Noise Elimination from Web Page in Web Content Mining

Khaing Wah Wah Linn, Sabai Phyu
University of Computer Studies, Yangon
khaingwah2linn@gmail.com, sabaiphyu72@gmail.com

Abstract

Nowadays, a large number of web pages contained useful information is often accompanied by a large amount of noise such as banner advertisements, navigation bars, copyright notices, etc. These noise data can seriously harm for web miners by extracting whole document rather than the informative content and also retrieve non-relevant results. It is also important to distinguish valuable information from noisy data within a single web page. The 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. When web documents are processed, the main content is surrounded by noise in the retrieved data. To tackle these issues, a noise elimination process is described by using html tags and main content is retrieved by using gomory-hu tree.

Keywords: *noise elimination, block splitting*

1. Introduction

Web content mining is the process of identifying user specific data from Text, Image, and Audio or Video data already available on the web. This process is alternatively called as web text mining, since text content is the most widely researched subjects on the World Wide Web. The technologies that are generally used in web content mining are Information retrieval and Natural language processing. Web Structure mining is another process of using graph theory to analyze the node and connection structure of a web site. Depending upon the type of web structural data, web structure mining has been divided into two fields. The first one is extracting patterns from hyperlinks on the web. The other one is mining the document structure. This involves using the tree-like structure to analyze and describe the HTML tags within the web page. Web usage mining is to identify user access patterns from Web usage logs. Web content mining identifies the useful information from the web contents/data/documents. However, such data in its broader form has to be further narrowed down to useful information. The web content data consists of structured data such as data in the tables, unstructured data such as free texts, and semi-structured data such as HTML documents.

2. Related Work

The simplest way to clean web pages is to remove metadata and tags the source data. A number of approaches have been reported in the literature for extracting information from web pages. Segmenting Web Pages & Detecting Noise is the technique [4] that normally Web page is consists of different blocks or areas, e.g., core content areas, navigation panels,

advertisements area, etc. It is automatically separated the area using this technique for several practical application. For example, in Web data mining, e.g., classification and clustering, identifying main content areas or removing noisy blocks (e.g., advertisements, navigation panels, etc.) enables one to produce much better results. It was shown in, that the information contained in noisy blocks can seriously harm Web data mining. To remove key information out of web pages that comprised of noisy information, a technique was introduced by Chao Wang et al. [5]. A.K. Tripathy and A.K. Singh [17] were introduced A technique which was employed to eliminate noise from web page. A tree structure, called the Pattern Tree was proposed to capture the general presentation styles and the definite essence of the pages in a specified Web site. A Pattern Tree called the Site Pattern Tree (SPT) was put up for the site, by sampling the pages of the site. A measure which was based on the information was then introduced to decide which parts in SPT represent noises, and which parts represent the core contents of the site. By mapping any Web page to the SPT, the noises were detected and eliminated from that particular Web page.

3. Method and Background Theory

3.1. Overview of Noise Elimination process

In this noise elimination process, two mechanisms that determine which region of current web page contain noise or mixture (data and noise region) is illustrated. Then another mechanism is devised on matching to determine how to process the three classes (noise, data and mixture) based on cases. Last, the various noise patterns are removed in current web page and show the extracted main content data. Figure 1 describes the detailed architecture of the noise deducting process.

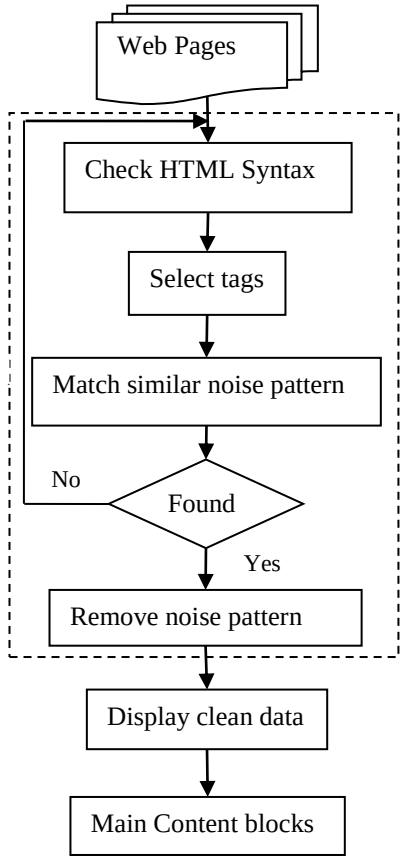

Figure 1. Noise Elimination

3.2. Web Page Noise

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 are interested in. These “features” may include pop-up advertisement, 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 extract, only relevant information, such items are irrelevant and should be removed. Therefore, the mechanism proposed in this work is to eliminate multiple noise patterns in web pages to reduce irrelevant and redundancy data. The HTML <DIV>, <Table>and <TD> tag is used to detect multiple noise patterns in current web page.

