

**An Integrated Approach for Content Extraction,
Word Segmentation and Information Presentation from
Thai Websites**

Wigrai Thanadechteemapat

This thesis is presented for the degree of Doctor of Philosophy of

Murdoch University, 2012

Declaration

I declare that this thesis is my own account of my research and contains as its main content work which has not previously been submitted for a degree at any tertiary education institution.

.....
(Wigrai Thanadechteemapat)

Abstract

This thesis presents an integrated approach for the presentation of an overview of key content from Thai websites. This approach is intended to address the information overload issue by presenting an overview to users so that they could assess whether the information meets their needs. This study has proposed rule-based techniques for Web content extraction, and they are capable to extract key content from single and multiple webpages. As there are currently no criteria in assessing the performance of content extraction from Thai websites, this study has proposed evaluation criteria based on the length of the extracted content. Experiment results in this study have demonstrated high accuracy with efficient performance. This study also proposed a Thai word segmentation approach based on the longest matching technique with the utilisation of a corpus to segment Thai words in the extracted key content. The results from the proposed technique have been compared to techniques submitted to the Benchmark for Enhancing the Standard for Thai Language Processing (BEST) contest at Thailand. Results from this work have demonstrated that the performance is consistently better than most of the results from the participants in the contest with an accuracy of between 95 to 97 percent. To select the segmented words for a tag cloud as presentation of the overview, statistical techniques for keyword identification from the key content of single and multiple webpages have been developed, and the techniques are based on the normalisation of the Term Frequency of the keywords. The identified keywords were compared with the key content and tags provided by the websites,

and the accuracy of the results was higher than the outputs obtained from the Term Frequency and Inverse Document Frequency (TFIDF) and Term Length Term Frequency (TLTF) techniques. The proposed techniques were evaluated based on *Precision, Recall* and *F-measure*. A Variable Tag Cloud approach has also been developed in order to provide the overview to the users with flexibility and user-determined number of keywords in the tag cloud. The approach is novel and it is believed that the findings in this research will benefit the Thai community and encourage more efficient access of information from Thai Websites.

Acknowledgments

I am grateful to have arrived at this stage of the research, where I have learnt how to discover and make contribution to knowledge. This only happened because I have received much support and encouragement from many kind and intelligent people whom I am forever indebted to.

First of all, I would like to express my sincere gratitude to my supervisor, Associate Professor Dr Lance Chun Che Fung, who has been most helpful, and he had offered much invaluable advice throughout my study. I personally feel like he is my parental guidance in supervising and advising me in every aspect of my stay here. He was always available, and he went many extra miles to assist me at any time including days, nights and weekends. I am extremely fortunate having met and worked under him. I hope my professional association with him will continue into the future.

I am also deeply thankful to my co-supervisor, Associate Professor Dr Kevin Kok Wai Wong, who provided me with useful support and guidance as well as many insightful questions during this research. In addition, I would like to thank Professor David Harries for his helpful support at the beginning of my study.

I greatly appreciate the support given by Murdoch University through the Murdoch International PhD Scholarship for the whole period of my study. In addition, I would also like to thank my homeland, Thailand, for the beautiful language, and the

National Electronics and Computer Technology Centre (NECTEC), for providing the Thai language data sets and results from the BEST competition, which I have used in this research.

Also, I am extremely thankful to Mrs Swee Lin Tan, who kindly spent her valuable time to read through this thesis as well as correcting the language in it at short notice. I also thank A.K. and Sau Chee Ch'ng for providing comfortable accommodation during my stay in Perth.

Finally, I would like to express my greatest appreciation to my mum and dad, although the later passed away during my study. Special thanks to my wife, my sister, and other friends for their support including past teachers for giving me the strong education foundation, and everyone who have encouraged me until the completion of this thesis. Last but not least, I wish to respectfully acknowledge the teachings of Gautama Buddha which had helped me keep calm and focused during difficult times when doing this thesis.

