Automatic Content Extraction and Visualization of Thai Websites for Improved Information Representation

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

***Abstract*—This paper presents an integrated approach to automatically provide an overview of content on Thai websites based on tag cloud. This approach is intended to address the information overload issue by presenting the overview to users in order that they could assess whether the information meets their needs. The approach has incorporated Web content extraction, Thai word segmentation, and information presentation to generate a tag cloud in Thai language as an overview of the key content in the webpage. From the experimental study, the generated Thai Tag clouds are able to provide an overview of the tags which frequently appear in the title and body of the content. Moreover, the first few lines in the tag cloud offer an improved readability.**

Keywords; Thai tag cloud, Web Content Extraction, Thai Word Segmentation, Keyword Extraction, Maximum Term Frequency

I. INTRODUCTION

Search engines currently are used as the de-facto tools to help users to address the issue of information retrieval, but a large amount of results are normally returned. One approach to help the users in searching for appropriate information from the search results is to provide an overview of the Web content in the form of information visualization. An example is the use of tag clouds [1], which refers to a group of words or tags being extracted from the original content to represent the characteristics of the content [2]. Currently, websites such TagCrowd[1] and Wordle[2]offer a service to generate tag clouds, but there are limitations. For example, most of the services support the generation of tag clouds only with text supplied by the users, and they cannot accept a URL as a reference to a particular webpage where the service is directed to. The main reason is that there are many steps concerned before a tag cloud can be generated from a webpage. In addition, those tag cloud services support only certain languages since there are unique characteristics and features for each language.

In order to generate the visualization of Web content based on tag clouds from URL directly, there are two major challenges. The first one is how to extract the key features or characteristics of the Web content from a website. It is necessary to carry out a process of *Web content extraction*. For certain languages such as Chinese, Japanese, Korean and Thai normally not supported by the current tag cloud services, an additional challenge is how to segment the words in order to present them in a tag cloud. In particular, Thai words do not have space separation as in English, which is the most popular language on the Web. *Word segmentation* is therefore required to process the Thai Websites for information presentation purposes and this forms an essential aspect in this project.

In this study, Thai websites have been targeted because this will help to increase the use of Internet among the Thai nationals. At present, the population of Thailand is about 1.1% of the world but the presence of websites in Thai language is about 0.3% [3]. This correlates with the number of Internet users in Thailand which stands at about 18 millions. Furthermore, as Thailand as a part of the ASEAN Community, will partake in the free flow of information, goods, services, capital and skilled labors among the ASEAN members by 2015[4]. Hence, it is important to improve the country's computer literacy and to increase the participation in the cyber world among the Thai people.

To this end, this study proposes an integrated approach to automatically generate an overview of key content from Thai webpages based on tag cloud. While tag clouds have already existed in many websites, but they are produced with words provided by the web authors. There is no clear association with the content in the webpage and the process could be tedious. The proposed approach in this paper incorporates novel *Web content extraction* and *Thai word segmentation techniques*. The objective is to address the fundamental problems of information presentation and extraction from Thai websites for better use by human readers. This approach also addresses the information overload issue by providing an overview of webpage(s) to the users so that they could assess whether the information meets their needs.

The structure of this paper begins with an introduction and the purposes of this study. Section II provides other related work on Web content extraction, Thai word segmentation, and information presentation based on tag cloud. The proposed

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

approach is then described in Section III, and Section IV reports the experimental results. Finally, Section V concludes the paper and provides a discussion on future work.

II. Related Work

A. Web Content Extraction

Web content extraction is intended to extract key content from one or multiple webpages. Web content extraction can be used to enhance the results from Web crawling, classification, Web data mining, and presentation of information as well as text-to-speech and translation tools [5].