Noise data of web documents can be categorized into two groups such as global noise and local noise. 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 work focuses on the local noise elimination method.

Information in a web page is not uniformly significant. For example, consider the web page in Figure 2. Dissimilar information in a web page has dissimilar importance weight according to its location, occupied area, content, etc. Thus, it is to assign importance to a region in a web page, and need to segment a web page into a set of blocks as shown in figure 3. [1, 4]

3.3. Fixed Noise

Fixed noise is usually descriptive the information on a webpage or a website. It consists of three sub-types as shown in figure 1:

1. Decorating noise like site logos and decorative graphics or text, etc.
2. Statement noise, such as copyright notices, privacy statements, license notices, terms, partners or sponsors statement and etc.
3. Page description noise like date, time and visit counters of the current page and etc. [7]

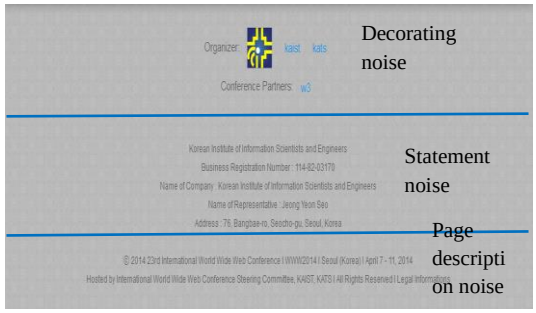

Figure 2. Example of Fixed noise

3.4. Navigational Assistance

Navigation Tool leadership is common in large websites as it helps users to the sites. It usually serves as intermediate guidance or a shortcut to the pages in a website. There are two main types of navigation guidance and leadership directory recommended guidance.

1. Guide line is usually a list of hyperlinks that lead to important index / portal page within a website. It usually reflects the subject categorization and / or topic hierarchies. The three guidance in styles are:

i. Global directory guidance shows the main topic categories of the current sites.

ii. Hierarchical leadership guide shows the hierarchical concept space of the current page within a given location.

iii. The Hybrid combines a guide leading the world wide leading directory and guide hierarchical leadership.

2. Recommendation guidance set web users with some potentially interesting sites. It comes in three styles

i. Advertise recommendation is usually a block of hyperlinks that lead to hot items for Web users. This is shown for commercial purposes. Those hot items are usually advertisements, offers and promotions.

ii. Site recommendation surfers set a few links to other potentially.

iii. Page surfer's recommendation set some links related to web's topics is in any way related to the current page. For example, it recommends the pages under the same category of the current page. It can also recommend some pages with similar or related topics. [7]

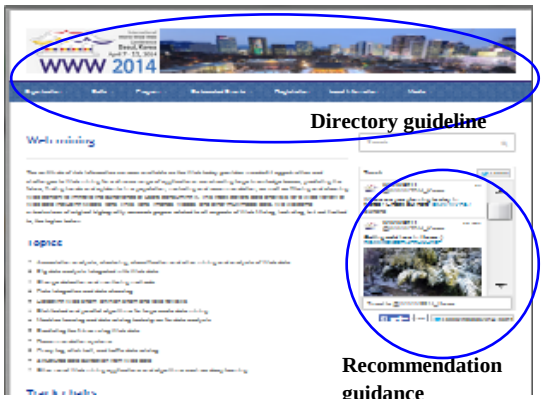

Figure 3. Example of Navigational Guidance Noise

3.5. Block Splitting

The rapid expansion of the Internet has made the World Wide Web a popular place for disseminating and collecting information. The innovation of the web creates numerous information sources published as HTML pages on the Internet. Search engines crawl the World Wide Web to collect web pages. These pages are stored and indexed. The user who is performing a search using search engine is interested in primary informative content of the web page data mining on the web thus becomes an important task for discovering useful information from the web. But a large part of these web pages is content that can not be classified as the primary informative content of the web page. Useful information on the web is often accompanied by advertisements, image-maps, plug-ins, logos, search boxes, category information, navigational links, related links, footers and headers, and copyright information. Although such information items are useful for human viewers and necessary for the web site owners, they often hamper automated information gathering and web data mining. These blocks are not relevant to the main content of the page. Such blocks are referred to as non-content blocks; these blocks are very common in web pages.

