MAIN CONTENT EXTRACTION FROM DYNAMIC WEB PAGES

PAN EI SAN ,NILAR AYE

University of computer studies (Yangon)
E-mail: paneisan1985@gmail.com

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 Extraction, Line-Block, TKD, TTR, ATTR

I. 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 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 d threshold method to classify whether the block is content or not.

II. MOTIVATION

For human, the behavior can be done relatively fast and accurate because they can use their knowledge, visual representation and layout of the web pages to distinguish the main content from other parts. WWW rapidly grow as it is accessible for public use through the web browser. Typically, a modern web document comprises of different kinds of content. Elimination of noisy and irrelevant contents from web pages has many applications,

- web page classification, clustering, web featuring,
- proper indexing of search engines,
- efficient focused crawlers,
- Cell phones and PDA browsing.

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. After shortly reviewing related work in Section III, we discuss background theory in section IV. Next, in section V we describe our proposed system in detail. In Section VI we give our evaluation and experiments other CE algorithms. Finally we offer the conclusion with a discussion of further work in Section VII.

III. RELATED WORKS

Istvan Endredy, Attila Novak presented an automatic text extraction procedure, GoldMiner, which by enhancing a previously published boilerplate removal algorithm, minimizes the occurrence of irrelevant duplicated content in corpora, and keeps the text more coherent than previous tools. The algorithm exploits similarities in the HTML structure of pages coming from the same domain. A new evaluation document set CleanPortalEval) is also presented, which can demonstrate the power of boilerplate removal algorithms for web portal pages.

C. Kohlschiitter, P. Fankhauser, and W. Nejdl proposed a merit of the boilerpipe algorithm is that its authors demonstrated experimentally that boilerplate content can be identified effectively by using a good combination of simple text properties. They used an annotated training corpus of 500 documents (mainly Google news) to find the most effective feature combination. They tried to extract articles with the help of shallow text features, using 8-10 different feature combinations, and then they evaluated their results. In their experiments, a combination of word and link density features gave the best results (its F-measure was: 92%). Furthermore, the method is very fast and it needs no preprocessing. Both the training set and the tool can be downloaded.

D.Cai, S.Yu, J.-R.Wen, and W.-Y. Ma proposed a new web content structure analysis based on visual representation is proposed in this paper. Many web applications such as information retrieval, information extraction and automatic page adaptation can benefit from this structure. This paper presented an automatic top-down, tag-tree independent approach to detect web content structure. It simulates how a user understands web layout structure based on his visual perception. Comparing to other existing techniques, our approach is independent to underlying documentation representation such as HTML and works well even when the HTML structure is far different from layout structure. Experiments show satisfactory results. In this paper proposed VIPS (Vision-based Page Segmentation) algorithm to extract the semantic structure for a web page. Such semantic structure is a hierarchical structure in which each node will correspond to a block. Each node will be assigned a value (Degree of Coherence) to indicate how coherent of the content in the block based on visual perception. The VIPS algorithm makes full use of page layout feature: it first extracts all the suitable blocks from the html DOM tree, and then it tries to find the separators between these extracted blocks. Here, separators denote the horizontal or vertical lines in a web page that visually cross with no blocks. Finally, based on these separators, the semantic structure for the web page is constructed. VIPS algorithm employs a top-down approach, which is very effective.

Xin Qi andJian Peng Sun, proposed a novel method to deal with Web page noise. Since no training data set and artificial annotation, this technique had a very wide of versatility. The noise elimination technique based on the following observation: a web page contains many information block, while the topic blocks usually have the characteristics of aggregate. Based on this observation, we propose a tree structure, called satisfiable sub-tree. The process of eliminating noisy information translates into finding the satisfiable sub-tree. Their proposed method is evaluated with several portals website.

