Detecting and Removing Noisy Data on Web Document using Text Density Approach

Hassan F. Eldirdiery
Faculty of Computer Science and Information Technology
University of Science and Technology, Omdurman, Sudan

A. H. Ahmed
Faculty of Computer Science and Information Technology
Al-Neelain University, Khartoum, Sudan

ABSTRACT
The web documents content are useful resources for many applications. However, this content could be classified into relevant content and irrelevant content with respect to the involved application. The irrelevant content, like advertisements banner, copyright information, and navigation menus assumed as noisy data. Noisy data that found among the content of the web document affects negatively the performance of most of applications that deals with the content of web pages. The process of detecting and removing noisy data is an important pre-processing step in many applications such as web page classifications, clustering of web pages and information retrieval tasks. We developed a unified algorithm able to detect automatically the noisy data and eliminate them out of the web page and produce a clear web document that could be used effectively in later steps. The suggested approach examined using a dataset composed of different classes. The results of the conducted experiments showed a significant enhancement in the problem of detecting and removing noisy.

General Terms
Algorithms, Performance, Experimentation

Keywords
Web Page Segmentation, Noise Removal, Information Extraction

1. INTRODUCTION
The internet is an extraordinary resource that full of the information, which is strongly needed by many users and applications for different purposes. However, this resource is composed of web pages that contain different types of information and extra data which affects the process of extracting the intended information from the web. The difficulties of extracting the relevance content from the web come from unstructured formatting of the files where the intended data resides.

Most of the web documents contain extra information which is not relevant to the main article, but used for visual layout concerns among the desired contents. The redundant information in web document, like advertisements banner, copyright information and navigation menus assumed as noisy data for many applications. Gibson et al. [1] presented study showed that noisy information represents between 40% and 50% of data on the web and that this volume is growing at a rate of approximately 6% per year.

These noisy data considered harmful to many applications that related to the web, since it affects and reduces the performance of their process. There are many applications work to collect and manipulate the relevant data on web pages, for example web mining applications, which include classification and clustering of web pages, detecting of replicated web pages and information retrieval applications. Since there is a rapid growth in the applications that generate web pages, the research and approaches that used to detect and remove the noisy data out of web documents still be active researches [3, 12, 13].

The problem of detecting and removing noisy data out of web document has been addressed from different approaches [2]. Some authors solved the problem by developing methods based on the analysis of the underlying structure of DOM tree, which is representation structure of the web page [3]. While some other algorithms used to detect and remove the noisy data by depending solely on the visual layout of the web page [4]. Also there are some methods developed based on text properties included in the web page file [5]. However, most of the developed algorithms have some limitations from the side of performance and accuracy measurements according to the approach they followed. The process of detecting and removing noisy data out of the web page in most of the approaches composed of two phases: (1) web page segmentation and (2) detecting and removing noisy segments phase. In the phase of segmentation, the content of the web document partitioned into separate blocks or segments.

In the process of segmenting the web document, most of the developed algorithms used to rely either on DOM tree approach or vision approach. However, the methods based on DOM tree take a significant amount of time to build the tree besides that the number of possible DOM layout patterns is virtually infinite, which definitely leads to errors with respect to the growth of web documents. While one of the drawbacks of using vision approach is the complexity of segmentation process. That since the process of rendering the page needs an access to the browser engine. And the second drawback also the requirements for external resources such as CSS files and images. While with respect to other approaches, the field of linguistics with its rich text properties can be utilized well in order to enhance the process of segmentation as suggested in [5], which consequently enhance the process of detecting and removing noisy data out of web pages.

The paper presented a method which detect and eliminate the noisy data out of web document and provide a clear document valid to be used in the involved applications. The proposed algorithm based on text-based segmentation approach. The method first segments the web document into separated blocks. Then it investigates these blocks to detect the noisy blocks from not noisy ones. The algorithm focuses to remove the noisy data those almost composed of text.

One of the benefits of the proposed algorithm that it needs only one web page in the process of detecting the noisy data. While in some other approaches the process of detecting noisy

data requires more pages. Also no needs to convert the web document file into DOM tree representation.

