Thanadechteemapat, W. and Fung, C.C. (2012) *Improving Webpage Content Extraction by extending a novel single page extraction approach: A case study with Thai websites.* In: International Conference on Machine Learning and Cybernetics, ICMLC 2012, 15 - 17 July, Xian, Shaanxi.

IMPROVING WEBPAGE CONTENT EXTRACTION BY EXTENDING A NOVEL SINGLE PAGE EXTRACTION APPROACH: A CASE STUDY WITH THAI WEBSITES

WIGRAI THANADECHTEEMAPAT, CHUN CHE FUNG

School of Information Technology, Murdoch University, Western Australia, 6150
E-MAIL: W.Thanadechteemapat@murdoch.edu.au, L.Fung@murdoch.edu.au

Abstract:

Web Content Extraction technique is proposed in this paper. The technique is able to work with both single and multiple pages based on heuristic rules. An Extracted Content Matching (ECM) technique is proposed in the multiple page extraction to identify the noises among the extracted results. Some features in this technique are also introduced in order to reduce processing time such as use of XPath, file compression, and parallel processing. Assessment of the performance is based on precision, recall and F-measure by using the length of extracted content. Initial results by comparing results from the proposed approach to extraction by manual process are good.

Keywords:

Web Content Extraction; Extracted Content Matching (ECM); XPath

1. Introduction

Information resources available on the Internet have been drastically increased over the last decade. The expansion of the Internet could be seen from a number of indicators. The first one is the number of Internet users, which has increased more than 480% to approximately two billions in the last eleven years [1]. The number of websites is another indicator, which has also increased by approximately 10 times to roughly 357 millions in the last ten years [2]. This huge repository of information has led to an issue known as *information overload*. There are many researchers attempting to address this issue by proposing various web information processing approaches such as Web content mining, Web clustering, Web classification, text summarization, and Web visualization.

A website is an electronic entity that provides one or more related webpages, while a webpage is an electronic document presented to the users through a web browser. Depending on the design, the content in the document may be blended with HTML tags and some programming scripts [3] so as to format and present the webpage through the browser. Moreover, in addition to the key informative content on a webpage, there are also non informative content such as advertisement, navigation words, header, footers and other objects on the same webpage. These types of non informative content can be considered as *noise* in the context of web information processing. In order to extract informative content, HTML tags, programming scripts and other forms of non informative content have to be filtered out.

One of the essential tasks for web information processing is *Web Content Extraction*, which deals with extracting only the informative content from a webpage, prior to be presented to the subsequent tasks of information processing. Many techniques have been proposed in Web Content Extraction. It appears that most of the reported techniques have to work with multiple webpages in order to detect the template used in the webpages, and not many techniques are able to work with only single page as the accuracy is compromised by the existence of noise.

This paper proposes a Web Content Extraction technique that is capable to work with both single and multiple pages. The proposed techniques are sub tasks of a Thai Web visualization research project and the objective of the project is to address the issue of information overload by providing an overview of Thai webpage(s) to the users. The Web Content Extraction is an important component of this project. The examples and results presented in this paper used using Thai websites as a case study, and the Web Content Extraction is mainly described in this paper.

The structure of this paper starts with an introduction and the aims of this research. Section 2 provides a summary of other related works, and the proposed techniques of Web Content Extraction are outlined in Section 3. The initial experimental results are reported in Section 4, and Section 5 presents the conclusion and discussion on future work.

2. Related Work

With respect to input of Web Content Extraction process, it could be grouped into two types: single and multiple pages. Some researchers proposed techniques [4] [5], which require multiple pages for the extraction process. Only some techniques [6] [7] are able to work on single page. In addition, some works proposed are only aimed to extract specific Web content such as shopping data, news.

"Automatic Identification of Informative Sections of Web Pages" was proposed by Debnath, et al [4]. Initially, webpage blocks are separated from the webpages based on rules. The blocks are then identified as either content blocks or non-content blocks with four algorithms. The block features have to be trained, while some HTML tags are set in an algorithm. At the time when this approach proposed, this rule might deem to be appropriate. However, Web design such as the use of Cascading Style Sheets (CSS) at present has changed and improved, so the algorithm might be affected and needs updates.

