Automatic Web Content Extraction For Generating Tag Clouds from Thai Web Sites

Wigrai Thanadechteemapat and Chun Che Fung
School of Information Technology
Murdoch University
Murdoch, Australia
W.Thanadechteemapat@murdoch.edu.au, L.Fung@murdoch.edu.au

***Abstract*—This paper proposes a novel Web content extraction approach based on heuristic rules and the XPath utility in XML. The main objective is to address the problem of Web visualization by generating tag clouds from Thai Web sites in order to provide an overview of the key words in the Web pages. This paper also proposes a detailed method to assess the Web content extraction technique on a single Web page by using the length of the extracted content. There are three main steps in the proposed technique: Web page elements and features extraction, Block detection, and Content extraction selection. The empirical results have shown this technique produces high accuracies.**

Keywords-component; Web Content Extraction; XPath; Tag clouds

I. INTRODUCTION

Ever since its introduction to the general public, the Internet has been expanding dramatically over the past years. One important indicator is the number of Web servers, which is more than 324 millions in May 2011 [1]. This has led to the availability of an overwhelming amount of information on the Internet. Many researchers have been working on how to address the issue of information overloading by using different approaches such as text summarization, Web clustering, Web classification, opinion mining, Web visualization and a range of other techniques.

The use of Tag clouds is an example of Web content visualization, which is an approach to provide the overall content of Web pages [2]. There are generally two methods to generate tag clouds. One is to generate them from key words in a database, which are created by Web content authors. The other one is to create the tag cloud directly from the text within the Web page. There are some Web sites such as Wordle [1], WordItOut [2] and WordCrowd [3] that provide services for Tag cloud generation. Most of them do not accept URLs of the Web pages and many steps are required prior to creating tag clouds from the Web. Hence, it is necessary to provide some means to extract the key features or characteristics of the Web content.

There are a few obvious Web characteristics, which have impacts on Web information processing for tag cloud generation. The first one is the Web page's structure, which has informative content blending with HTML tags as well as some programming scripts for presenting the information through the Web browsers. The tags and scripts have to be separated from the content prior to the Web information processing. The other characteristic is related to the information presented on the Web pages, which contain not only the relevant content information, but it also includes other readable information on the same page such as advertisement, navigation icons, footers, etc. These categories of readable information can be considered as *noise* with respect to the processing. This leads to the need of Web content extraction in order to "filter" out the noise and extracting the relevant information only.

Although some tag cloud services are able to receive URLs as an input parameter, the other readable information will be included in the tag cloud results. It is therefore necessary to incorporate a stage of Web content extraction into the generation of tag cloud in order to offer more accurate overview of the Web pages to the users.

This paper proposes a novel Web content extraction technique, and it is applied to generate tag clouds from Thai Web pages as an example. The objective is to address the problem of Thai Web visualization using tag clouds based on Thai language. This technique is a part of a proposed visualization approach, which is intended to decrease the time required by Thai Internet users by providing overviews of the Web content. In addition, none of the current tag cloud services are able to support Thai websites due to the unique characteristics and features of Thai language. The main problem is the lack of spacing and separation between the Thai words, and Thai word segmentation is required in order to present the key words in the tag clouds. However, this paper only focuses on the Web content extraction technique and solutions for the issue of Thai word segmentation have been reported by the authors [3] and other researchers in other venues. The structure of this paper starts with an introduction on the background and aims of this paper. A summary of other related works is provided in Section II. Section III outlines the proposed technique of Web content extraction, and the empirical results are reported in Section

[1] http://www.wordle.net/
[2] http://worditout.com/
[3] http://tagcrowd.com/

IV. The last section concludes this paper followed by a discussion on future work.

II. Related work

Web content extraction can be generally grouped into two categories: single page and multiple page extraction. Some approaches work only on a single Web page, but some work on more than one page. Multiple page extraction may not work on a single Web page if many Web pages are required to identify the common template in order to extract the content of each page. Furthermore, some work focused on particular types of Web content such as news [4]. On the other hand, different techniques such as rule based [4-6] and machine learning [7-9] approaches have been employed in the Web content extraction. The rule based approach may work with Document Object Model (DOM) [10] due to the tree structure of the Web pages. The followings are some of the related work.