The results of the conducted experiments reflect the importance of eliminating noisy data out of web documents. The algorithm examined with a dataset composed of different web pages in order to test the effectiveness of the clean web pages on the process of classification. The metrics recall and precision used with experiments to measure the accuracy of using the suggested method.

The rest of the paper organized as follows. Section 2 presents the related work and Section 3 provides the proposed method. Section 4 gives the experimental set up and results. Section 5 concludes the work.

2. RELATED WORK

The problem of cleaning the web document by extracting the noisy data out of the required pages has been addressed using many different approaches. Bar-Yossef et al. [6] used a technique based on that segments the DOM tree of the web page into pagelets. A pagelet is a self-contained logical region within a page that has a well defined topic or functionality. When specific element contains k links, it is likely represent an independent idea or topic. This method of segmentation relies on the tags of HTML. However, the tag almost contains attributes that represents the visual layout of the content and can be used to improve the process of segmentation. Yi et al. [3] proposed a method that removes noisy features in order to improve data mining results. They make use of structure tree (SST) which is similar to the structure of DOM tree, then the noisy nodes detected through a process of evaluating to the diversity presentation styles and contents of each node in SST. But process of building the desired tree takes a significant amount of time. Debnath et al. [7] proposed an approach similar to the one in [3]. They also select portions of web pages, called blocks, which have an importance level above a given threshold. They estimate the importance for each individual block. Song et al. [8] proposed a learning method that detects template among web pages. The method automatically assigns importance weights to hierarchically arranged segments in web pages, termed as blocks. The method requires blocks in sample web pages to be labeled by users based on their judgment of the importance of each block. Next, using the labeled blocks as training data, classifiers are used to automatically label other unseen pages. Template identification can be seen as a specific instance of this problem, since templates can be regarded as sets of blocks that have very low importance. However, the presented approach works to manually label specific blocks. Also Viera et al. [9] proposed a method that detect template by finding the optimal mappings between DOM trees of web pages. The method is based on finding a mapping between the underlying tree structures of web pages in order to detect identical nodes in the trees and sub trees that contain these nodes. They represent each HTML as a labeled ordered rooted tree that corresponds to its underlying DOM tree. When a given sub-tree that spans from the document root is detected in both input pages, this sub-tree regarded as a template. Once this sub tree is found, then it can be easily located and removed from other pages. But, not all the constructed web pages follow the same underlying tree structure. E. S. Laber et al. [10] proposed an approach that identify and extract relevant content from web pages those contain news. Their method defines relevant as the textual sections that more objectively describe the main event in the article. This includes the title and the main body section, and excludes comments about the story and presentation elements. However, the method works only on news web pages. Sharma and Bhatia [11] developed an algorithm that based on page replacement to remove noisy data from web page. The method uses page replacement of the last recently used strategy, where pages not used for long time will remain unused for a long of time. The results showed that the proposed method is less complex than DOM tree method. Dutta et al. [12] suggested an approach that uses the regular expression to remove the noisy tags out of the HTML file as a first step, and then constructs site style tree in the next step. Raheja and Katiyar [13] developed an algorithm that works through three phases. Phase 1 starts by developing a web page in the form of table composed of n rows and 1 column, then inserts the required data into the developed table as internal table. In this phase each internal table is marked with an ID attribute. Phase 2 converts the constructed table into XML format which later used in phase 3 with XSL to extract the data using the feature filter. Pappas et al.[14] suggested a method that make use of visual and non-visual characteristics of web page to remove noisy data in three major category of pages which contain user-generated content (News, Blogs, Discussion). The method utilizes DOM tree with Style-Density tree with two optimal thresholds in order to classify the type of the web page. While our suggested algorithm works on all types of web documents.

3. THE PROPOSED METHOD

The developed algorithm called BDBNE (Block Density-Based Noise Extractor), processes the target web page file and segments it into multiple blocks. Then it analyzes the extracted blocks in order to distinguish the noisy blocks from not noisy blocks using pre-computed threshold value. The algorithm works on the HTML file of the web page. It used the sequence of characters outside HTML tags to construct the blocks and ignores the sequence of characters inside HTML tags. The process of detecting noisy blocks based on the text density of the block. The text density is the number of words within a particular 2-dimentional area. Here the boundary of the text is the extracted block.