Kim, et al proposed "Unsupervised learning of mDTD extraction patterns for Web text mining" [5]. This approach mainly worked on structured webpages such as shopping sites. Document Type Definition (DTD) is modified and used in order to extract content patterns on webpages. Predefined rules by a human expert are set in DTD, which is used for training a classifier from webpages. The classifier is finally used to extract content from learning patterns. Chen et al proposed "An adaptive bottom up clustering approach for Web news extraction" [6]. Only single webpage was applicable and it was based on domain specific extraction with predefined rules. The process first starts by detecting the news areas based on the rules. In the meantime, the lowest level areas are then merged with its higher level based on space and visual properties until reaching a predefined threshold. The news areas are then verified based on position as well as space and format continuity. These techniques were not capable to work with both single and multiple page extraction.

3. Proposed Techniques

This paper proposes automatic single and multiple webpages content extraction techniques. The details of the single page extraction technique [8] have been described in another publication. In the reported work, although the single page extraction produces reasonable results, the accuracy of the extraction could be improved by extending the approach to multiple pages. Furthermore, parallel processing can be incorporated in the multiple page extraction approach in order to increase the processing speed and throughput.

3.1. Single Page Extraction

The proposed single page extraction technique is based on heuristic rules, and it also introduced XPath [9] language to retrieve and compare the elements on a webpage. This can reduce processing time as there is no need to traverse every node in a DOM tree of a HTML page, or the need to analyze the tree structure for comparison purpose.

The followings are the three main steps of the proposed single page extraction technique and it is also shown in Figure 1.

1) Web Page Elements and Features Extraction: This step is for the extraction of the features from each element in a webpage and to eliminate other unrelated elements.

After downloading a webpage, the page is transformed to a DOM tree. Specific rules [8] are applied in order to eliminate the unnecessary elements such as programming scripts.

2) Block Detection: This step is for the detection and selection of blocks or groups of elements on the page, as well as calculating the blocks' attributes.

XPath is incorporated in this process without executing a tree structure comparison. This will reduce the processing time as the XPath value is considered as a string. The lowest level blocks are then selected based on their attributes.

3) Content Extraction Selection: This step is for extracting informative content based on the calculated attributes by using a statistical approach.

Finally, the extracted contents of each webpage are collected. However, there are some non informative content found during the experiment. In order to improve the accuracy, multiple page extraction is then considered.

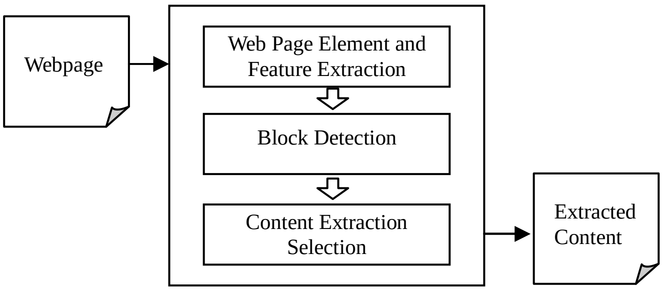
Webpage
Web Page Element and Feature Extraction
Block Detection
Content Extraction Selection
Extracted Content

Figure 1. An overview of the proposed single page extraction technique

3.2. Multiple Page Extraction

Most of the non informative content in the webpages within a website are similar. This leads to the proposal of a technique called *Extracted Content Matching (ECM)* for multiple page extraction. Moreover, parallel processing is also incorporated in this technique in order to increase the processing speed. The proposed multiple page extraction process is explained as follows:

1) Crawler: The Crawler is an initial module used for downloading the webpages. There are three main features as described below.

- *File compression:* The crawler requests compressed file(s) from the web server if it supports this feature. This feature can reduce downloading time.
- *Automatic Encoding:* Information on the Internet has been produced in different languages such as English, Chinese, or Thai. The crawler automatically encodes the content based on language encoding defined on page level, HTTP header, default setting in the crawler, respectively.
- *Automatic fixing the link address:* A link address in anchor tag (<a>) is defined by the webmaster and it could be in various formats. Some could be defined as full addresses such as "http://dailynews.co.th/newstartpage/index.cfm". In such case, the crawler can follow the address easily. The addresses can also be defined in short forms such as "index.cfm", "/index.cfm", or "../index.cfm". The short address has to be resolved to the full address in order to facilitate the crawler to look up the appropriate resource.