Jinlin Chen, at al [4] proposed "An adaptive bottom up clustering approach for Web news extraction". This approach is domain specific, but it can work on a single page with predefined rules. The process starts by identifying the news areas by using some rules. Meanwhile, the lowest level areas are then merged to 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.

Lei Fu, et al [7] proposed "Conditional Random Fields (CRFs) Model for Web Content Extraction". This approach at first has to transform the Web content in order to fit the requirement of CRFs by using predefined heuristic rules. The rules are for retrieving text nodes in the DOM tree of the web page. The text nodes are manually judged based on certain conditions such as minimum word counts and minimum URL counts in nodes etc to identify whether or not the nodes contain information and not navigation parts. Next, some features are defined for CRFs model during the training procedure. This approach requires some rules and conditions with judgment by human before the training step. Moreover, the number of training data set affects the accuracy, and it is time consuming in the preparation.

Sandip Debnath, et al [9] proposed "Automatic Identification of Informative Sections of Web Pages". This approach is required to work on multiple pages and classifier, which is trained with block or area features. Firstly, blocks are separated from the Web page based on rules. The blocks are then identified whether they are content blocks or non-content blocks with different methodologies. According to experimental results, algorithms were invented for identifying content areas. Some HTML tags are set in an algorithm whilst this rule might appropriate at the time when this approach was proposed. At present, Web design such as the use of Cascading Style Sheets (CSS) has been changed and improved, so the algorithm might be affected and needs updates.

III. The Proposed Technique

A. Web Page Element and Feature Extraction

This paper proposes XPath [11] language for retrieving and grouping elements on Web page without traversing every node in a DOM tree of a HTML page. This means reducing processing time of the tree analysis such as tree structure comparison.

The first step as shown in Figure 1 is to download the Web page(s) and then transform the HTML document to nodes in a DOM tree. Document node represents the entire document, and each HTML tag is represented as an element node, which may have other node types inside, such as element node, attribute node, text node, or comment node. After the DOM tree is built, only the body node is extracted. The following rules (*R1*) are applied to eliminate the non-informative content nodes.

- Exclude every element node and its children if the node has a style attribute with "display:none" value. The reason is these nodes are not displayed on Web browsers.
- Exclude script, noscript, embed, object, img, style, comment, and form node as well as the node with non-text such as line feed ('\n'), or null because Web content are not supposed to exist in these nodes.

After the non-informative nodes have been eliminated, only leaf nodes are extracted from the rest of the nodes as well as extracted their features based on the followings rules (*R2*).

- Store XPath for processing in the Block Detection.
- Store Text and convert special characters to readable text if needed, such as converting from "&" to "&".
- Mark node as a footer of the page if the content begins with predefined keywords such as "powered by".
- Mark node if it is anchor tag.

The extracted leaf nodes and their features (*All Leaf Nodes*) are then processed in the next section.

B. Block Detection

This paper also proposes a new technique for detecting groups of element nodes or blocks based on XPath.

After the node and feature extraction is processed, the features of each leaf node are calculated and accumulated into XPath blocks. There are attributes in each block as shown below.