Some of the extracted blocks contain nonsense information such as blank characters and symbols. So the algorithm identifies these blocks as invalid blocks and excludes them before proceeding to the next step. The process of distinguishing the invalid blocks from valid blocks based on comparing the content of each block with a pattern of characters.

The value of the threshold, which used to differentiate the noisy blocks from not noisy blocks, is computed automatically through the processes of the algorithm. The phases of BDBNE represented in the following and the *pseudo code* of the main algorithm presented in Figure 2.

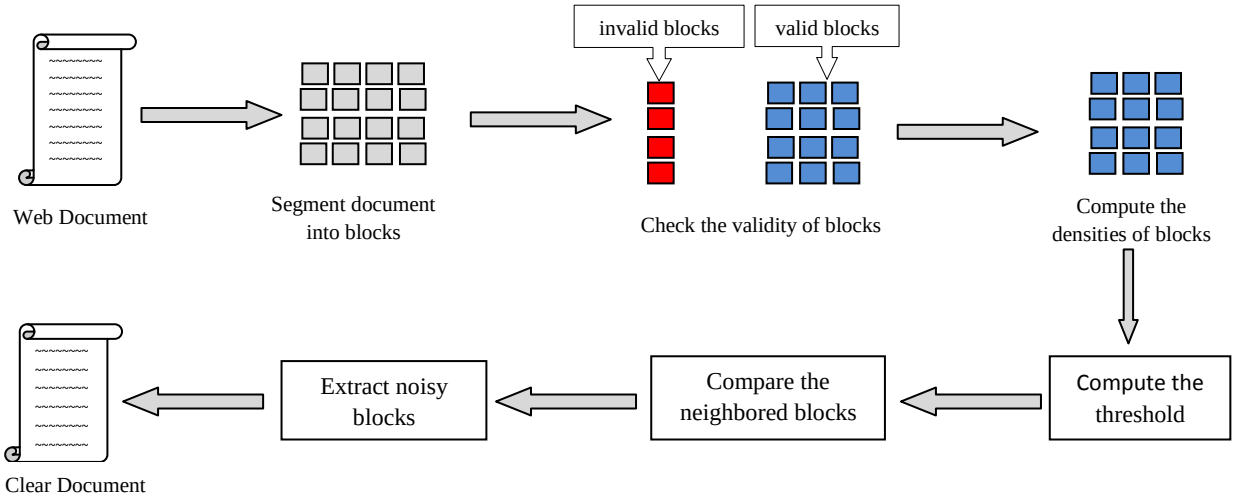

Figure 1: The phases of BDBNE

Pseudo Code of BDBNE algorithm

BDBNE Algorithm: Detect Noisy Blocks
Input: set of valid blocks B_i , i = 0, 1, 2, …, n
Input: threshold value t
Output: set of noisy blocks
for each block in the set B_i
  if density(B_i) != density(B_{i+1}) then
    $epsilon = \frac{|density\ (B_i) - density\ (B_{i+1})|}{max\ (density\ (B_i), density\ (B_{i+1}))}$
    if (epsilon > = t)
      if (lines(B_i) < = lines(B_{i+1}))
        add B_i to noisy blocks set
        add i to noisy indices set
      else
        add B_{i+1} to noisy blocks set
        add i+1 to noisy indices set
      end if
    else
      i++
  end if
next for
    return noisy indices set;

4. EXPERMENTAL SET UP

Many experiments were conducted to test and evaluate the proposed method effectiveness toward detecting and removing noisy blocks out of a web page and producing a clear document. The validity and accuracy of the proposed algorithm were checked using the measures recall, precision and F1 scores from the field of information retrieval. The dataset that used in the experiments composed of multiple pages from different web sites. The datasets were collected from the following web sites:

www.amazon.com: Library website

www.pcmag.com: PC Magazine for software, computers, hardware, news, reviews, and opinions

www.cnet.com: website that provides product reviews, prices, software downloads, and technology news

