BOILERPLATE REMOVAL AND CONTENT EXTRACTION FROM DYNAMIC WEB PAGES

Pan Ei San

University of Computer Studied, Yangon

ABSTRACT

Web pages not only contain main content, but also other elements such as navigation panels, advertisements and links to related documents. To ensure the high quality of web page, a good boilerplate removal algorithm is needed to extract only the relevant contents from web page. Main textual contents are just included in HTML source code which makes up the files. The goal of content extraction or boilerplate detection is to separate the main content from navigation chrome, advertising blocks, and copyright notices in web pages. The system removes boilerplate and extracts main content. In this system, there are two phases: Feature Extraction phase and Clustering phase. The system classifies the noise or content from HTML web page. Content Extraction algorithm describes to get high performance without parsing DOM trees. After observation the HTML tags, one line may not contain a piece of complete information and long texts are distributed in close lines, this system uses ***Line-Block concept*** *to determine the distance of any two neighbor lines with text and* ***Feature Extraction*** *such as text-to-tag ratio (TTR), anchor text-to-text ratio (ATTR) and new content feature as Title Keywords Density (TKD) classifies noise or content. After extracting the features, the system uses these features as parameters in threshold method to classify the block are content or non- content.*

KEYWORDS:

Content, line-block, feature, extraction

1. INTRODUCTION

Today, the internet matures, thus the amount of data available continues to increase. The artifacts of this ever-growth media provide interesting new research opportunities that explore social interactions, language, art, and politics and so on. In order to effectively manage this ever-growing and ever-changing media, content extraction methods have been developed to remove extraneous information from web pages. Extracting useful or relevant information from Web pages thus becomes an important task. Also irrelevant information is contained in these Web pages. A lot of researches on WWW need the main contents of web pages to be gathered and processed efficiently. Web page content extraction technology is a critical step in many technologies. Content Extraction (CE) is just the technique to clean the documents from extraneous information and to extract the main contents.

Nowadays, web pages become much more complex than before, so CE becomes more difficult and nontrivial. Template based algorithms and template detection algorithms also perform poorly because of web page's structure being changed more frequently and web page's being generated

DOI : 10.5121/ijcsea.2014.4603

dynamically. Traditionally, Document Object Model (DOM) based algorithms and vision based algorithms may get better results but they always consume a lot of computing resource. Parsing DOM tree is a time consuming task. Vision based algorithms need to imitate browsers to render HTML documents, which will consume much more time. This system is implemented to remove noises or boilerplate based on Line-Block concept, content features and uses the threshold method to classify whether the block is content or not.

Usually, apart from the main content blocks, web pages usually have such blocks as navigation bars, copyright and privacy notices, relevant hyperlinks, and advertisements, which are called noisy blocks. Modern web pages have largely abandoned the use of structural tags within a web page and adopted an architecture which makes use of style sheets and or tags for structural information. Most current content extraction techniques make use of particular HTML cues such as tables, fonts, size and line, etc., and since modern web pages no longer include these cues, many content extraction algorithms have begun to perform poorly. One difference between our approach and other related work is that no assumption about the particular structure of a given webpage, nor does look for particular HTML cues. In our approach, the system uses the Line-Block concept to improve preprocessing step. And then, the system calculates the content features as Text-to-Tag Ratio (TTR), Anchor-Text to Text Ratio and the new feature Title Keyword Density (TKD). This state is called featured extraction phase. After feature extraction, the system use this features' values to classify the block is content or not by using threshold method. The system's objectives are followed:

- To develop a web content extraction method that given an arbitrary HTML document
- To extract the main content and discard all the noisy content
- To get high performance of noises detection without parsing DOM trees
- To decrease the consuming time of preprocessing step such as noise detection and classification of blocks.
- To enhance accuracy of information retrieval of Web data

Contributions: Four main contributions can be claimed in our paper:

1. To propose Extended Content Extraction algorithm that contains line block concepts, boilerplate detection and extraction of main content block
2. To reduce the web page's preprocessing time that used Line-Block concept.
3. To reduce the loss of important data by adding the new feature Title Keyword Density (TKD)
4. To retrieve the more important blocks that use threshold method.

The paper is structured as follows. In this paper, we discuss background theory in section 2. Next, in section 3 we describe our proposed system in detail. In Section 4 we give our evaluation and experiments other CE algorithms. Finally we offer the conclusion with a discussion of further work in Section 5.