A webpage is an electronic document that contains various types of contents such as text, picture, interactive multimedia, and/or hyperlinks. These contents appear on a webpage in different parts such as key content, header, footer, navigation, or advertisement etc. More than likely, the user is only interested in the key content so as to acquire the information they are looking for [6]. Depending on the design and layout of a webpage, different parts may contain a range of information in different formats being presented to the users. The design and layout of the webpages may follow certain templates in order to create a sense of uniformity design. Researchers such as Davi de Castro Reis et al. [7] have chosen to use the template as a key factor in extracting the main content. However, some websites such as Blog may not apply the same template for all the pages within the website. Therefore, Web content extraction based on identifying the template may not be applicable in this situation.

The approaches for Web content extraction can be generally grouped into two categories [8] which are either based on single page, or, multiple page extraction. Moreover, some approaches work on particular types of content such as news. In addition, different techniques such as machine learning [9] and rule based [10] approaches have also been utilized. Document Object Model (DOM) has been employed in the rule based approach due to the tree structure of the webpages, which enable the handling of the elements on each webpage with ease.

B. Thai Word Segmentation

Thai word segmentation in the proposed approach is required to determine the boundary between the Thai words. Thai word segmentation contributes to other related fields as it is fundamental for processes in Natural Language Processing (NLP) which include Thai word correction, Thai sentence extraction, and summarization.

By definition, a word is the smallest language unit that has specific meaning [11], and it can be usually recognized based on different features [12, 13] such as:

- Orthography: Written marks such as a space can be used to segment words, but it is not strict criterion. For example, “ice cream” is considered as a word even it has a space in between. In the case of Thai language, there is no written mark to segment Thai words.
- Phonology: A part of speech or a unit of pronunciation can be used to segment the words. For example, when spoken, English words may have emphasis which indicate the segmentation. However, this does not apply to Thai language in the spoken or written formats.
- Lexicon: This refers to items contained in a dictionary. The lexical items can be different forms such as grammatical forms, phrasal verbs, and prepositional verbs. This is applicable to Thai language, and it is therefore used in this study.

While it is not easy to deal with the issue of word segmentation due to the characteristics of the languages, different results may also be produced by the manual approach [13, 14] and this leads to the lack of consistency and accuracy. Ideally, Thai word segmentation techniques should produce results either simple words or compound words [13]. The simple words or morpheme refer to words having one minimal meaningful unit of word. An example is the word รถ which means car. Compound words consist of more than one morpheme or one word stem, and they may offer different meanings when combined. An example is, เสียสละ, which means *sacrifice* and it is made up by two words: เสีย – *broken*, and, สละ – *to discard*. The meaning of the compound word is different from the individual meanings of the two words. However, if the compound word has no separate meaning, they could be segmented into multiple words. An example is คนจน which means *poor people*, made up by the words คน – *people*, and จน – *poor*. In such case, the segmentation will be sufficient by dealing with simple words and this is the approach adopted in this study.

Techniques applied to Thai word segmentation can be classified as two categories: dictionary based, and non dictionary based. Dictionary based approaches require a list of items in a common dictionary, and example techniques are longest matching [15], maximum matching [16] and decision tree. The non dictionary based do not need the common dictionary, and techniques are, for instance, rule-based [17], Hidden Markov Model (HMM) [18] and Native Bayesian [19]. This technique proposed in this approach is based on the longest matching technique and the utilization of a corpus. These are relatively simple to implement and experimentation with the other techniques are planned for future work.

C. Information Presentation based on Tag Cloud

A tag is referred to words extracted from an original content, and they are displayed collectively in a tag cloud. It is a means to represent the characteristics of the original content [2] and this is a form of visualization or summary [20, 21] as they provide conceptual information of the original content [22]. The importance of the tags could be visualized based on size, color and style. The idea is to draw the user’s attention with the perceived importance of the tags. The tags can be normally chosen based on statistical characteristics such as associating the number of occurrences to define the “weight” of each tag.

The tags can be generated by two methods [23]. The first one is to use pre-defined words from a database, which are usually created by the Web content authors. The tag clouds

generated from this method are often seen in Content Management Systems (CMS) since the systems aim to present an overview of their content to the users. The other method refers to the tag clouds generated directly from the content on a webpage. Apparently, not many websites offer this method due to technical issues, and one of the issues is how to extract key content from webpages. In addition, none of the current services support the generation of Thai tag clouds. Therefore, this study aims to fulfill the gap by the integration of Web content extraction and Thai word segmentation of webpages based on a given URL. The approach is described in the following section.