Figure 4(a). Sample Web Page containing Noise

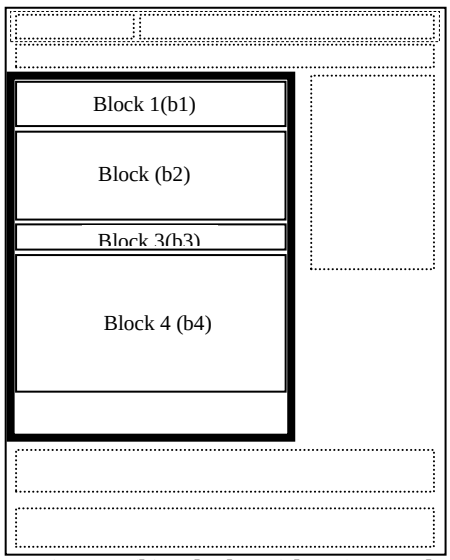

Figure 4(b). Blocks with noises and main contents

The Figure 4(a) is taken an example of a sample web page which consists of local noises such as images, multiple links, etc. and also the main content useful for mining. The dotted line represented in the Figure 4(b) is denoted as local noises. The useful main contents for web content mining are highlighted with dark lines. The main content has some sub-contents which are divided into blocks b1, b2, b3 and b4 using block splitting operation.

3.5.1 Block Splitting Algorithm

Step 1: Select the web page
Step 2: Identifying Local and Primary web page
Step 3: Find out Local Noise
Step 4: If Tag <DIV> or <TD> Then Read Content
Else Exit
Step 5: If identified <DIV> or <TD> tag then The store content in the each block values
Step 6: Continue this process up to end of the web page
Step 7: Exit

In Figure 4(a), an example of a sample web page is taken which consists of local noises such as images, multiple links, etc., and the main content useful for mining.

3.6 Duplicate Detection

Duplicate document or pages are not a problem in traditional IR. However, in the web, it is a significant issue. There are different types of pages and contents on the web. Copying a page is usually called duplication or replication, and entire site is called mirroring.

Several methods can be used to find duplicate information. On the web, one seldom finds extract duplicates. One efficient duplicate detection technique is based on n-grams. An n-gram is simply a consecutive sequence of words of a fixed window size n.

$$sim(d_1, d_2) = \frac{|S_n(d_1) \cap S_n(d_2)|}{|S_n(d_1) \cup S_n(d_2)|} \quad (1)$$

A threshold is used to determine whether d_1 and d_2 are likely to be duplicates of each other. [11]

3.7 Gomory-Hu tree

A *Gomory-Hu tree* (also known as a *cut tree*) is an O(n)-space data structure which represents the pair wise edge connectivity of all pairs of vertices in an undirected graph. More precisely, it is a weighted tree T on V, with the property that the pair wise edge connectivity between any two vertices s and t in the graph equals the minimum weight of an edge on the unique s-t path in T. Further, the partition of the vertices produced by removing this edge from T is a minimum s-t cut in the graph, i.e. a cut of cardinality equal to the s-t edge connectivity. An undirected graph has at least one Gomory-Hu tree, but it might not be unique; on the other hand, examples by Bencz´ur[Ben95] show that Gomory-Hu trees need not exist for directed graphs. Gomory-Hu trees have many applications in multi-terminal network flows.

Theoretically Gomory-Hu algorithm is an optimal algorithm. However, since it partitions the graph following the minimum cut criteria, it tends to generates outliers, very small sub graph or singleton vertices.

Given an undirected, weighted graph *G . .V; E; c.* a cut-tree
T . .V; F;w. is a tree with edge-set *F* and capacities *w* that fulfills the following properties.
1. Equivalent Flow Tree: For any pair of vertices *s; t* 2 *V*,
f .s; t. in *G* is equal to *fT .s; t.*.
2. Cut Property: A minimum *s-t* cut in *T* is also a minimum cut in *G*.