News.sudanvisiondaily.com: electronic newspaper

www.biomedcentral.com: Medicine Journal

5. EXPERMENTAL RESULTS

5.1 Testing Consistency

Our proposed algorithm validated against the consistency and accuracy measures. To evaluate consistency it means to test BDBNE if it has the ability to segment a web document into blocks and extract the noisy blocks from all blocks. The test started using a dataset of an electronic newspaper called Sudan Vision Daily. The data contains pages from four sections, news, reports, science and business. Figure 2 represents a sample of a web page from the dataset. In the figure we can see some rectangular with red boundaries around some parts of the page, these rectangular determine the noisy parts on the page. After running the algorithm, these parts will be detected as noisy blocks.

Figure 2: Web Page contains noisy data

The results of running the algorithm presented in Figure 3. From the figure we can notice that most content of the pages full of noisy parts. The average of noisy blocks represents more than 80% of the total blocks.

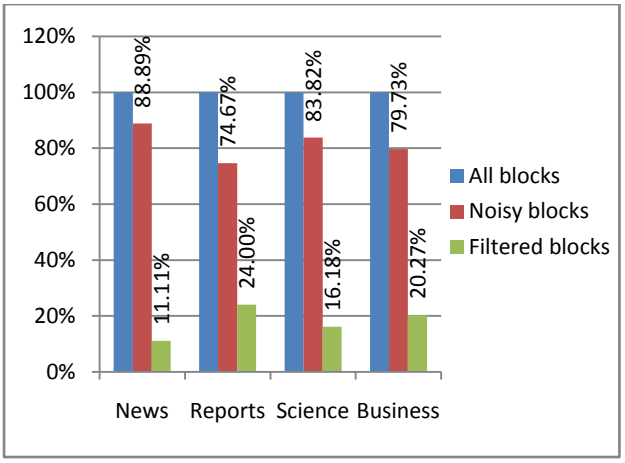

Figure 3: Chart represents the blocks of Sudan Vision Daily dataset

5.2 Testing Accuracy

This section evaluates the accuracy of the algorithm. The measures used are recall, precision and F1 score from the field of information retrieval. The algorithm segments the blocks into noisy and not noisy classes. The class of noisy blocks refers to it as invalid blocks and the class of clear blocks as valid blocks. The measure F_β is a measurement that combines recall and precision, which mostly computed with the value of $\beta = 1$and then called F-1 score. Three different pages from the dataset of BMC journal were selected to test the accuracy of BDBNE in the process of detecting invalid blocks from valid blocks. And Table 1 represents the results of precision, recall and F1-Score.

Table 1: Precision, Recall and F1-Score results

Web Page	Precision	Recall	F-1 Score
Page1	85.23%	94.94%	89.82%
Page2	96.42%	92.02%	94.17%
Page3	93.27%	92.82%	93.05%

5.3 The Impact on Classification

To evaluate the impact of BDBNE on classification, a dataset composed of different web documents were collected from Amazon, PCMag and CNet web sites. The sites contain many pages about different kinds of products. The web sites contain many classes of products, such as Camera, Laptop, Printer and TV. F-1 score used to measure the results of classification of the dataset after cleaning the web pages from noisy data. To implement the classification, the experiment used Support Vector Machine from the field of machine learning.

The experiment started by building a classifier that based on one class of pages from different web sites. For example, the class can be laptop from Amazon, PCMag and CNet. The dataset of the chosen class partitioned into two sets, one for training and the other for testing. The training dataset of the intended class is used to build the classifier. Then the classifier is used to classify the testing dataset of the intended class. The process of classification is done for four different classes. Table 2 shows the results of classification after cleaning the web pages from noisy data. The results proved that our suggested algorithm positively affects the performance of classifier.