The crawler starts with downloading a seed page or webpage from an URL given by a user. Links are automatically extracted and resolved from the seed page. The crawler is then run in parallel by downloading all the extracted links as well as performing file compression and automatic encoding at the same time.

In the present experiment, the crawler is limited to downloading of two levels only: a seed page and all the pages links in the seed page. These pages are only downloaded if the pages are in the same domain of the seed page.

2) Single Page Extraction: After the webpages are downloaded, each webpage is then passed to the single page extraction process, which is also run in parallel. As a result, extraction of the content from each webpage is executed in this stage.

3) Extracted Content Matching (ECM): The Extracted Content, referred as EC in this paper, are the extracted elements on each webpage produced from the single page extraction, are required in this step. EC can be either informative content or non informative content. Examples of non informative elements in social media webpages are "share", "like" and "top of page". If these words appear as extracted content from a webpage, it is likely that they will also appear as EC in other pages within the same website. Therefore, the EC could be checked by matching them among the results from the multiple pages. If the same EC occur multiple times from different pages, it indicates that it is likely to be a non informative content. According to the initial experiment, it is observed that if a matched EC occurs in less than 3% of all the pages, the matched EC should be an informative element. The initial results are shown and discussed in Section VI.

An overview of the proposed single and multiple page extraction is shown in Figure 2.

3.3. Evaluation

Assessment of the accuracy of multiple page extraction is similar to the measurement of single page extraction as proposed in [8]. It is worth noting that there is no well established or commonly accepted means of measurement at this moment of time. The measurement adopted in this study is therefore based on *precision*, *recall* and *F-measure* of the extracted web content. These measurements are normally used in information retrieval and they are shown in expressions (1), (2), and (3), respectively. An important parameter used in this proposal is the use of the **length** of the informative and non-informative EC's. Based on a single page and by applying the rules for determining the nature of the elements, they are concatenated into three combined components known as LEC, LEP and LM. Definitions of these parameters are given in the expressions. The lengths of these components are then used to deduce the values of precision, recall and F-measure.

To assess the performance of the proposed technique,

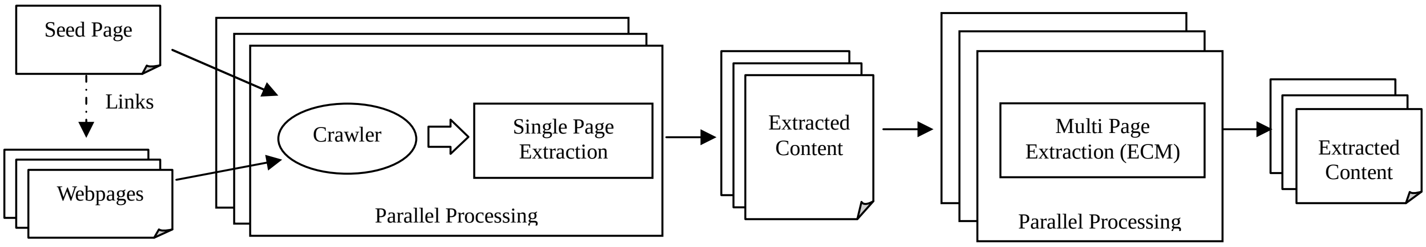

Figure 2. An overview of the proposed single and multiple page extraction

the extracted content is compared with the expected content in the webpage manually. Examples of the results are shown in Table 1.

$$\text{Precision} = (\text{LEC} - (\text{LEC} - \text{LEP} + \text{LM})) / \text{LEC} \quad (1)$$

$$\text{Recall} = (\text{LEC} - (\text{LEC} - \text{LEP} + \text{LM})) / \text{LEP} \quad (2)$$

$$\text{F-measure} = 2 * ((\text{Precision} * \text{Recall}) / (\text{Precision} + \text{Recall})) \quad (3)$$

Whereby

- LEC refers to the length of extracted content from the process.
- LEP refers to the length of expected content manually deduced from the web browser.
- LM refers to the length of missing content without spaces that the process cannot extract.