Here, *f .s; t.* is the value of a maximum *s-t* flow in *G*, and *fT .s; t.*is the corresponding value in *T*. [8]

4. Proposed System Design

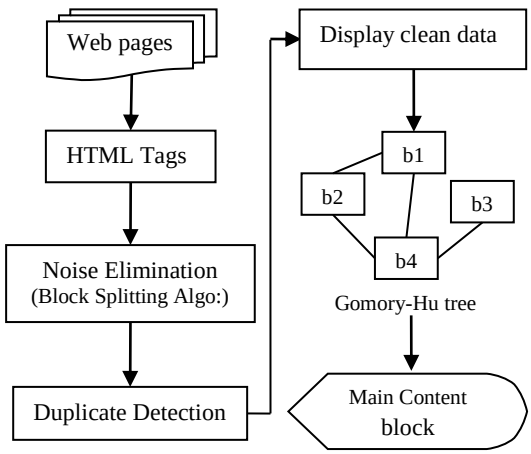

Figure 5. Main Content Block Extraction

The proposed system is shown in figure 5. In this system, noise tags of the web page are eliminated by using block splitting algorithm, detect duplicate blocks with equation (1) and achieve the clean data of the web page. Then, gomory-hu tree is constructed by using these data. Finally, the maximum flow of graph is assumed as the main content block of the web page.

5. System Performance

This system has been tested with full sites downloaded from following links.

Table1. Results for selected Web Site

URL	Precision (%)	Recall (%)
www.wikipedia.org	99	97
www.educause.edu	98	90
www.cbsnews.com	100	98
www.itap.purdue.edu	92.3	87.5

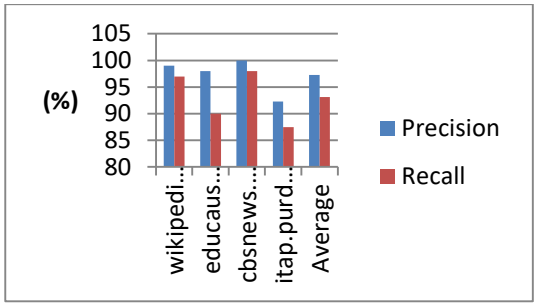

Figure 6. The precision and Recall Chart for selected web sites

6. Conclusion

To improve the performance of web content mining, an approach is proposed which removes noises from web pages. The irrelevant data considered as primary noises have been removed using block splitting operation. From the resultant blocks, the duplicate blocks are removed by computing the n-grams for each block; two parameters are computed such as Keyword Redundancy and Title word Relevancy for knowing the importance of each block. Then the noise blocks are removed by using the threshold value. After removing the noise blocks, the remaining blocks considered as important blocks are extracted using Gomory-Hu tree algorithm. The main objective for removing noise from a web page is to improve the performance of the search engine. It is very essential to differentiate important information from noisy content that may misguide users' interest.

References

[1] Bing Liu. Web Content Mining. The 14th International WWW Conference(WWW-2005), Chiba, Japan.

[2] Cai, D.,Yu, S., Wen, J.R., Ma, W.Y., "VIPS: A vision-based segmentation algorithm". 2003.

[3] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schutze, " Introduction to Information

Retrieval", Cambride. Uniersity Press, New York, USA, 2008, 2009.

[4] A. Arasu and H. Garcia-Molina. Extracting structured data from web pages. In p*roceedings of SIGMOD 2003*, 2003.

[5] Chao Wang, Jie Lua, and Guangquan Zhanga, "Mining Key Information of Web Pages: A Method and Its Application", Expert Systems with Applications, Vol.33, No.2, pp.425-433, August 2007.

[6] Deng C., Shipeng Y., Ji-Rong W., Wei-Ying M., "Extraction Content Structure for Web Pages based on Visual Representation", Microsoft Research Asia, China.

[7] P Sivakumar "Noise removal from web page PAGE", 2014

[8] Amit Chauhan, Himanshu Uniyal, Dr.Bhasker Pant, "Cleaining Web Pages for Relevant Text Extraction and Text Categorization", Graphic Era University, India.

[9] Deng C., Shipeng Y., Ji-Rong W., Wei-Ying M., "Block-based Web Search", Microsoft Research Asia, China.

[10] Swe Swe Nyein, "Mining Contents in Web Page Using Cosine Similarity", University of Computer Studies, Yangon, Myanmar.

[11] "Web Data Mining" , Springer, page -190.

[12] Lin, S.-H. And Ho, J.-M. 2002. Discovering informative content blocks from web documents. In Proceedings of the 8th ACM SIGKDD Knowledge Discovery and Data Mining, Edmonton, Canada.

[16] C. Li, J. Dong, and J. Chen, " Extraction of Informative Blocks from Web Pages Based on VIPS", January 2010.

[17] A. K. Tripathy and A. K. Singh, "An Efficient Method of Eliminating Noisy Information in Web Pages for Data Mining", In Proceedings of the Fourth International Conference on Computer and Information Technology (CIT'04), pp. 978 – 985, September 14-16, Wuhan, China, 2004.