List of Publications

A total of ELEVEN publications have been published based on findings in this study. Most of the papers were included in proceedings of international conferences hosted by IEEE, while the rest were published by conferences with various entities such as IET and local universities. The list of publications is given below.

- (1) W. Thanadechteemapat and C. C. Fung, "Automatic content extraction and visualization of Thai websites for improved information representation." in the *Proceedings of the 2012 IEEE International Conference on Systems, Man, and Cybernetics, IEEE SMC 2012*, Seoul, Korea, October 14-17, 2012. pp. 2229-2234.
- (2) W. Thanadechteemapat and C. C. Fung, "Improving webpage content extraction by extending a novel Single Page Extraction approach: A case study with Thai websites." in the *Proceedings of the 11th International Conference on Machine Learning and Cybernetics, ICMLC 2012*, Xi'an, China, July 15-17, 2012. pp. 1263-1267.
- (3) W. Thanadechteemapat and C. C. Fung, "Automatic web content extraction for generating tag clouds from Thai Web sites", in the *Proceedings of the 8th IEEE International Conference on e-Business Engineering, IEEE ICEBE 2011*, Beijing, China, October 19-21, 2011. pp. 85-89.
- (4) W. Thanadechteemapat and C. C. Fung, "Thai word segmentation for visualization of Thai Web sites", in the *Proceedings of the 10th International*

Conference on Machine Learning and Cybernetics, ICMLC 2011, Guilin, China, July 10-13, 2011. pp. 1544-1549.

- (5) C. C. Fung and W. Thanadechteemapat, "Discover information and knowledge from websites using an integrated summarization and visualization framework", in the *Proceedings of the Third International Conference on Knowledge Discovery and Data Mining 2010, WKDD 2010, Phuket, Thailand, January 9-10, 2010. pp. 232-235.*
- (6) C. C. Fung, W. Thanadechteemapat, and K. P. Wong, "Summarizing information from Web sites on distributed power generation and alternative energy development", in the *Proceedings of the 8th IET International Conference on Advances in Power System Control, Operation and Management, APSCOM 2009, Kowloon Shangri La Hotel, Hong Kong, November 8–11, 2009. pp. 1-6.*
- (7) W. Thanadechteemapat and C. C. Fung, "A Web assessment approach based on summarisation and visualisation", in the *Proceedings of the 10th Postgraduate Electrical Engineering & Computing Symposium, PEECS 2009, ECU, Western Australia, October 1, 2009.*
- (8) C. C. Fung, W. Thanadechteemapat, and D. Harries, "Acquiring knowledge and information on alternative energy from the World Wide Web", in the *General Meeting: Proceedings of the 2009 IEEE Power Engineering Society, IEEE PES 2009, Calgary, Alberta, Canada, July 26-30, 2009. pp. 1-9.*
- (9) C. C. Fung, W. Thanadechteemapat, and K. P. Wong, "iWISE, An intelligent Web interactive summarization engine", in the *Proceedings of the 8th*

International Conference on Machine Learning and Cybernetics, ICMLC 2009, Baoding, Hebei, China, July 12-15, 2009. pp. 3457-3462.

- (10) W. Thanadechteemapat and C. C. Fung, "A study on the deployment of Web technologies by business websites on sustainable energy", in the *Proceedings of the 7th International Conference on e-Business, INCEB 2008*, Bangkok, Thailand, November 6-7, 2008.
- (11) W. Thanadechteemapat and C. C. Fung, "A survey on the use of Web technologies in the promotion of sustainable energy", in the *Proceedings of the 9th Postgraduate Electrical Engineering & Computing Symposium, PEECS 2008*, UWA, Western Australia, November 4, 2008.

Summary of Contributions

The following are the main contributions from this study. References have been made with respect to the Chapter, Section and Paper number.

