the more file size, the longer the processing time. The processing time wholly depend on the web pages' file size. The user should choose only the required web pages which will be used on their web mining.

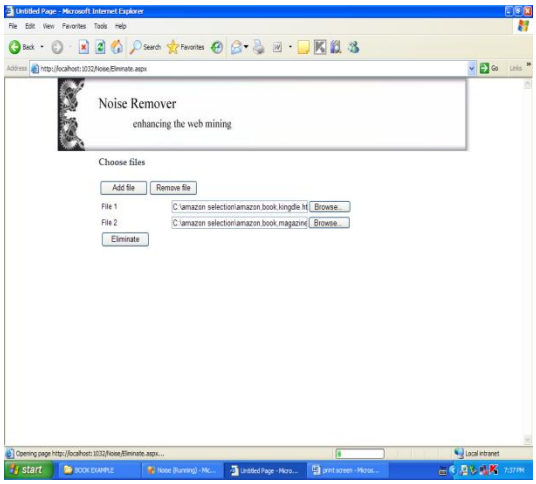

Figure 6 Process menu form

Figure 7 represents noise remove form. In the figure, if the user clicks “show result” button, he can see the web pages without noise.

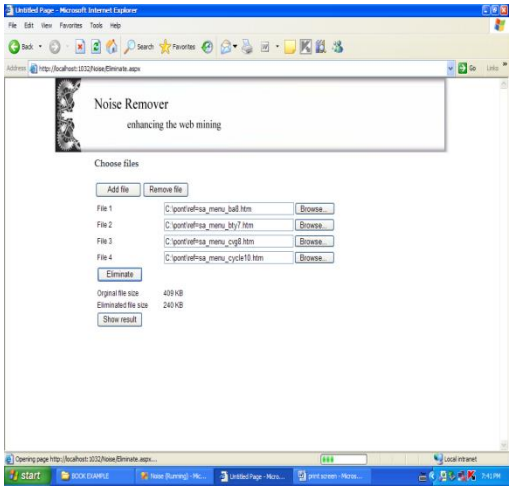

Figure 7 Noise remove form

4. Experimental Result

The experimental result shows the differences of file size on web pages before and after cleaning noise. It also compares the percentage of file size depending upon the amount of noise removed. The presented method is more effective than template based method [1]. The template detection method cleans all the pages from 5 sites altogether. The presented web page cleaning method cleans the web page in each individual site separately. And so, the presented cleaning technique proves to be more effective.

In the given commercial web page, product description could be defined as the main contents and another navigational links and decoration pictures are defined as noise and they are eliminated.

The performance of web mining results can be improved by eliminating noise. This noise elimination technique is effective for web mining researchers. However, the visitors (users) can not browse another link to get other information. The presented system works on HTML pages only.

The original web page can be found in figure 8. In this page, the user will see unnecessary links.

There are some portions needed to remove from web page. After removing noise, the user can see this as shown in figure 9.

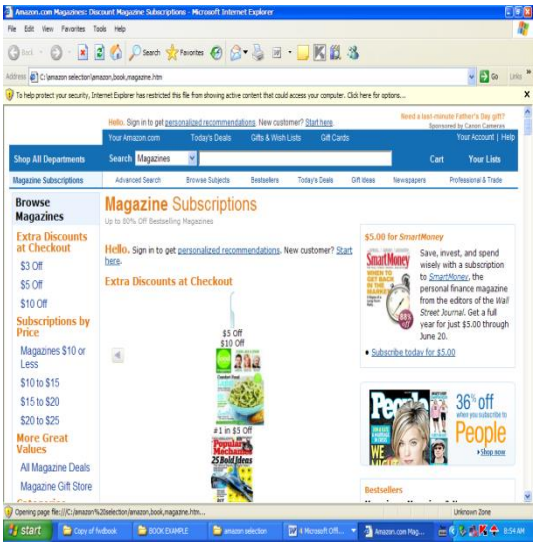

Figure 8 Original web page

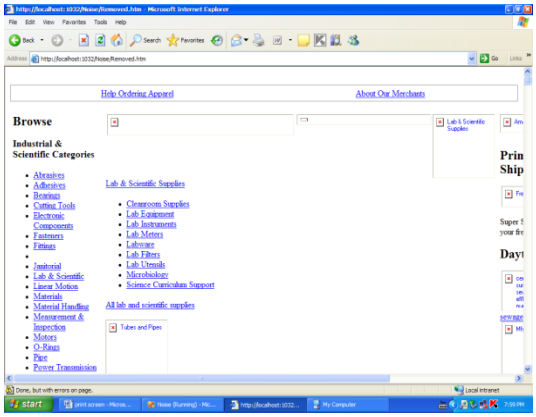

Figure 9 Show result form

5. Conclusion

This paper presents that the local noises in web page seriously harm web mining, and so web page cleaning algorithm is used to convenient web mining researchers. The site style tree (SST) is applied to capture the common presentation style and actual contents of the web page. The SST provides to see clearly which portions of the web page are necessary or unnecessary.

Information based measure is utilized to mark noise. Noise detection algorithm is also used to detect noise on web pages. The cleaning technique can protect misleading from web mining. If web mining without removing noise on web pages is performed, the researchers can get poor output result. By using the presented system, this is able to improve the results of web data mining.

References

[1] Bar-Yossef, Z. and Rajagopalan, S. *Template Detection via Data Mining and its Applications*, WWW 2002, 2002.

[2] Beeferman, D., Berger, A. and Lafferty, J. *Statistical models for text segmentation*. Machine learning, 34(1-3), 1999.

[3] Davision, B.D. *Recognizing Nepotistic links on the Web*. Proceeding of AAAI 2000.

[4] Jushmerick, N. *Learning to remove Internet advertisements*, AGENT-99, 1999.

[5] Yang, Y. and Pedersen, J.O. *A comparative study on feature selection in text categorization*. ICML-97, 1997.