- Number of all element nodes (*All Elements*)
- Number of all anchor element nodes (*All Href*)
- Length of text from all element nodes (*All T- Length*)
- Length of text from all anchor element nodes (*All Href T-Length*)
- Anchor Node Ratio (*ANR*) is calculated from (*All Href*) divided by (*All Elements*)
- Anchor Text Ratio (*ATR*) is calculated from (*All Href T-Length*) divided by (*All T- Length*)

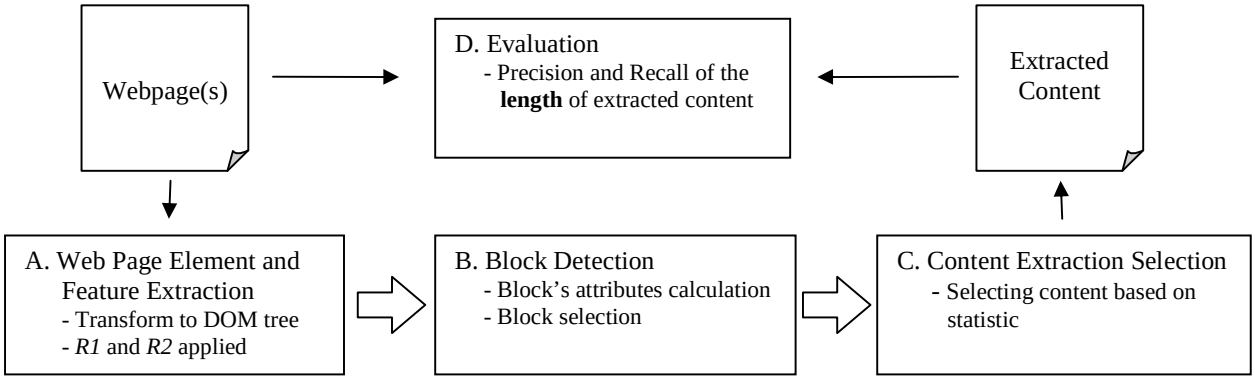

Figure 1. An overview of the proposed technique

- Anchor Ratio (*AR*) is ((*ATR*) * .75) + ((*ANR*) * .25) whereby the predefined values come from empirical experiment.

All attributes of each XPath block are calculated by starting from XPath of a parent of each leaf node and step up until the parent is a body element node.

At this stage, there is a list of XPath and its attributes built, and block detection is then performed based on the following steps.

1) Block flitering: All blocks that have *All Elements* more than 1 are selected. *All Elements* being 1 means XPath referring to text; not a block or a group of element nodes.

2) The lowest level block selection: The smallest blocks should be detected as it is convenient to do upper grouping. This step uses XPath to compare together like a normal string. There is no need to traverse and compare nodes of the DOM tree.

C. Content Extraction Selection

The selection is based on statistical approach. Every leaf node in *All Leaf Nodes* is considered as Web content only if:

- its footer feature is false; otherwise, this node is considered as a footer, and
- its text length is less than 2 characters, and
- (*AR*) of its group is less than 60%. The higher (*AR*) means the high possibility to be navigation parts, or (*All Href*) is less than 2 and (*All Elements*) is more than 2. This rule is for Web page's subject text, which is anchor in the same block of many element nodes.

Finally, Web contents are extracted and used in the next process of Thai Web visualization, which comprises Thai word segmentation and tag cloud generation.

D. Evaluation

Detailed measurement of Web content extraction on single page is presently unavailable. Only results of overall number of Web pages are therefore proposed. Consequently, this paper proposes the measurement on single page based on *precision* and *recall* of the length of extracted Web content, and this can be extended to work on multiple pages.

In order to evaluate the accuracy of the proposed technique, *precision* and *recall* as well as *F-measure* are calculated on the length of the expected content and the length of the extracted content. The three measurements are shown in expressions (1), (2), and (3), respectively.

The extracted content is derived from the Web content extraction process while the expected content is manually taken from the content presented on Web browser. Spaces are removed from both contents when determining the length in order to reduce errors during copy and paste of the expected content.

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

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

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

Whereby

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

IV. EXPERIMANTAL RESULTS

Observation in this experiment has provided some insights. Some noises were shown in the results and they are repeated in the same Web site template. This implies the number of Web pages examined in the same template does not have impacts on the accuracy. Three Web pages in different Thai Web sites were randomly used to assess the proposed technique. In addition, there is no detailed measurement or baseline model on single page of other techniques, so it is not suitable to compare with other results. The empirical results of this proposed technique are shown in Table1.