III. The Integrated Approach

The proposed integrated approach is aimed to benefit users in assessing the information on a Thai website based on tag cloud, as an overview based on either single or multiple webpages. There are three main modules in this approach and an illustration of the approach is shown in Fig. 1.

A. Web Content Extraction

Techniques of Web content extraction based on heuristic rules have been integrated in this proposal and they can extract key content from both single and multiple pages. A previously published technique on single page extraction has been reported by the authors in [8] and the technique has been further improved to address multiple page extraction. A summary of the techniques are given as follows:

1) Single Page Extraction Technique

a) Web Page Element and Feature Extraction

This step is intended to extract features from each element in a webpage and to eliminate other unrelated elements. The step starts with downloading a webpage and transforming the page to a DOM tree. Specific rules described in [8] are applied in order to eliminate the unnecessary elements such as programming scripts.

b) Block Detection

A block is referred to a group of elements. This step is for the detection and selection of blocks on the page, as well as calculating the attributes of the blocks.

c) Content Extraction Selection

The last step is to extract key content based on the calculated attributes of the blocks by using statistical approaches such as the ratio between the number of characters and hyperlinks inside the blocks.

2) Multiple Page Extraction

A technique called *Extracted Content Matching (ECM)* has been proposed for multiple page extraction by improving the single page extraction. This technique is aimed to eliminate noises from the set of extracted key content. Most of the non informative extracted content on the webpages are similar and examples are some standard text icons such as *search* and *vote*. It is likely that they will also appear in the key content of other pages within the same website. Therefore, each element in a set of the extracted key content could be checked by matching all extracted elements from other webpages. If the same elements occur in multiple times from different pages, they could be treated as a noise. In addition, parallel processing is also incorporated in this technique in order to increase the processing speed. After this stage, the extracted key content can be passed to next module for word segmentation.

B. Thai Word Segmentation

The technique used in this module is based on the longest matching technique by using a refined hybrid corpus instead of a dictionary. This technique has already been published in [23] and the objective is to segment Thai words in the extracted key content. The corpus should be verified whether the segmented words included in the corpus are consistent before it is utilized. If there are inconsistent words found, they have to be resolved. The process of refinement refers to the identification and resolving the inconsistent segmented words in the corpus. After the corpus is refined, it is then sorted in descending order. In addition, there are two more collections: Thai named-entities, and titles of person names. They are used in the segmentation process, and they are also sorted in descending order.

The proposed technique starts with separating the key content into small pieces of text by using a *new line* character as a separation so that the text units could be processed in parallel. The process then fetches each entry in the two collections mentioned above, and from the refined corpus in order to look for any match with each text unit. If there is a match, the matched strings will be marked as a word. Finally, the segmented words in each piece of text are merged in the same order of the original key content, and the key content now consists of segmented Thai words. The key content will be transformed to present in a tag cloud in the next module.

C. Automatic Thai Tag Cloud Generation

In order to present an overview of the single webpage or webpages, some of the Thai segmented words should be selected as keywords so that they can present the characteristic of the original content. This leads to the need of a method called *keyword extraction* that is used to extract the keywords from the key content.

a) Keyword Extraction

With respect to the process of automatic Thai tag cloud generation, the popular *term frequency and inverse document frequency (TF*IDF)* technique [24] was applied to perform keyword extraction on the given Thai key content on single and multiple webpages. The *TF*IDF* technique works with a collection of documents and this is not applicable to a single webpage. In this study, a single page is decomposed into multiple documents by considering the content is being separated by specific characters such as paragraph breaks, title, footer and header…etc. However, it was found during the experiments that the accuracy of *TF*IDF* were quite low. This may be due to the reason that the separated documents are lacking consistency between them. To the best understanding of the authors, there is no available technique for keyword extraction from Thai Web content of a single webpage. An alternate method other than *TF*IDF* is therefore required.