J. Pomikalek proposed the jusText algorithm splits HTML content into paragraphs at block-level tags that are generally used to partition HTML content into logical units, such as <p>, <td>, <h1> etc. Using various features of these blocks of text such as the number of links (an idea from boilerpipe), words and

stopwords, the algorithm performs a rule-based classification of the blocks using various thresholds and a language-dependent list of function words tagging each unit 'good', 'almost good', 'bad', or 'too short'. The latter tag applies to units too short to categorize reliably. After initial classification, 'almost good' and 'too short' units surrounded by 'good' ones are reclassified as 'good'. The text to be extracted consists of all units classified as 'good' in the final classification. The algorithm performs quite well even for extreme pages. However, inspection of the corpus generated by using the jusText algorithm to filter crawled news portals revealed that many expressions that obviously come from a single article and should not occur more than once, like The feeding-bottle is a potential source of hazard, were still extremely strongly over-represented.

IV. 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

- Navigation
- Lists of links
- Copyright notices
- Template material, such as header and footers
- Advertisements
- Web spam, such as automated postings by spammers
- Forms
- Duplicate material, such as quotes of the previous posts in a discussion forum

A. 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.

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.

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.

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.

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

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.

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

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. The text extracted from between <title> and </title> tag to count the occurrence of title keyword density (TKD). For example: Title= "Bladder cancer: Exciting drug break" is tokened to {bladder, cancer, exciting, drug, break}.

B. 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. $\tau=\lambda\ \sigma$, 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 assign the TTR threshold as 30 and ATTR's threshold as 0.2 and TKD' threshold as 2.

V. PROPOSED SYSTEM

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

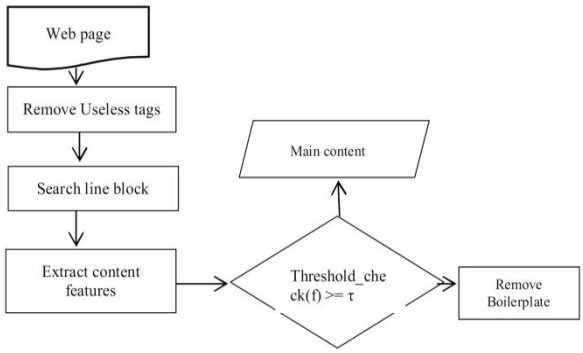


Figure1. Proposed system for content extraction

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.

C. 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

VI. EXPERIMENTAL RESULTS

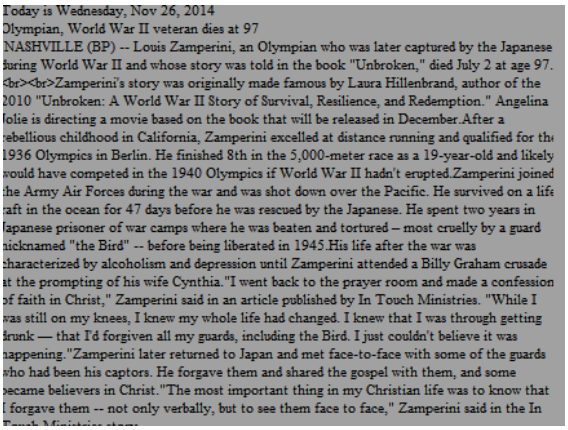
In this paper proposes a new content extraction algorithm. It differentiates noisy blocks and main content blocks. We present here the experimental results to testify the effect of algorithm. In many web pages, so many links the main content that they can produce enough noise. At the same time, so many 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 to reduce when removing the boilerplate blocks; the testing page has only 8.82 KB. So, our proposed system can be reduced the storage space than original file size.

Table1. File sizes for 736 original HTML files compared to the sizes for the extracted text

	HTML	Extracted Text
File Size (KB)	9,231	2288

We are also able to show a space savings of 70% when the extracted text is compared to the original HTML. Table 1 shows the result of savings.

Figure2. Original Web page


Today is Wednesday, Nov 26, 2014
Olympian, World War II veteran dies at 97
NASHVILLE (BP) -- Louis Zamperini, an Olympian who was later captured by the Japanese during World War II and whose story was told in the book "Unbroken," died July 2 at age 97.

