

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 (ICSIT 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', *In 16th 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
<i>Chapter 2: Thai Text Indexing</i>	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
<i>Chapter 3: Frequent Max Substring Technique</i>	Successfully developed the frequent max substring technique to perform Thai text indexing for language-independent technique.	P1, P3, P7, P8, P9, P10, P11
<i>Chapter 4: Hybrid Method: Integration of the Frequent Max Substring Technique and Thai Language-Dependent Technique</i>	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
<i>Chapter 5: Non-Segmented Document Clustering Using Self-Organizing Map and the Frequent Max Substring Technique</i>	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
<i>Chapter 6: Non-Segmented Text Problems</i>	Successfully implemented the proposed technique with some other non-segmented texts like Chinese and genome sequence.	P6

Contents

Declaration	i
Acknowledgments	ii
Abstract	iv
List of Publications Related to this Thesis	vii
Contributions of this Thesis	viii
List of Figures	xv
List of Tables	xx
Chapter 1: Introduction and Overview	1
1.1 Overview	1
1.2 Objectives	8
1.3 Contributions	9
1.4 Thesis outline	11
Chapter 2: Thai Text Indexing	14
2.1 Introduction	14
2.2 Overview of Thai text indexing	16
2.3 Linguistic characteristics of the Thai language	18
2.4 Thai text indexing techniques	22
2.4.1 Using language-dependent method for Thai text indexing	23
2.4.1.1 Inverted index construction	26
2.4.1.2 Thai word segmentation	28
2.4.1.3 Thai stopword removal	48
2.4.2 Using language-independent method for Thai text indexing	50
2.4.2.1 An n-gram inverted index	50
2.4.2.2 Suffix array approach	56
2.5 The retrieval process	60
2.5.1 A query	61

2.5.2 Query processing	65
2.5.3 Searching	65
2.5.3.1 Search using inverted index	66
2.6 Limitations of Thai text indexing techniques	67
2.7 Conclusion	72
Chapter 3: Frequent Max Substring Technique	74
3.1 Introduction	74
3.2 Substring indexing based on suffixes	77
3.2.1 Suffix trie	81
3.2.2 Suffix tree	85
3.2.3 Suffix array	88
3.2.4 The comparison of suffix trie, suffix tree and suffix array	91
3.3 Frequent substring indexing with Vilo's method	93
3.3.1 Frequent substring indexing based on Vilo's technique	95
3.4 Frequent max substring technique	103
3.4.1 Frequent suffix trie structure or FST structure	106
3.4.1.1 The frequent suffix trie construction	108
3.4.2 Algorithms	109
3.5 Experimental studies	124
3.5.1 Text collection	125
3.5.2 Evaluation of indexing	126
3.5.2.1 Space efficiency	127
3.5.2.2 Time efficiency	134
3.5.3 Evaluation of retrieval performance	138
3.6 Conclusion	150
Chapter 4: Hybrid Method: Integration of the Frequent Max Substring Technique and Thai Language-Dependent Technique	154
4.1 Introduction	154
4.2 Related works	154
4.3 Hybrid method	156
4.4 Experimental studies	163
4.5 Conclusion	170

Chapter 5: Non-Segmented Document Clustering Using Self-Organizing Map and the Frequent Max Substring Technique	172
5.1 Introduction	172
5.2 Document clustering	173
5.3 Keyword extraction	174
5.4 Document clustering algorithms and related works	176
5.5 SOM based clustering using the frequent max substring technique for non-segmented texts	179
5.6 Experimental studies and comparison results	184
5.7 Conclusion	190
Chapter 6: Non-Segmented Text Problems	191
6.1 Introduction	191
6.2 Non-segmented language problems	191
6.2.1 Characteristics of the Chinese language	192
6.2.2 Related works	194
6.2.3 Applying the frequent max substring technique to the Chinese language	196
6.2.4 Experimental results	197
6.3 Genome sequencing	203
6.3.1 Characteristics of the genome sequence	204
6.3.2 Related works	206
6.3.3 Applying the frequent max substring technique to genome sequencing	208
6.3.4 Experimental studies and comparison results	213
6.4 Conclusion	221
Chapter 7: Conclusions	222
7.1 Contribution and outcomes	222
7.2 Future work and directions	228
Appendixes	229
Appendix A: Addresses of Thai text collection	229

Appendix B: Details of Thai text collection	232
Appendix C: Comparison of number of indexing terms extracted from five indexing techniques	234
Appendix D: Comparison of index sizes used by five indexing techniques	236
Appendix E: Comparison of indexing times used by five indexing techniques	238
Appendix F: Comparison of number of indexing terms extracted from Vilo's method (contracted form 'Vilo') and the frequent max substring technique (contracted form 'FM') at given frequency threshold values between 2 and 10	240
List of References	243

List of Figures

Figure 1.1.	Some drawbacks of existing Thai text indexing techniques	7
Figure 2.1.	A general indexing and retrieval system (numbers beside each dash-line box indicate sections that cover corresponding topic)	14
Figure 2.2.	Example of the Thai language	18
Figure 2.3.	Thai character set	19
Figure 2.4.	Four levels of appearance of Thai characters	20
Figure 2.5.	Existing Thai text indexing techniques (numbers beside each box indicate sections that cover corresponding topic)	22
Figure 2.6.	General process of indexing Thai text documents using word inverted index (numbers beside each box indicate sections that describe corresponding topic)	24
Figure 2.7.	Example of the word inverted index	25
Figure 2.8.	Example of vocabulary and posting file of document containing the string s ‘การประกอบการ’	26
Figure 2.9.	Illustration of building word inverted index for documents containing the string s ‘การประกอบการ’	28
Figure 2.10.	Different ways to insert word separators in Thai texts	30
Figure 2.11.	BNF of Thai word formation rules	32
Figure 2.12.	Types of Thai characters	35
Figure 2.13.	All rules for TCC Segmentation	36
Figure 2.14.	Example of Thai character cluster compared with correct segmentation	37
Figure 2.15.	Result of longest match for string s ‘เจมส์เป็นเด็กตากลมที่ต้องการนำโคลงเรือ’	38
Figure 2.16.	Segmentation of: ‘การวางแผนไทยเป็นการกายภาพบำบัดเพื่อรักษาคนไข้อย่างหนึ่ง’	40
Figure 2.17.	Illustration of maximum matching algorithm procedure	41
Figure 2.18.	Example of string of characters tagged as word-beginning (B) or intra-word (I) characters	46