- Investigation and development of Web content extraction techniques that are able to extract Web content in Thai language efficiently from both single page and multiple pages. [Chapter 3] (2, 3)
- Proposal and development of evaluation criteria for content extraction from single webpage and the criteria are applicable to multiple pages. [Chapter 3] (2, 3)
- Addressed the problem of Thai word segmentation based on the longest matching technique with a refined corpus instead of a dictionary. [Chapter 4] (4)
- Development and proposal of techniques for keyword identification from single and multiple webpages based on a normalisation of Term Frequency of the keywords [Chapter 5] (1)
- Automatic generation of Thai website information overview from single and multiple pages based on a proposed Variable Tag Cloud approach. [Chapter 5] (1, 9)
- Background literature survey and proposal of the framework have been carried out and reported. [Chapter 2] (5, 6, 7, 8, 10, 11)

Table of Contents

Abstract	I
Acknowledgments.....	III
List of Publications	V
Summary of Contributions	VIII
Table of Contents	IX
List of Tables.....	XIV
List of Figures	XVI
Table of Abbreviations	XIX
1. Introduction.....	1
1.1 Motivation.....	1
1.2 Proposal	7
1.2.1 Web Content Extraction	9
1.2.1.1 A Proposed Technique for Single Page Extraction	9
1.2.1.2 A Proposed Technique of Multiple Page Extraction.....	9
1.2.1.3 Experimental Data Set and Evaluation	10
1.2.2 Thai Word Segmentation	10
1.2.3 Keyword Identification	11
1.2.4 Information Presentation with a Variable Tag Cloud.....	11
1.3 Thesis Organisation.....	12
2. Background.....	15

2.1	Introduction	15
2.2	Web Content Extraction	16
2.2.1	Design and Layout of Webpages.....	18
2.2.2	Related Terminologies on the Web	21
2.2.3	Web Content Extraction Techniques	24
2.2.3.1	Single Page Extraction	25
2.2.3.2	Evaluation.....	29
2.2.3.3	Multiple Page Extraction	30
2.3	Thai Word Segmentation.....	32
2.3.1	Definition of Words.....	32
2.3.2	Word Segmentation Techniques	34
2.4	Keyword Identification	38
2.5	Information Presentation	41
2.5.1	Applications of Data Visualisation	42
2.5.2	Information Presentation based on Tag Cloud	46
3.	Web Content Extraction.....	50
3.1	Introduction	50
3.2	Single Page Extraction	51
3.2.1	Webpage Element and Feature Extraction	52
3.2.2	Block Detection	54
3.2.3	Selection of Elements in Blocks for Content Extraction	55
3.2.4	Evaluation.....	56

3.2.5	Experimental Results and Discussion	57
3.2.6	Summary of the Proposed Single Page Extraction Technique	62
3.3	Multiple Page Extraction.....	63
3.3.1	Crawler	66
3.3.2	Applying Single Page Extraction	67
3.3.3	Extracted Content Matching (ECM)	67
3.3.4	Evaluation.....	68
3.3.5	Experimental Results and Discussion	68
3.3.6	Summary of Multiple Page Extraction	73
3.4	Summary.....	74
4.	Thai Word Segmentation	76
4.1	Introduction	76
4.2	Corpus Details	78
4.2.1	Description of Corpus.....	78
4.2.2	Corpus Preparation	83
4.2.3	Resolving Inconsistent Segmentation	86
4.3	Thai Word Segmentation Technique	90
4.4	Evaluation of Word Segmentation Results.....	92
4.5	Experimental Results and Discussion	93
4.6	Summary	97
5.	Information Presentation with a Variable Tag Cloud.....	99
5.1	Introduction	99