A statistical technique, called the *Average Maximum Term Frequency (AMTF),* was developed to extract the Thai

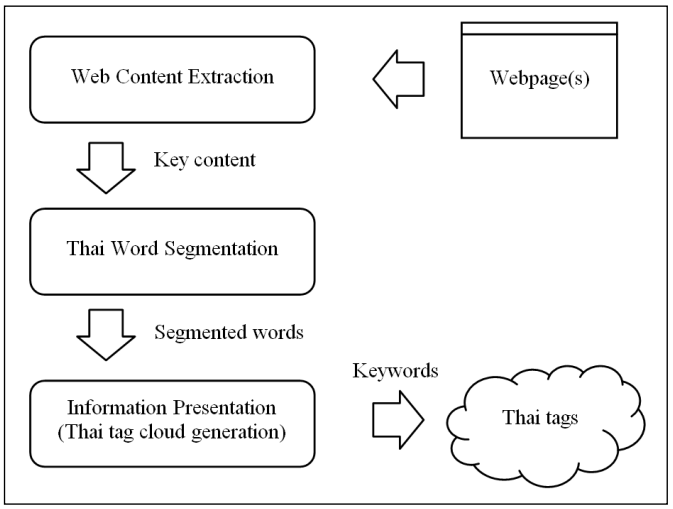

Figure 1. The process of the integrated approach

keywords on a single webpage. AMTF is based on the normalization of the *term frequency* by using *maximum term frequency* [25], which is a well-studied technique [26]. The idea of AMTF is to define a weight for each Thai segmented word by using an average of the maximum term frequency of the same word in the different elements, or the documents in the *TF*IDF* approach. It was found that this approach yields better accuracy and therefore it is adopted in this work.

*b) Evaluation of the AMTF and the TF*IDF approaches*

To evaluate the proposed approaches, a data set of 20 webpages from four Thai websites is used. Each of these webpages has tags, or keywords, provided by the web authors while not many Thai websites provide both keywords and content. These webpages comprise of different categories of content such as news, fashion, entertainment, politics….etc. In order to assess the performance of the proposed techniques, the key content of each page is extracted by ignoring the irrelevant information. The extracted key contents are then segmented and applied to the two approaches, AMTF and TF*IDF as inputs. The keywords produced by these approaches are then compared with the tags from the website. Table I shows a comparison of the results between TF*IDF and AMTF. Based on metrics of Precision, recall and F-Measure, the results of AMTF are found to be better than those due to TF*IDF's.

While it appears that the automatic process does not produce 100% accuracy, it should be recognized that the tags were provided by the web authors through a manual process. There is no assurance that the tags are the keywords based on the highest frequency of occurrence. They were just tags perceived to be important by the authors. In addition, the proposed approach provides consistency in the segmentation and extraction processes and the correct key content has found to be included in the output.

Moreover, the Thai tags provided by the webpages through the manual process usually consist of named-entities, compound words, and single words, as well as words that are NOT included in the original content. This affected the accuracy of the result. On the other hand, the segmented words produced by this study are in the form of smallest meaning units. When comparing between the results from the developed approach to the tags in the website, it was observed that a compound word in the web could be segmented into smaller words as shown in Fig. 2. In the example, the "บัตรทอง" compound word in the tag provided in the Web was identified as two segmented smaller units: "บัตร" and "ทอง" from the processed output. Similarly, the "30บาทรักษาทุกโรค" word was recognized as "30", "บาท", "รักษา", "ทุก" and "โรค". This shows the versatility of the proposed approach in determining the compound words as well as the segmented smaller single words.

TABLE I. COMPARISON OF THE RESULTS BETWEEN TF*IDF AND AMTF FOR THAI KEYWORD EXTRACTION

Technique	Precision	Recall	F-Measure
AMTF	48.51%	24.72%	28.71%
TF*IDF	31.04%	1.70%	3.16%

IV. EXAMPLE RESULTS

The integrated approach is able to provide an overview of the content from Thai webpages based on tag cloud. An example webpage[3] shown in Fig. 3 is used to illustrate the approach. The webpage is from the *Manager Online*, where was the highest ranked Thai news website according to the list[4] of Top Websites in Thailand provided by Alexa.

