School of Information Technology

A Frequent Max Substring Technique for Thai Text Indexing

Todsanai Chumwatana

This thesis is presented for the Degree of

Doctor of Philosophy of

Murdoch University

May 2011

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.

Todsanai Chumwatana

May, 2011

Acknowledgments

I would like to take this opportunity to acknowledge and thank the following people and organizations that helped me to complete this thesis.

First of all I would like to thank my principal supervisor, Associate Professor Dr Kevin Wong, for his support, guidance, comments and encouragement throughout the period of my PhD research. From the first day of my study, Professor Kevin Wong always offered me opportunities and taught me how to do good research. Throughout my study, Professor Kevin Wong always walked beside me and supported me. I am very grateful for his helpful advice and his efforts to explain things to me.

I would also like to thank my associate supervisor, Dr Hong Xie, for his suggestions. Besides my principal supervisor and associate supervisor, I wish to express my gratitude to Associate Professor Dr Lance Chun Che Fung. Professor Fung has always offered me and his students support, opportunities and mental stimulation. Many thanks to my PhD colleagues and Thai students who have supported me and given me wonderful friendships. Thanks to all of them for a memorable time.

My research would not have become a reality without the financial support of my family during the period of my study. I cannot forget to thank my lovely family: my father Ponlasak Chumwatana, my mother Tassanee Kaewbavon and my two older sisters Nattaya Mohjhaw and Chalakod Chumwatana for everything. They have always supported and encouraged me and lifted my spirits since I opened my eyes to see the world.

Finally, I would like to thank the School of Information Technology, Murdoch University, for providing me with all the necessary facilities for my research.

Abstract

This research details the development of a novel methodology, called the frequent max substring technique, for extracting indexing terms and constructing an index for Thai text documents.

With the rapidly increasing number of Thai digital documents available in digital media and websites, it is important to find an efficient Thai text indexing technique to facilitate search and retrieval. An efficient index would speed up the response time and improve the accessibility of the documents. Up to now, not much research in Thai text indexing has been conducted as compared to more commonly used languages like English. The more commonly used Thai text indexing technique is the word inverted index, which is language-dependent (i.e. requires linguistic knowledge). This technique creates word document indices on document collection to enable an efficient keyword based search. However, when using the word inverted index technique, Thai text documents need to be parsed and tokenized into individual words first. Therefore, one of the main issues is how to automatically identify the indexing terms from the Thai text documents before constructing the index. This is because the syntax of Thai language is highly ambiguous and Thai language is non-segmented (i.e. a text document is written continuously as a sequence of characters without explicit word boundary delimiters). To index Thai text documents, most language-dependent indexing techniques have to rely on the performance of a word segmentation approach in order to extract the indexing terms before constructing the index. However, word segmentation is time consuming and segmentation accuracy is heavily dependent either on the linguistic knowledge used in the underlying segmentation algorithms, or on the

dictionary or corpus used in the segmentation. It is for this reason that most language-dependent indexing techniques are time consuming and require additional storage space for storing dictionary or corpus or manually crafted rules resource.

Apart from the language dependant indexing techniques, some language-independent techniques have been proposed as an alternative indexing technique for Thai language such as the n-gram inverted index and suffix array approaches. These approaches are simple and fast as they are language-independent, and do not require linguistic knowledge of the language, or the use of a dictionary or a corpus. However, the limitation of these techniques is that they require more storage space for extracting the indexing terms and constructing the index.

To address the above limitations, this thesis has developed a frequent max substring technique that uses language-independent text representation, which is computationally efficient and requires small storage place. The frequent max substring technique improves the performance in terms of construction time over the language-dependent techniques (i.e. the word inverted index) as this technique does not require text pre-processing tasks (i.e. word segmentation) in extracting the indexing terms before indexing can be performed. This technique also improves space efficiency compared to some existing language-independent techniques. This is achieved by retaining only the frequent max substrings, which are strings that are both long and frequently occurring, in order to reduce the number of insignificant indexing terms from an index.

To demonstrate that the frequent max substring technique could deliver its performance, experimental studies and comparison results on indexing Thai text documents are presented in this thesis. The technique was evaluated and compared in term of indexing

efficiency and retrieval performance. The results show that the frequent max substring technique is more computationally efficient when compared to the word inverted index, and also that it requires less space for indexing when compared to some language-independent techniques.

Additionally, this thesis shows that the frequent max substring technique has an advantage in terms of versatility, as it can also be combined with other Thai language-dependent techniques to become a novel hybrid language-dependent technique, in order to further improve the indexing quality. This technique can also be used with a neural network to enhance non-segmented document clustering. The frequent max substring technique also has the flexibility to be applied to other non-segmented texts like the Chinese language and genome sequences in bioinformatics due to its language-independency feature.

List of Publications Related to this Thesis

P1. T. Chumwatana, K. W. Wong and H. Xie, 'Using Frequent Max Substring Technique for Thai Text Indexing', accepted for publication in the *Australian Journal of Intelligent Information Processing Systems (AJIIPS).*

P2. T. Chumwatana, K. W. Wong and H. Xie, 'A SOM-Based Document Clustering Using Frequent Max Substrings for Non-Segmented Texts', *In the Journal of Intelligent Learning Systems and Applications (JILSA),* 2010.