Figure 2. An example of Thai Web Page

Moreover, the percentage values depended on the length of content on each page, whilst the noises are the same amount of length. This means this technique should be improved on noise flittering and block detection in order to increase accuracy.

TABLE I. RESULTS FROM THE PROPOSED TECHNIQUE

Empirical Results			
Web sites	***Precision***	***Recall***	***F-Measure***
News : www.matichon.co.th	100%	100%	100%
Wiki : th.wikipedia.org	93.16%	100%	96.43%
Information : www.dmc.tv	95.04%	100%	97.44%

The proposed technique produces 100% of recall for the three Web sites while there is only one Website having 100% of precision. This means this technique is able to extract all Web content even though some noises have been included and affected the accuracy of the extraction.

Examples of the actual results are shown in Figure 2 to 4. Thai Web page[4] from Matichon Online is shown in Figure 2. There are not only navigation parts on the top and bottom, but there are also advertisements on the top and right side embedded with links. Figure 3 shows the extracted Thai content from the Web page on Figure 2. In addition, a tag cloud is shown in Figure 4 which was generated by using the extracted content and a Thai word segmentation process which is not described in this paper.

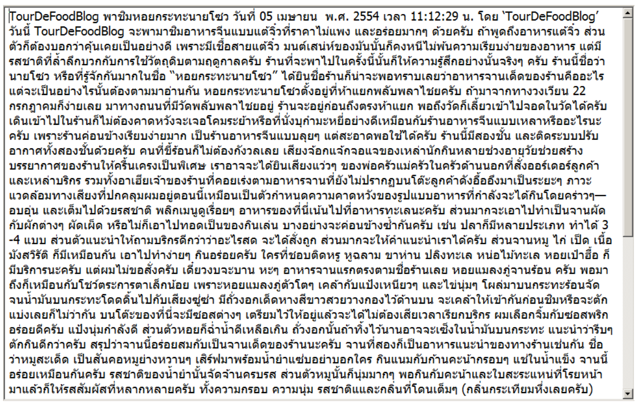
TourDeFoodBlog พาชิมหอยกระทะนายโซว วันที่ 05 เมษายน พ.ศ. 2554 เวลา 11:12:29 น. โดย 'TourDeFoodBlog'
วันนี้ TourDeFoodBlog จะพามาชิมอาหารจีนแบบแต้จิ๋วที่ราคาไม่แพง และอร่อยมากๆ ด้วยครับ ถ้าพูดถึงอาหารแต้จิ๋ว ส่วน
ตัวก็ต้องบอกว่าคุ้นเคยเป็นอย่างดี เพราะมีเชื้อสายแต้จิ๋ว มนต์เสน่ห์ของมันนั้นก็คงหนีไม่พ้นความเรียบง่ายของอาหาร แต่มี
รสชาติที่ล้ำลึกบวกกับการใช้วัตถุดิบตามฤดูกาลครับ ร้านที่จะพาไปในครั้งนี้นั้นก็ให้ความรู้สึกอย่างนั้นจริงๆ ครับ ร้านนี้ชื่อว่า
นายโซว หรือที่รู้จักกันมากในชื่อ "หอยกระทะนายโซว" ได้ยินชื่อร้านก็น่าจะพอทราบเลยว่าอาหารจานเด็ดของร้านคืออะไร
แต่จะเป็นอย่างไรนั้นต้องตามมาอ่านกัน หอยกระทะนายโซวตั้งอยู่ที่ห้าแยกพลับพลาไชยครับ ถ้ามาจากทางวงเวียน 22
กรกฎาคมก็ง่ายเลย มาทางถนนที่มีวัดพลับพลาไชยอยู่ ร้านจะอยู่ก่อนถึงตรงห้าแยก พอถึงวัดก็เลี้ยวเข้าไปจอดในวัดได้ครับ