In Fig. 3, the key content appears in different areas, as showed by the dashed-line boxes while the other areas contain non key contents. The webpage was analyzed and extracted by the three proposed modules, and the resultant Thai tag cloud is shown in the lower half of Fig. 4 within the box. The tag cloud comprised tags from top 45% weights of the keywords. It is noted that the first part of the tag cloud is consistent with the title of the key content as illustrated in the diagram since the tags are ordered by their appearance. This aspect is unique and different as words in tag clouds normally appear randomly. This offers better readability as the sequence of the tags is in

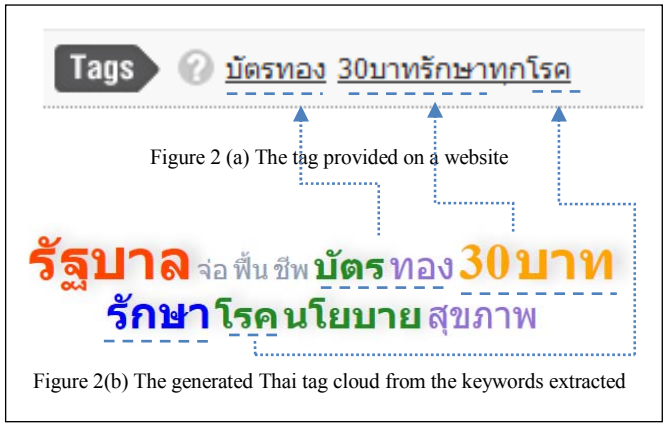

Figure 2 (a) The tag provided on a website

Figure 2(b) The generated Thai tag cloud from the keywords extracted

Figure 2. A comparison between a tag cloud of the extracted keywords and tags given on a website

[3] http://www.manager.co.th/Home/ViewNews.aspx?NewsID=9550000056013 - last accessed on 6 May 2012

[4] http://www.alexa.com/topsites/countries/TH - last accessed on 2 May 2012

the same order as the original title. The rest of the tags represent main words in the body part of the content. In this manner, the generated tag cloud may be considered as an overview or summary of the key content since the tags contain both the title and body of the key content. This observation remained correct when multiple pages of the same theme were considered as shown in Fig 5. The tag cloud appeared as a summary of the content in the multiple pages.

However, this observation is not applicable to the case when multiple webpages of different themes are considered. It is expected that the tag cloud generated will be a broad overview of all the pages instead of appearing as a single theme.