Zamperini's story was originally made famous by Laura Hillenbrand, author of the 2010 "Unbroken: A World War II Story of Survival, Resilience, and Redemption." Angelina Jolie is directing a movie based on the book that will be released in December.After a rebellious childhood in California, Zamperini excelled at distance running and qualified for the 1936 Olympics in Berlin. He finished 8th in the 5,000-meter race as a 19-year-old and likely would have competed in the 1940 Olympics if World War II hadn't erupted.Zamperini joined the Army Air Forces during the war and was shot down over the Pacific. He survived on a life raft in the ocean for 47 days before he was rescued by the Japanese. He spent two years in Japanese prisoner of war camps where he was beaten and tortured -- most cruelly by a guard nicknamed "the Bird" -- before being liberated in 1945.His life after the war was characterized by alcoholism and depression until Zamperini attended a Billy Graham crusade at the prompting of his wife Cynthia."I went back to the prayer room and made a confession of faith in Christ," Zamperini said in an article published by In Touch Ministries. "While I was still on my knees, I knew my whole life had changed. I knew that I was through getting drunk — that I'd forgiven all my guards, including the Bird. I just couldn't believe it was happening."Zamperini later returned to Japan and met face-to-face with some of the guards who had been his captors. He forgave them and shared the gospel with them, and some became believers in Christ."The most important thing in my Christian life was to know that I forgave them -- not only verbally, but to see them face to face," Zamperini said in the In

Figure3. Content Result

D. 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 736 web pages. In this dataset contains the following web site as BBC, nytimes and so on. CleanEval 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.

E. Evaluation of Content Extraction Algorithm

As our aim is to find parameter values for CE algorithms in order to obtain best extraction results, we need to describe first how to actually measure and evaluate the extraction performance. The entire evaluation is based on a set of test documents, for which we provide a gold standard for the actual main content. The test documents are fed into the CE algorithms and their output is compared to the gold standard. The comparison can essentially be based on the texts in the computed extract and the gold standard. For each single test document, its text is represented as a sequence of words and the overlap between computed an actual main content is defined to be the longest common (not necessarily continuous) subsequence of words. Using the number of words in the computed extract e, in the gold standard g and in their overlap e∩g= lcss (e,g). We can define the common IR measures recall (r), precision (p) and the F1 measure (f1) for CE evaluation.

Example

Original text ="Title Copyright Some text in the body Commercial"

Gold Standard="Title Some text in the body"

CE algorithm provided="Title Copyright Some text in"

Common subsequence="Title Some Text in"

Recall r=4/6

Precision p=4/5

F1 f1=8/11

Table 2. Experimental Result

Web Sites	Page Number	Right Number	Wrong Number	Accuracy
BBC	150	145	5	96.7%
nytime	150	143	7	95.3%
nypost	150	142	8	94.6%
suntimes	150	140	10	93.3%
techweb	136	127	9	93.3%

The biggest advantage of this evaluation approach-beside its objective measures-is that it can be completely automated. Once the gold standard is defined, the evaluation process can be run without any human interaction. This is a key feature for the optimization via an evolutionary approach.

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 plan, we continue to classify the web page and search engine for information retrieval.

REFERENCES

[1] X.Qi and J. Sun,"Eliminating Noisy information in Webpage through Heuristic Rules", International Conference on Computer Science and Information Technology (ICCSIT 2011), IPCSIT vol. 51,page 137_141. Singapore, 2011.

[2] M.Baroni,S.Sharoff, "https://cleaneval.sigwac.org.uk/annotation_guidelines.html ", Jan ,2007.

[3] I.Endredy, A.Novak , "More Effective Boilerplate Removal-the GoldMiner Algorithm", ISSN 1870-9044; pp. 79–83,Polibits (48) 2013.

[4] J. Pomikalek, “Removing boilerplate and duplicate content from web´corpora [online],” Ph.D. dissertation, Masarykova univerzita, Fakulta informatiky, 2011.

[5] D.Cai, S.Yu, J.-R.Wen, and W.-Y. Ma, "VIPS: a vision-based page segmentation algorithm", Microsoft Technical Report, MSR-TR-2003-79,2003.

[6] C. Kohlschiitter, P. Fankhauser, and W. Nejdl, “Boilerplate detection using shallow text features,” in Proceedings of the third ACM international conference on Web search and data mining, pp. 441-450, ser. WSDM ’10. New York, NY,USA: ACM, 2010,. [Online].

★★★