TABLE 1. EXAMPLES OF THE PROPOSED MEASUREMENT

	Page1	**Page2**	**Page3**
Length of non informative *EC*	200	200	200
LEC	1,700	1,500	700
LEP	1,500	1,300	500
LM	0	0	0
Precision	90.91%	89.66%	75%
Recall	100%	100%	100%
F-Measure	95.24%	94.55%	85.71%

In Table 1, the values of the three combined contents from Pages 1, 2 and 3 were extracted from the same website. For instance, the length of the expected informative content (LEP) is 1,500 characters in Page1, and there are 200 characters of non informative EC. The value of LEC in Page1 was 1,700 from the extraction process. Similar assessments are done on Pages 2 and 3. This measurement is used as a guideline for assessment of single page extraction and as an indicator to improve on the extraction technique. Furthermore, this measurement is able to present how much improvement on the accuracy of the proposed single and multiple page extraction techniques.

4. Experimental Results

In this experiment, the non informative ECs on the webpages in the same website are repeated on all the webpages. Initial results from the multiple page extraction technique on a Thai website are presented in Table 2. In this website, there are over 200 webpages linked into the seed page. Out of these pages 15 were randomly downloaded and extracted. Each webpage is extracted by single and multiple page extraction techniques to assess the improvement of the proposed multiple page extraction techniques as compared to the single page extraction approach.

As indicated in Table 2, the proposed techniques produced 100% of recall, and the proposed multiple page extraction is able to filter out the noise. There are only two pages that did not yield 100% of precision due to non informative content which were not shown on the web browser. Results based on URL's No.4 and No.5 in Table 2 are shown in Figure 3. The solid rectangles on Figure 3 enclosed areas of extracted informative content, and the dashed rectangle areas shown the extracted non informative content produced based on the single page extraction technique. However, there were frames in the right-hand-side of the page having been eliminated in the extraction process.

TABLE 2. EXPERIMENTAL RESULTS OF SINGLE AND MULTIPLE PAGE EXTRACTION[#]

No	URLs[*]	Single Page Extraction			Multiple Page Extraction		
		Precision	Recall	F-Measure	Precision	Recall	F-Measure
1	…page=content&categoryID=420&contentID=153533	94.50%	100%	97.17%	100%	100%	100%
2	…page=content&categoryID=419&contentID=153386	82.11%	100%	90.18%	98.75%	100%	99.37%
3	…page=content&categoryID=38&contentID=153389	77.50%	100%	87.32%	100%	100%	100%
4	…page=content&categoryID=424&contentID=153390	77.37%	100%	87.24%	100%	100%	100%
5	…page=content&categoryID=23&contentID=153325	91.52%	100%	95.57%	100%	100%	100%
6	…page=content&categoryID=460&contentID=152668	83.45%	100%	90.98%	100%	100%	100%
7	…page=content&categoryID=447&contentID=153327	86.16%	100%	92.56%	100%	100%	100%
8	…page=content&categoryID=333&contentID=153290	90.43%	100%	94.97%	100%	100%	100%
9	…page=content&categoryID=414&contentID=153241	91.55%	100%	95.59%	100%	100%	100%
10	…page=content&contentId=153564&categoryID=420	66.38%	100%	79.79%	100%	100%	100%
11	…page=content&contentId=153560&categoryID=420	61.35%	100%	76.04%	100%	100%	100%
12	…page=content&contentId=153559&categoryID=420	77.84%	100%	87.54%	100%	100%	100%
13	…page=content&contentId=153558&categoryID=420	78.80%	100%	88.15%	100%	100%	100%
14	…page=content&contentId=153557&categoryID=420	86.19%	100%	92.58%	99.08%	100%	99.54%
15	…page=content&contentId=153543&categoryID=420	77.51%	100%	97.33%	100%	100%	100%
		81.51%	**100%**	**89.53%**	**99.86%**	**100%**	**99.93%**

[#] These results were produced on 29 July 2011.

[a] Every URL begins with "http://dailynews.co.th/newstartpage/index.cfm?"
e.g. No2. is http://dailynews.co.th/newstartpage/index.cfm?page=content&categoryID=419&contentID=153386

The extracted non informative content areas are repeated in Figure 3, and they have been filtered out in the process. It can be seen from the diagrams that the proposed techniques have correctly extracted the informative content on URL's No.4 and No.5 from Table 2. Similar results have also been achieved for the other pages.