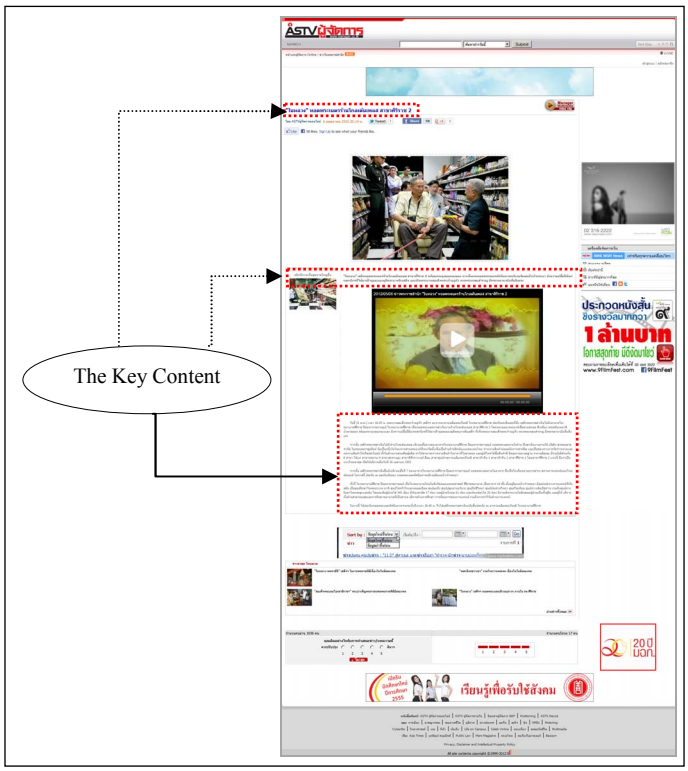

Figure 3. A webpage used in thie approach to show the key content

V. CONCLUSION AND FUTURE WORK

In this paper, an integrated approach aimed to present an overview of content from Thai Webpages based on tag cloud is reported. The approach has incorporated Web content extraction and Thai word segmentation. The objective was to address the fundamental problems of information presentation and extraction for Thai websites. Furthermore, this approach was intended to address the information overload issue by providing an overview of single or multiple webpages to the users so that they could assess whether the information meets their needs. 20 webpages have been used to evaluate the performance of the proposed techniques. Based on the experiments, the automatic generated Thai tag clouds are generally able to present an overview of both single and multiple webpages. Thai tag clouds of single page are capable of providing an overview as the extracted tags usually contain both the title and body of the content. In addition, the tags in the first few lines offer better readability as their sequence is similar to the original content. As a result, the users are able to appreciate the key themes of the article provided by this approach. On the other hand, the tag clouds of multiple pages could present a broader picture based on the key themes from the categories of contents as there may be no single key theme among the multiple pages.

Figure 4. A generated Thai tag cloud being compared with the title of the actual content on a website

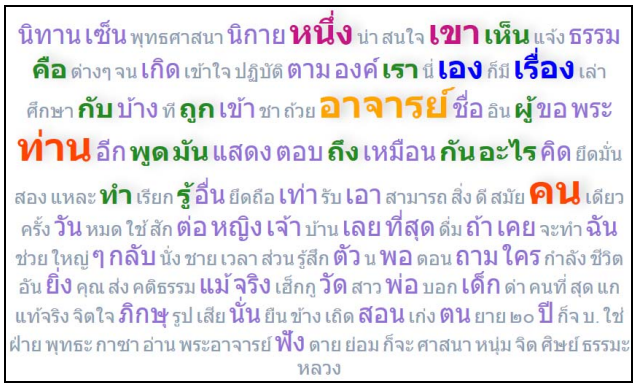

Figure 5. A Thai tag cloud automatically generated from multiple webpages[5]

As regard to future work, rules based techniques of Web content extraction could be used to analyze and extract key content from webpages, however, some noises have found to be included in the extracted key content. The accuracy of the

[5] www.whatami.net/zen/zen.html - last accessed on 8 May 2012

techniques could be improved by employing clustering techniques to identify groups of content elements from the webpages. On the other hand, the Thai word segmentation technique was based on the longest matching technique and refined corpus. Wrong segmented words could be produced due to the lack of context consideration. Therefore, other grammatical elements such as part-of-speech, statistical approach and rules based on characteristics of the Thai language could be considered to be included in the refined corpus. Finally, qualitative assessment of the results from the proposed method by human users could be another way to evaluate this proposed work.

ACKNOWLEDGMENT

Wigrai Thanadechteemapat is supported by a Murdoch University International PhD Scholarship. The support for this work is greatly appreciated.

REFERENCES

[1] 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.

[2] McKie, S. *Scriptclud.com: Content clouds for screenplays*. in *Semantic Media Adaptation and Personalization, Second International Workshop on*. 2007. p. 221-224.

[3] W3Techs, Q-Success Web-based Services. *Usage of content languages for websites, March 2012*. 2012 March 30 [cited 2012 March 30]; Usage of content languages for websites]. Available from: http://w3techs.com/technologies/overview/content_language/all.

[4] ASEAN-Secretariat, *Roadmap for an ASEAN community 2009-2015*, ed. Office, P.A. 2009, Jakarta: ASEAN Secretariat. 128.

[5] Waqar, M. and Z. S. Khan. *Web 2.0 content extraction*. in *Internet Technology and Secured Transactions (ICITST), 2010 International Conference for*. 2010. p. 1-3.