2. BACKGROUND THEORY

There are non-informative parts outside of the main content of a web page. Navigation menus or advertisement are easily recognized as boilerplate, for some other elements it may be difficult to decide whether they are boilerplate or not in the sense of the previous definition. The **CleanEval guidelines** instruct to remove boilerplate types such as [1] Navigation, lists of links, Copyright notices, template material, such as header and footers, Advertisements ,Web spam, such as automated postings by spammers, Forms and Duplicate material, such as quotes of the previous posts in a discussion forum

2.1. Related Concept of Line

In this section, we describe the some concepts about the line of HTML source documents and content-feature that we use in our system.

A. Line

A HTML tag is a continuous sequence of characters surrounded by angle brackets like <html> and <a>. Hyperlink is one tag of HTML tag set. A complete Hyperlink tag has two markups: <a> as the open tag and as the close tag. A line is a HTML source code sequence from original HTML documents with texts and complete HTML tags (especially Hyperlinks tags). Anchor text of a line is the text between hyperlink tag's opening tag '<a>' and closing tags ''. Text of a line is the plain text of a line. It is all the continuous sequence characters between angle brackets '>' and '<'. If a line has no angle brackets, then all characters in this line are text of this line.

B. Define Line-Block

Line-Block is a line or some continuous lines, in which the distance of any two neighbor lines with text. Block means that it defines between open tag <>and end tag</>. In this paper, we define and use the important block tags as p, div, h1, h2 and so on.

C. Content Features

Definition (Text-to-tag ratio (TTR): TTR is the ratio of the text length in the block is divided by the total sum of tags in this block.

$$TTR = \frac{text.length}{sum(tag)} \qquad (1)$$

Definition (Anchor text-to-text ratio (ATTR): ATTR is the ratio of the length of the anchor text is divided by the text length in the block.

$$ATTR = \frac{anchortext.length}{text.length} \qquad (2)$$

Definition (Title Keyword Density (TKD)): A web page title is the name or heading of a Web Site or a Web Page. If there is more number if title words in a certain block, then it means that the corresponding block is of more importance.

$$TKD_k = 1 - \left| \frac{m_k}{\left\lceil m_k + \sum_{i=1}^{|m_k|} F(m_k) \right\rceil} \right| \qquad (3)$$

Where, m_k is Number of Title keywords and F (m_k) means Frequency of title keyword m_k in the block.

2.2. Selecting Content from Threshold method

Finally, we get three feature values for each line block and need the best the parameters as thresholds to remove the noise block. It is increasing of precision and a sharp decrease of recall. τ is threshold. τ=λ σ, where λ is constant parameter and σ is standard deviation. Following steps are to find the mean value, find the variance and find the standard deviation for calculates to get σ. Here, different web pages may have different kinds of content, so if we set the thresholds as constants it will lead to skew determinations. We analyze the TTR threshold as 30, ATTR's threshold as 0.2 and TKD' threshold as 2.

3. PROPOSED SYSTEM

In our proposed system include the four steps. They are defined as follows:

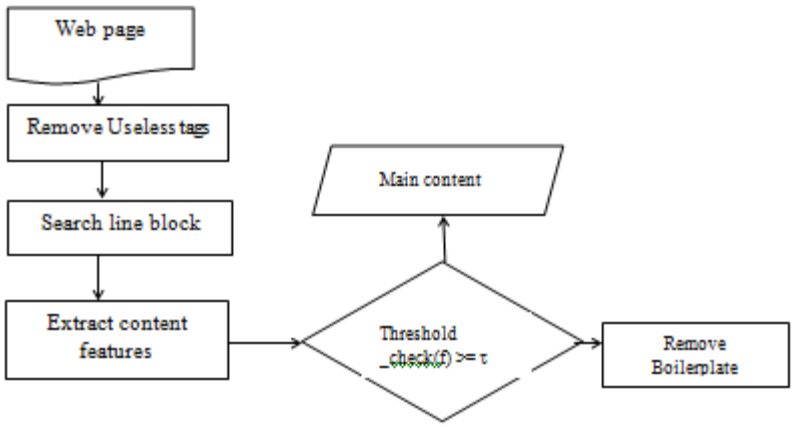

Figure1. Proposed system for content extraction

Step1. Preprocessing the web page tags

The tags filtered in this step, contains <head>, <script>, <style>, Remark and so on.

Step2.Define Line-Block

Line-block is a line or some continuous lines, which the distance of any two neighbor lines with text. The system reads line and makes the block using line-block concept. By sampling merging the lines, the system gets the line-blocks.

Step3. Feature Extraction

Next, the system calculates features for each block to determine whether they are content or not. TTR and ATTR are calculated as their formulas. For TKD, the system uses the title keywords in a block. Title Keyword Density (TKD) calculates to solve the loss of important information. There may be possibility that tag with less density also has the some important information. To remedy this the system a list of keywords from the title of the page and check if keyword density is greater than the threshold then the system add it the output block.