เดินเข้าไปในร้านก็ไม่ต้องคาดหวังจะเจอโคมระย้าหรือที่นั่งบุกำมะหยี่อย่างดีเหมือนกับร้านอาหารจีนแบบเหลาหรืออะไรนะ
ครับ เพราะร้านค่อนข้างเรียบง่ายมาก เป็นร้านอาหารจีนแบบลุยๆ แต่สะอาดพอใช้ได้ครับ ร้านนี้มีสองชั้น และติดระบบปรับ
อากาศทั้งสองชั้นด้วยครับ คนที่ขี้ร้อนก็ไม่ต้องกังวลเลย เสียงจ้อกแจ้กจอแจของเหล่านักกินหลายช่วงอายุวัยช่วยสร้าง
บรรยากาศของร้านให้ครื้นเครงเป็นพิเศษ เราอาจจะได้ยินเสียงแว่วๆ ของพ่อครัวแม่ครัวในครัวด้านนอกที่สั่งออร์เดอร์ลูกค้า
และเหล่าบริกร รวมทั้งอาเฮียเจ้าของร้านที่คอยเร่งตามอาหารจานที่ยังไม่ปรากฏบนโต๊ะลูกค้าดังอื้ออึงมาเป็นระยะๆ ภาวะ
แวดล้อมทางเสียงที่ปกคลุมผมอยู่ตอนนี้เหมือนเป็นตัวกำหนดความคาดหวังของรูปแบบอาหารที่กำลังจะได้กินโดยคร่าวๆ—
อบอุ่น และเต็มไปด้วยรสชาติ พลิกเมนูดูเรื่อยๆ อาหารของที่นี่เน้นไปที่อาหารทะเลนะครับ ส่วนมากจะเอาไปทำเป็นจานผัด
กับผักต่างๆ ผัดเผ็ด หรือไม่ก็เอาไปทอดเป็นของกินเล่น บางอย่างจะค่อนข้างซ้ำกันครับ เช่น ปลาก็มีหลายประเภท ทำได้ 3
-4 แบบ ส่วนตัวแนะนำให้ถามบริกรดีกว่าว่าอะไรสด จะได้สั่งถูก ส่วนมากจะให้คำแนะนำเราได้ครับ ส่วนจานหมู ไก่ เป็ด เนื้อ
มังสวิรัติ ก็มีเหมือนกัน เอาไปทำง่ายๆ กินอร่อยครับ ใครที่ชอบติดหรู หูฉลาม ขาห่าน ปลิงทะเล หน่อไม้ทะเล หอยเป๋าฮื้อ ก็
มีบริการนะครับ แต่ผมไม่ขอสั่งครับ เดี๋ยวงบจะบาน หะๆ อาหารจานแรกตรงตามชื่อร้านเลย หอยแมลงภู่จานร้อน ครับ พอมา
ถึงก็เหมือนกับโชว์ตระการตาเล็กน้อย เพราะหอยแมลงภู่ตัวโตๆ เคล้ากับแป้งเหนียวๆ และไข่นุ่มๆ โผล่มาบนกระทะร้อนจัด
จนน้ำมันบนกระทะโดดดิ้นไปกับเสียงซู่ซ่า มีถั่วงอกเด็ดหางสีขาวสวยวางกองไว้ด้านบน จะเคล้าให้เข้ากันก่อนชิมหรือจะตัก
แบ่งเลยก็ไม่ว่ากัน บนโต๊ะของที่นี่จะมีซอสต่างๆ เตรียมไว้ให้อยู่แล้วจะได้ไม่ต้องเสียเวลาเรียกบริกร ผมเลือกจิ้มกับซอสพริก
อร่อยดีครับ แป้งนุ่มกำลังดี ส่วนตัวหอยก็ฉ่ำน้ำดีเหลือเกิน ถั่วงอกนั้นถ้าทิ้งไว้นานอาจจะเซ็งในน้ำมันบนกระทะ แนะนำว่ารีบๆ
ตักกินดีกว่าครับ สรุปว่าจานนี้อร่อยสมกับเป็นจานเด็ดของร้านนะครับ จานที่สองก็เป็นอาหารแนะนำของทางร้านเช่นกัน ชื่อ
ว่าหมูสะเด็ด เป็นสันคอหมูย่างหวานๆ เสิร์ฟมาพร้อมน้ำยำแซ่บอย่าบอกใคร กินแนมกับก้านคะน้ากรอบๆ แช่ในน้ำแข็ง จานนี้
อร่อยเหมือนกันครับ รสชาติของน้ำยำนั้นจัดจ้านครบรส ส่วนตัวหมูนั้นก็นุ่มมากๆ พอกินกับคะน้าและใบสะระแหน่ที่โรยหน้า
มาแล้วก็ให้รสสัมผัสที่หลากหลายครับ ทั้งความกรอบ ความนุ่ม รสชาติและกลิ่นที่โดนเต็มๆ (กลิ่นกระเทียมหึ่งเลยครับ)