[6] Xunhua, Liu, et al. *On Web page extraction based on position of DIV*. in *Computer and Automation Engineering (ICCAE), 2010 The 2nd International Conference on*. 2010. p. 144-147.

[7] Davi de Castro Reis, et al., *Automatic Web news extraction using tree edit distance*, in *Proceedings of the 13th international conference on World Wide Web*. 2004, ACM: New York, NY, USA. p. 502-511.

[8] Wigrai Thanadechteemapat and Chun Che Fung. *Automatic Web content extraction for generating tag clouds from Thai Web sites*. in *e-Business Engineering (ICEBE), 2011 IEEE 8th International Conference on*. 2011. Beijing, China. p. 85 - 89.

[9] 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.

[10] 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.

[11] *Longman English dictionary online*. 2011 [cited 31 March 2011]; Available from: http://www.ldoceonline.com/dictionary/word_1.

[12] Trask, Larry. *What is a word?* 2004 [cited 30 March 2011]; Available from: http://www.sussex.ac.uk/linguistics/documents/essay_-_what_is_a_word.pdf.

[13] Aroonmanakun, Wirote. *Thoughts on word and sentence segmentation in Thai*. in *the Seventh Symposium on Natural Language Processing*. 2007. Pattaya, Thailand: Citeseer. p. 85-90.

[14] Aroonmanakun, Wirote, *Collocation and Thai word segmentation*, in *the Fifth Symposium on Natural Language Processing & The Fifth Oriental COCOSDA Workshop*. 2002. p. 68--75.

[15] Yuen Poowarawan and Imarrom Wiwat. *Dictionary-based Thai syllable separation*. in *the Ninth Annual Meeting on Electrical Engineering of the Thai Universities*. 1986. Khonkaen, Thailand.

[16] Meknavin, S., et al. *Feature-based Thai word segmentation*. in *the Natural Language Processing Pacific Rim Symposium 1997*. 1997. Phuket, Thailand: Citeseer.

[17] Sutheebanjard, P. and W. Premchaiswadi. *Thai personal named entity extraction without using word segmentation or POS tagging*. in *Natural Language Processing, 2009. SNLP '09. Eighth International Symposium on*. 2009. p. 221-226.

[18] Bheganan, Poramin, et al., *Thai Word segmentation with Hidden Markov Model and Decision Tree*, in *Advances in Knowledge Discovery and Data Mining*. 2009. p. 74-85.

[19] Haruechaiyasak, Choochart, et al. *A comparative study on Thai word segmentation approaches*. in *Electrical Engineering/Electronics, Computer, Telecommunications and Information Technology, 2008. ECTI-CON 2008. 5th International Conference on*. 2008. p. 125-128.

[20] Seifert, C., et al. *On the beauty and usability of tag clouds*. in *Information Visualisation, 2008. IV '08. 12th International Conference*. 2008. p. 17-25.

[21] Hearst, M. A. and D. Rosner. *Tag clouds: Data analysis tool or social signaller?* in *Hawaii International Conference on System Sciences, Proceedings of the 41st Annual*. 2008. p. 160-160.

[22] Pérez García-Plaza, Alberto, et al., *Reorganizing clouds: A study on tag clustering and evaluation*. Expert Systems with Applications, 2012. 39(10): p. 9483-9493.

[23] Wigrai Thanadechteemapat and Chun Che Fung. *Thai word segmentation for visualization of Thai web sites*. in *Machine Learning and Cybernetics (ICMLC), 2011 International Conference on*. 2011. Guilin, China. p. 1544-1549.

[24] Melucci, Massimo and Ricardo Baeza-Yates, *Advanced topics in Information Retrieval*. 2011, Springer: Dordrecht.

[25] Salton, Gerard and Christopher Buckley, *Term-weighting approaches in automatic text retrieval*. Inf. Process. Manage., 1988. 24(5): p. 513-523.

[26] Christopher D. Manning, et al., *Introduction to Information Retrieval*. 2008, Cambridge, England: Cambridge University Press.