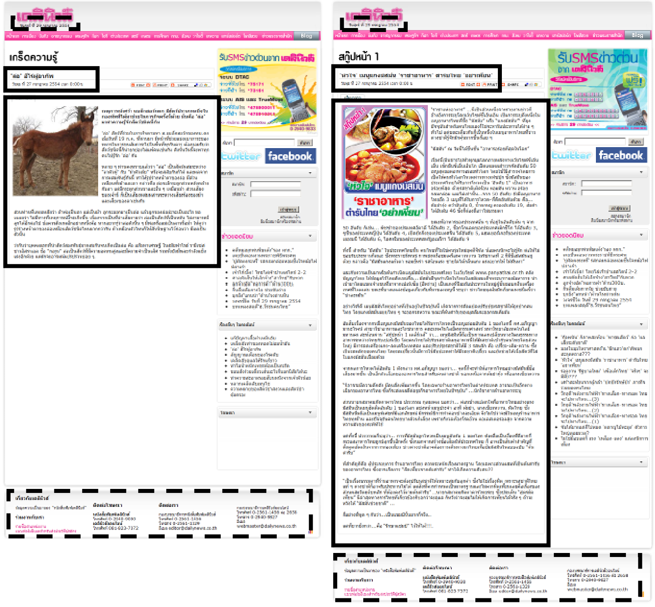

Figure 3. Examples of URL No. 4 and 5 from Table 2 and areas of extracted informative and non informative content

5. Conclusion and Discussion

Many different approaches have been proposed in order to address the issue of information overload due to the rapid growth of the Internet. One of the approaches is Web Content Extraction, and many techniques have been proposed on how to extract the informative content. This paper proposes both single and multiple page extraction techniques and some initial results are reported. The single page extraction technique is based on heuristic rules, and this technique also employs XPath language to manipulate elements on webpage.

This paper proposed an *Extracted Content Matching (ECM)* technique for extracting content from multiple pages based on a previously proposed single page extraction technique. Furthermore, parallel processing is incorporated in order to increase the processing speed. Other features included are file compression, automatic encoding, and automatic fixing of the link addresses.

Initial results from a Thai website are reported, and they were assessed by using precision, recall and F-measure based on the length of the extracted contents. The results have shown that the proposed *ECM* approach is able to improve the accuracy of the extracted content. The techniques produce 100% accuracy for 13 of 15 webpages in a website. The study will be extended by examining other websites.

References

[1] *World Internet Users and Population Stats*. 2011 March 31 [cited 2011 July 28]; Available from: http://www.internetworldstats.com/stats.htm.

[2] Netcraft. *July 2011 Web Server Survey*. 2011 8 July 2011 [cited 2011 July 28]; Available from: http://news.netcraft.com/archives/2011/07/08/july-2011-web-server-survey.html.

[3] Chun Che Fung and Wigrai Thanadechteemapat. *Discover Information and Knowledge from Websites Using an Integrated Summarization and Visualization Framework*. in *Third International Conference Knowledge Discovery and Data Mining, 2010. WKDD '10*. 2010. p. 232-235.

[4] Debnath, S., et al., *Automatic identification of informative sections of Web pages*. Knowledge and Data Engineering, IEEE Transactions on, 2005. 17(9): p. 1233-1246.

[5] Kim, Dongseok, et al., *Unsupervised learning of mDTD extraction patterns for Web text mining*. Information Processing & Management, 2003. 39(4): p. 623-637.

[6] Jinlin, Chen, et al. *An adaptive bottom up clustering approach for Web news extraction*. in *Wireless and Optical Communications Conference, 2009. WOCC 2009. 18th Annual*. 2009. p. 1-5.

[7] Fu, Lei, et al. *Web Content Extraction based on Webpage Layout Analysis*. in *Information Technology and Computer Science (ITCS), 2010 Second International Conference on*. 2010. p. 40-43.

[8] Wigrai Thanadechteemapat and Chun Che Fung. *Automatic Web Content Extraction For Generating Tag Clouds from Thai Web Sites*. in *The 8th IEEE International Conference on e-Business Engineering (ICEBE 2011)*. 2011. Beijing, China.

[9] *XML Path Language (XPath) 2.0 (Second Edition)*. 2011 January 3 [cited 2011 May 12]; Available from: http://www.w3.org/TR/xpath20/.