5.2	Keyword Identification from Single Page	100
5.2.1	Keyword Identification Technique	101
5.2.2	Evaluation.....	104
5.2.3	Experimental Results and Discussion.....	107
5.3	Variable Tag Cloud for Single Page	111
5.3.1	Display of Keywords in a Variable Tag Cloud	112
5.3.2	Visualisation of the Variable Tag Cloud	114
5.4	Keyword Identification from Multiple Pages.....	124
5.4.1	Keyword Identification Technique	124
5.4.2	Evaluation.....	125
5.4.3	Experimental Results and Discussion.....	126
5.5	Variable Tag Cloud for Multiple Pages	130
5.6	Summary.....	133
6.	Conclusion and Future Work	136
6.1	Conclusion.....	136
6.1.1	Web Content Extraction.....	137
6.1.2	Thai Word Segmentation	137
6.1.3	Information Presentation with a Variable Tag Cloud	138
6.2	Future Work.....	140
	References	142
	Appendices.....	155

Appendix A. Results from Single Page Extraction in Chapter 3 (Section 3.2.6)	155
Appendix B. Results from Multiple Page Extraction in Chapter 3 (Section 3.3.6)	156
Appendix C. Compared Results of the Use of a Dictionary and the Corpus with the Proposed Technique of Thai Word Segmentation in Chapter 4 (Section 4.1).....	158
Appendix D. Detailed Results of Keyword Identification from Single Webpage in Chapter 5 (Section 5.2.3)	158
Appendix E. Detailed Results of Keyword Identification from Multiple Webpages in Chapter 5 (Section 5.4.3).....	161
Appendix F. Variable Tag Cloud for Multiple Pages from Sanook! in Chapter 5 (Section 5.5)	166

List of Tables

Table 3.1. Example results of Single Page Extraction from three pages within a website with the same template	58
Table 3.2. Results of Single Page Extraction from three different websites	59
Table 3.3 Experimental results of Single and Multiple Page Extraction	70
Table 4.1. Types of annotations in the BEST 2009 word-segmented corpus	81
Table 4.2. Samples of inconsistent segmented words in the corpus	88
Table 4.3. Improvement on the resolving inconsistent segmented words	89
Table 4.4. Results on academic article data set	94
Table 4.5. Results on encyclopaedia data set	94
Table 4.6. Results on news data set	95
Table 4.7. Results on novel data set	95
Table 5.1. Statistics of the data set used to evaluate the proposed keyword identification technique from single page	108
Table 5.2. A comparison of the best results from AMTF, TLTF and TFIDF techniques	109
Table 5.3. 7 styles of tags in a tag cloud	113
Table 5.4. The number of identified keywords in different percentage of their weights in a webpage from Sanook!	116
Table 5.5. Statistics of the data set used to evaluate the proposed keyword identification technique from multiple webpages	127

Table 5.6. The best results produced by KID, TFIDF and TLTF for Sanook! and MThai website based on keyword identification from multiple webpages	128
Table 5.7. The number of identified keywords in different percentage of their weights from multiple pages in MThai.....	131

List of Figures

Figure 1.1. The number of websites from August 1995 to March 2011	2
Figure 1.2. Estimation of information generated and hosted from 2006 to 2020 by IDC published by The Economist	3
Figure 1.3. An overview of the integrated approach for information presentation with a Variable Tag Cloud	8
Figure 2.1. Examples of the same design used in a Thai news website, Dailynews.....	19
Figure 2.2. A layout of the design in Figure 2.1	19
Figure 2.3. An example of two different designs in the same website, OK nation.....	21
Figure 2.4. An example of a section of HTML code	23
Figure 2.5. The DOM tree of the HTML code in Figure 2.4.....	23
Figure 2.6. An example Treemap obtained from marumushi.com	44
Figure 2.7. An example tag cloud to compare 2002 State of the Union Address by U.S. President Bush with 2011 State of the Union Address by President Obama	45
Figure 3.1. An overview of the proposed technique of Single Page Extraction	52
Figure 3.2. An example of Thai webpage from Thai news, Matichon Online	61
Figure 3.3. Extracted Thai web content from the proposed technique of Single Page Extraction	62
Figure 3.4. An overview of the proposed Multiple Page Extraction approach	65