Table 2: F-1 Scores and accuracies after cleaning web pages

Class	F-1 Score	Accuracy
Camera	0.997	0.998
Laptop	0.988	0.989
Printer	0.989	0.989
TV	0.997	0.998

6. CONCLUSIONS AND FUTRE WORK

The paper presented a method that detect and remove the noisy data from web document. The suggested algorithm based on the text density approach. The implementation of the technique proved that the textual content of the web document could be used properly to clean the web page from noisy data. Our method simply works only on a single web page. So it positively affects the storage and speed factors of the running algorithms. The conducted experiments evaluated the effectiveness of the algorithm and the results showed a significant improvement in the process of detecting and removing noisy data out of web document. For future work, we will work to improve and refine the algorithm by combining other rich data types, such as video, audio and graphics.

7. REFERENCES

[1] David Gibson, Kunal Punera, and Andrew Tomkins. 2005. The Volume and Evolution of Web Page Templates, IW3C2, ACM Press, pp. 830-839.

[2] K. Viera, A. S. da Silva, N. Pinto, Edleno S. de Moura, J. M. B. Cavalcanti, J. Freire. 2006. A Fast and Robust Method for Web Page Template Detection and Removal. In Proceedings of the International Conference on Information and Knowledge Management, ACM Press, pp. 258-267.

[3] Lan Yi, Bing Liu, and Xiaoli Li. 2003. Eliminating Noisy Information in Web Pages for Data Mining, SIGKDD, ACM Press, pp. 296-305.

[4] E. Akpinar and Y. Yesilada. 2012. Vision based page segmentation: Extended and improved algorithm. eMINE Technical Report Deliverable 2 (D2), Middle East Technical University, Ankara, Turkey.

[5] C. Kohlschutter, W. Nejdl. 2008. A Densitometric Approach to Web Page Segmentation. In Proceeding of the 17th ACM conference on Information and knowledge management, ACM Press, pp. 1173-1182.

[6] Z. Bar-Yossef and S. Rajagopalan. 2002. Template detection via data mining and its applications. In proceedings of the International Conference on the World Wide Web, ACM Press, pp. 580-591.

[7] S. Debnath, P. Mitra, and C.L. Giles. 2005. Automatic Extraction of Informative Blocks from Web pages. In ACM Symposium on Applied Computing, pp. 1722-1726.

[8] R. Song, H. Lui, J.-R. Wen, and W.-Y. Ma. 2004. Learning block importance models for web pages. In proceedings of the International Conference on the World Wide Web, ACM Press, pp. 203-211.

[9] K. Viera, A. S. da Silva, N. Pinto, Edleno S. de Moura, J. M. B. Cavalcanti, J. Freire. 2006. A Fast and Robust Method for Web Page Template Detection and Removal. In Proceedings of the International Conference on Information and Knowledge Management, ACM Press, pp. 258-267.

[10] E. S. Laber, C. Souza, I. Jbour, E. Amorim, E. Cardoso, R. Renteria, L. Tinoco, C. Dias. 2009. Fast and Simple Method for Extracting Relevant Content from News Webpages. In Proceedings of the ACM Conference on Information and Knowledge Management, ACM Press, pp. 1685-1688.

[11] R. Sharma and M. Bhatia. 2014. Eliminating the Noise from Web Pages using Page Replacement Algorithm. International Journal of Computer Science and Information Technologies, Vol. 5 (3) , IJCSIT, pp. 3066-3068.

[12] Amit Dutta, Sudipta Paria,Tanmoy Golui, Dipak Kumar Kole. 2014. Noise Elimination from Web Page Based on Regular Expressions for Web Content Mining. Advanced Computing, Networking and Informatics, Vol. 1, Springer.

[13] Neeraj Raheja and V.K.Katiyar. 2013. Noise Reduction Approach based on n x 1 Table and XSL Display Method for Efficient Web Data Extraction. International Journal of Computer Applications, Vol. 64.

[14] N. Pappas, G. Katsimpras and E. Stamatatos. 2012. Extracting Informative Textual Parts from Web Pages Containing User-Generated Content. In Proceedings of the 12th International Conference on Knowledge Management and Knowledge Technologies, Article No. 4, ACM.