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

List of Tables

List of Figures

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