Step 4: Clustering Main contents or not

After calculating the content features, the system determines whether the block is content or not based on these features values. In this step, the system uses threshold methods to classify the main content or non -content and analyze the results. The threshold method uses standard derivation method. Threshold methods use three thresholds for TTR, ATTR and TKD. If TTR>TTR's threshold and ATTR<ATTR's threshold and TKD>=TKD's threshold then the block is main content. Otherwise, the block is noise block. Finally, the system extracts more accurately main contents.

3.1. Proposed Algorithm

- Input: D
- Output: mC
  DF ← filter_useless_tags (D)
  DB ← break_orginal_lines(DF)
  DL ← get_lines(DB)
- LB ← get_line_blocks(DL)
- **For** all block **in** LB **do**
- f ← get_feature (block)
- **If** threshold_check(f) >= τ **then**
- mC.append (block.text)
- **End for**

4. EXPERIMENTAL RESULTS

In this paper proposes a new content extraction algorithm. It differentiates noisy blocks and main content blocks. We present here the experimental results to testify the effect of algorithm. In many web pages, so many links the main content that they can produce enough noise. At the same time, so many links in the text reduces the weight of the text, but Title Keyword density (TKD) effectively supplements the weight of main text. In this example the original web page has 48.4 KB in figure 2 to reduce when removing the boilerplate blocks; the testing page has only 8.82 KB in figure 3. So, our proposed system can be reduced the storage space than original file size.

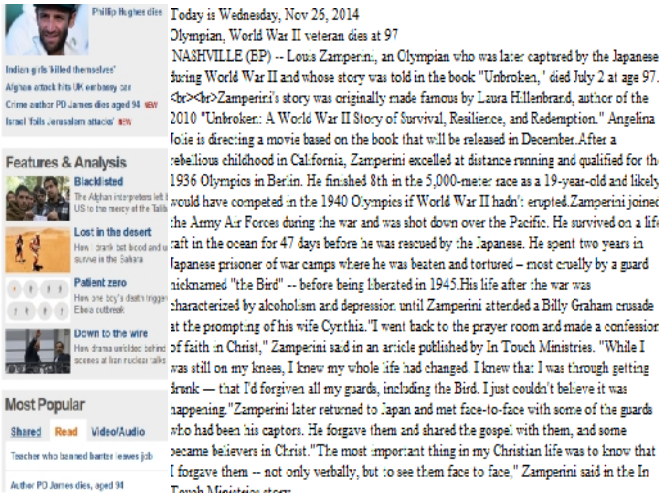

Figure2. Original Web page Figure3. Content Result

4.1. Data sets

The test data sets we use are from development and evaluation data sets from the CleanEval competition. They both hand-labeled gold standard set of main contents files; the amount of documents in each source are total of 606 web pages. In this dataset contains the following web site as BBC, nytimes and so on. CleanEval [1] is a shared task and competitive evaluation on the topic of cleaning arbitrary web pages. Besides extracting content, the original CleanEval competition also asked participants to annotate the structure of the web pages: identify lists, paragraphs and headers. In this paper, we just focus on extracting content from arbitrary web pages and use the 'Final dataset'. It is a diverse data set, only a few pages are used from each site, and the sites use various styles and structure.

5. CONCLUSION

The structures of webpages become more complex and the amount of data to be processed is very large, so Content Extraction (CE) remains a hot topic. We propose a simple, fast and accurate CE method. We do not parse the DOM trees to get a high performance. We can get the main contents from HTML documents and research can be done on the original files, which widens the direction of CE research. However, our approach uses some parameters and depends on the logic lines of HTML source code. In the future work, we continue to classify the web page and search engine for information retrieval.

REFERENCES

[1] M.Baroni, S.Sharoff https://cleaneval.sigwac.org.uk/annotation_guidelines.html, Jan 2007
[2] S.Gupta, G.E. Kaiser, D.Neistadt, and P.Grimm. Dom-based content extraction of html documents. In WWW, pages 207-214,2003
[3] S.Gupta, G.E. Kaiser and S.J.Stolfo. Extracting context to improve accuracy for html content extraction. In WWW (special interest tracks and posters), pages-1114-1115, ACM, 2005.
[4] S.Gupta, G.E.Kaiser, P.Grimm, M.F.Chiang, and J.Starren. Automating content extraction of html documents. World Wide Web, 8(2):179-224, 2005.
[5] T.Weninger, W.H.Hsu and J.Han. CETR-Content Extraction via tag ratios. In proceedings of WWW'10, pages 971-980, New York, NY, USA, 2010.

AUTHOR

Pan Ei San is working in University of Computer Studies. She has participated and presented two papers in national conferences. She is working in the area of web content mining and web classification.