P3. T. Chumwatana, K. W. Wong and H. Xie, 'Using Frequent Max Substring Technique for Thai Keyword Extraction used in Thai Text Mining', *In Proceedings of the 2nd International Conference on Soft Computing, Intelligent System and Information Technology (ICSIIT 2010),* Bali, Indonesia, 1-2 July 2010.

P4. T. Chumwatana, K. W. Wong and H. Xie, 'Non-segmented Document Clustering Using Self-organizing map and Frequent Max Substring Technique', Lecture Notes in Computer Science, Springer Berlin/Heidelberg, LNCS 5864, 2009, pp. 691–698.

P5. T. Chumwatana, K. W. Wong and H. Xie, 'Non-segmented Document Clustering Using Self-organizing map and Frequent Max Substring Technique', *In16th International Conference on Neural Information Processing (ICONIP 2009),* Bangkok, Thailand, 2009.

P6. T. Chumwatana, K. W. Wong and H. Xie, 'Indexing Non-Segmented Texts using n-Gram Inverted Index and Frequent Max Substring: A Comparison of Two Techniques,' *In 10th Postgraduate Electrical Engineering & Computing Symposium (PEECS2009),* Perth, Australia, 2009.

P7. T. Chumwatana, K. W. Wong and H. Xie, 'An Automatic Indexing Technique for Thai Texts using Frequent Max Substring,' *In The 8th International Symposium on Natural Language Processing, 2009 (SNLP '09),* Bangkok, Thailand, 2009.

P8. T. Chumwatana, K. W. Wong and H. Xie, 'An Efficient Text Mining Technique', *In 9th Postgraduate Electrical Engineering & Computing Symposium (PEECS2008),* Perth, Australia, 2008.

P9. T. Chumwatana, K. W. Wong and H. Xie, 'Thai Text Mining to Support Web Search for E-commerce', *In The 7th International Conference on e-Business 2008 (INCEB2008),* Bangkok, Thailand, 2008.

P10. T. Chumwatana, K. W. Wong and H. Xie, 'Frequent Max Substring Mining for Indexing', *International Journal of Computer Science and System Analysis (IJCSSA), India,* 2008.

P11. T. Chumwatana, K. W. Wong and H. Xie, 'Frequent Max Substring Mining', *In 8th Postgraduate Electrical Engineering & Computing Symposium (PEECS2007),* Perth, Australia, 2007.

Contributions of this Thesis

The contributions in this thesis which have already been published and reported are described below and summarized in Table 1.

A survey and review of various techniques in the Thai text indexing area has been completed. This work forms the basis of Chapter 2. Parts of this work have been published in conference papers P3, P7, P9 and journal paper P1.

The development of the novel indexing technique, called the frequent max substring technique, forms a part of Chapter 3. Results from this work are reported in conference papers P3, P7, P8, P9, P11 and journal papers P1 and P10. Some of these publications also include the experimental studies, comparison results and discussion on indexing Thai text documents. Another contribution documented in Chapter 3 is the establishment of a methodology for evaluating the frequent max substring technique by comparing it to other indexing techniques such as the word inverted index, n-gram inverted index, Vilo's technique and suffix array. The comparison was based on indexing efficiency and retrieval performance.

The contribution in Chapter 4 is the establishment of the integration of the frequent max substring technique with other Thai language-dependent techniques to create a novel language-dependent technique. The hybrid method is used for extracting and indexing meaningful indexing terms from Thai text documents. Parts of this chapter have been published in journal paper P1.

The work on an integrated method using the frequent max substring technique with self-organizing map (SOM) for non-segmented document clustering is published in conference paper P5 and lecture notes in computer science paper P4. Conference paper P5 was later extended to journal paper P2, which has been described in Chapter 5. Journal paper P2 was published in *Journal of Intelligent Learning Systems and Applications* and it showed that the frequent max substring technique can be used with self-organizing map to enhance non-segmented document clustering in order to improve the efficiency of Thai information retrieval.

Chapter 6 discussed the application of the frequent max substring technique to other non-segmented texts such as non-segmented languages (the Chinese language) and genome sequences. This demonstrated that the frequent max substring technique is versatile as it can be used not only for Thai text indexing but also for indexing other non-segmented texts. This work is reported in conference paper P6. The paper also presents some comparison results and discussion on indexing non-segmented texts.

Table 1 Summary of the Contribution of the Thesis

Chapter	Contributions	Paper No
***Chapter 2*: Thai Text Indexing**	Presents a literature survey on previous research in the Thai text indexing area and identifies the limitations of existing Thai text indexing techniques.	P1, P3, P7, P9
***Chapter 3*: Frequent Max Substring Technique**	Successfully developed the frequent max substring technique to perform Thai text indexing for language-independent technique.	P1, P3, P7, P8, P9, P10, P11
***Chapter 4*: Hybrid Method: Integration of the Frequent Max Substring Technique and Thai Language-Dependent Technique**	Successfully developed a hybrid method by combining the frequent max substring technique and language-dependent technique to perform Thai text indexing using linguistic knowledge.	P1
***Chapter 5*: Non-Segmented Document Clustering Using Self-Organizing Map and the Frequent Max Substring Technique**	Developed an integrated method using the frequent max substring technique with self-organizing map (SOM) to enhance the non-segmented document clustering.	P2, P4, P5
***Chapter 6*: Non-Segmented Text Problems**	Successfully implemented the proposed technique with some other non-segmented texts like Chinese and genome sequence.	P6

Contents

List of Figures

List of Tables