Figure 3. Extracted Thai Web Content from the proposed technique

Figure 4. An example of Thai Web page

[4] http://matichon.co.th/news_detail.php?newsid=1301976014&grpid=&catid=09&subcatid=0901

V. CONCLUSION AND DISSCUSSION

Many researchers have been addressing the information overload on the Internet by using diverse approaches. Web visualization is one of them in order to provide an overview of the content on Web pages, and tag cloud is considered as a common tool for content visualization. However, most of the tag cloud services are not able to extract content directly from Web page due to the structures of the Web pages and various types of information presented on Web browsers.

A novel Web content extraction is proposed in this paper. DOM and XPath are employed to represent a tree structure of the HTML document, while XPath is used to retrieve the elements from the Web page. There are three main steps in this technique. The first one is Web page element and feature extraction. This step is to eliminate non-informative content elements based on some rules and to extract the features of the leaf nodes. Next, block detection is used to identify the proper blocks or groups of elements in the Web page based on calculations of the attributes of each block. The last step is content extraction selection. This step uses rules to select only informative elements based on block attributes. Furthermore, this paper proposes use of length of extracted content in precision and recall measurement for assessment of accuracy of the proposed approach.

The empirical results show that this technique is able to extract Web content with high accuracy over 96% from a number of Thai Web sites. According to experiment, the accuracy can be increased further if noise filtering and block detection are improved. Moreover, this technique is expected to work on multiple Web pages in the future. The accuracy should be also increased because noises can be detected by being compared among the pages. Further experiments and development will be carried out to support the hypotheses.

REFERENCES

[1] Netcraft. May 2011 Web Server Survey. 2011 2 May 2011 [cited 2011 May 13]; Available from: http://news.netcraft.com/archives/2011/05/02/may-2011-web-server-survey.html.

[2] 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.

[3] Wigrai Thanadechteemapat and Chun Che Fung, Thai Word Segmentation For Visualization Of Thai Web Sites, in International Conference on Machine Learning and Cybernetics 2011 (ICMLC 2011). 2011: Guilin, China. p. (Publishing).

[4] 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.

[5] Dingkui, Yang and Song Jihua. Web Content Information Extraction Approach Based on Removing Noise and Content-Features. in Web Information Systems and Mining (WISM), 2010 International Conference on. 2010. p. 246-249.

[6] 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.

[7] Fu, Lei, et al. Conditional Random Fields Model for Web Content Extraction. in Computing in the Global Information Technology (ICCGI), 2010 Fifth International Multi-Conference on. 2010. p. 30-34.

[8] Htwe, T. and Hla Khin. Noise removing from Web pages using neural network. in Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on. 2010. p. 281-285.

[9] 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.

[10] W3C Document Object Model. 2009 6 Jan [cited 2011 May 12]; Available from: http://www.w3.org/DOM/.

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