Figure 3.5. Examples of URL No. 4 and 5 from Table 3.3 and areas of informative and non-informative extracted content	73
Figure 4.1. An example excerpt from the corpus provided by the BEST project	80
Figure 4.2. Corpus preparation process	84
Figure 4.3. Process of resolving inconsistent segmentation	86
Figure 4.4. Main algorithm of the corpus refinement	87
Figure 5.1. The overall approach of information presentation from Thai websites	101
Figure 5.2. An illustration of calculating the Precision and Recall from outputs of the keyword identification process	106
Figure 5.3. Content area of the example webpage from Sanook! website	116
Figure 5.4. A line chart showing the number of identified keywords in different percentage weights	117
Figure 5.5. An illustration of Variable Tag Clouds generated based on different percentage weights	120
Figure 5.6. Comparison of Thai tag cloud, actual content and defined tags	122
Figure 5.7. An illustration of Variable Tag Clouds from multiple pages in MThai	132
Figure 5.8. A Variable Tag Cloud at 5% weight from multiple pages in MThai	133
Figure A. 1 An illustration of Variable Tag Clouds from multiple pages in Sanook!	166
Figure A. 2. A Variable Tag Cloud at 5% weight from multiple pages in Sanook!	167

Figure A. 3. A Variable Tag Cloud at 10% weight from multiple pages in

Sanook! 167

Figure A. 4. . A Variable Tag Cloud at 15% weight from multiple pages in

Sanook! 167

Table of Abbreviations

Abbreviation	Original phrase or terms
AList	An Additional List of keywords (used in Variable Tag Cloud generation)
All Elements	Number of all the element nodes (used in Web content extraction)
All Href	Number of all the anchor element nodes (used in Web content extraction)
All Href T-Length	Length of the characters in all anchor element nodes (used in Web content extraction)
All T- Length	Length of the characters in all element nodes (used in Web content extraction)
AMTF	Average Normalised Term Frequency (used in keyword identification)
ANR	Anchor Node Ratio (used in Web content extraction)
AR	Anchor Ratio (used in Web content extraction)
ASEAN	Association of Southeast Asian Nations
ATR	Anchor Text Ratio (used in Web content extraction)
BEST	Benchmark for Enhancing the Standard for Thai Language Processing
CK	The number of correct identified keywords (used in keyword identification)
CMain	The BEST 2009 word-segmented corpus (used in Thai word segmentation)

Abbreviation	Original phrase or terms
CMS	Content Management Systems
CName	A collection of special words (used in Thai word segmentation)
CSS	Cascading Style Sheets
CTest20	The 20% of the files in the BEST Corpus (used in Thai word segmentation)
DK	The number of occurrence of the correct identified keywords in the tags (used in keyword identification)
ECM	Extracted Content Matching (used in Web content extraction)
ELists	Keyword lists from each webpage (used in keyword identification)
HTML	Hypertext Markup Language
HTTP	Hypertext Transfer (or Transport) Protocol
ICT	Information and Communications Technology
IDC	International Data Corporation
IDF	Inverse Document Frequency
KID	Keyword Identification based on Individual Documents (used in keyword identification)
LEXP	The Length of the EXpected content (used in Web content extraction)
LEXT	The Length of the EXtracted Content (used in Web content extraction)
LM	The Length of Missing relevant content (used in Web

Abbreviation	Original phrase or terms
	content extraction)
NAiST	Natural Language Processing and Intelligent Information System Technology
NECTEC	The National Electronics and Computer Technology Center
NI	The number of segmented words in the tags Not being Included in the content (used in keyword identification)
NLP	Natural Language Processing
PCCS	The practical colour coordinate system
SWT	The number of Segmented Words in the Tags provided in a webpage (used in keyword identification)
TF	Term Frequency
TFIDF	Term Frequency and Inverse Document Frequency
TK	The total number of identified keywords (used in keyword identification)
TLTF	Term Length Term Frequency
URL	Uniform (or universal) Resource Locator
WWW	World Wide Web
XPath	XML Path language