Figure 2.19.	Character types for building a feature set used by machine learning approach	47
Figure 2.20.	General process of indexing Thai text documents using n-gram inverted index	51
Figure 2.21.	Sets of 1-gram, 2-gram, 3-gram, ..., N-gram overlap sequence of document d containing the string s ‘การประกอบกร’	53
Figure 2.22.	Example of building n-gram inverted index for documents containing the string s ‘การประกอบกร’	55
Figure 2.23.	Illustration of suffix array from string $s =$ ‘การประกอบกร’	57
Figure 2.24.	Illustration of a suffix array from Figure 2.23, which has been sorted in alphabetical order	58
Figure 2.25.	All longest common prefixes with their length and term frequency from suffix array	59
Figure 2.26.	General retrieval system (numbers beside each box indicate sections that cover corresponding topic)	60
Figure 3.1.	All suffixes of string s ‘positivelives\$’	81
Figure 3.2.	Suffix trie of string $s =$ ‘positivelives\$’	83
Figure 3.3.	Example of suffix tree of string s ‘positivelives\$’	86
Figure 3.4.	Suffix tree of string s ‘positivelives\$’ represented by a pair of integers denoting starting and ending positions	87
Figure 3.5.	Illustration of suffix array from string $s =$ ‘positivelives\$’	89
Figure 3.6.	Illustration of suffix array from Figure 3.5, sorted in alphabetical order	90
Figure 3.7.	Discovering frequent substrings of string $s =$ ‘positivelives\$’ having at least two occurrences in string s . Nodes generated into trie represent substrings λ , e, i ,s, v, iv, ve and ive	99
Figure 3.8.	Frequent suffix trie structure for string $s =$ ‘positivelives\$’	109
Figure 3.9.	Frequent suffix trie structure using proposed algorithm	115
Figure 3.10.	Frequent suffix trie structure for string $s =$ ‘การประกอบกร\$’	117
Figure 3.11.	Frequent suffix trie structure using proposed algorithm for string $s =$ ‘การประกอบกร\$’	120

Figure 3.12. Example of indexing multiple documents using frequent max substring technique	122
Figure 3.13. Graph showing number of indexing terms extracted from two techniques: suffix array and proposed frequent max substring technique	127
Figure 3.14. Graph showing number of indexing terms extracted from two techniques: Vilo's technique and proposed frequent max substring technique	128
Figure 3.15. Graph showing number of indexing terms extracted from three techniques: word inverted index, 3-gram inverted index and proposed frequent max substring technique	129
Figure 3.16. Comparison of index size from two techniques: suffix array and proposed frequent max substring technique	131
Figure 3.17. Comparison of index size from two techniques: Vilo's technique and proposed frequent max substring technique	132
Figure 3.18. Comparison of index size from three techniques: word inverted index, 3-gram inverted index and proposed frequent max substring technique	132
Figure 3.19. Comparison of indexing time of two techniques: word inverted index and proposed frequent max substring technique	134
Figure 3.20. Comparison of indexing time of two techniques: 3-gram inverted index and proposed frequent max substring technique	134
Figure 3.21. Comparison of indexing time of two techniques: suffix array and proposed frequent max substring technique	135
Figure 3.22. Comparison of indexing time of two techniques: Vilo's technique and proposed frequent max substring technique	135
Figure 3.23. Precision and recall for given example information request	139
Figure 3.24. Example of querying text collection by using a phrase query in word inverted index technique	147
Figure 3.25. Example of querying text collection in 3-gram inverted index technique	148
Figure 3.26. Example of querying text collection using exact phrase query in frequent max substring technique	149

Figure 3.27.	Example of querying text collection using a segmented query in frequent max substring technique	150
Figure 4.1.	A system architecture	157
Figure 4.2.	Sample of training corpus which is POS tagged text	161
Figure 4.3.	Example of POS tagging of frequent max substrings	162
Figure 4.4.	Number of indexing terms extracted from two language-dependent techniques	167
Figure 5.1.	Example of document vectors in 3-dimension	175
Figure 5.2.	Example of document matrix at given frequency threshold value θ is equal to 2	181
Figure 5.3.	Document cluster map	182
Figure 5.4.	Neuron network architecture	183
Figure 5.5.	Self-organizing map	184
Figure 5.6.	SOM contains nine neurons and a group of similar documents from collection of 50 Thai documents	185
Figure 6.1.	Example of Chinese texts	193
Figure 6.2.	FST structure using frequent max substring technique on Chinese text documents	197
Figure 6.3.	Example of bi-gram terms from document d	199
Figure 6.4.	Comparison of number of indexing terms extracted from bi-gram based indexing and frequent max substring technique	201
Figure 6.5.	Example of nucleotide structure of some species' genes	204
Figure 6.6.	Relationships of genome	205
Figure 6.7.	Frequent suffix trie structure of string $s = \text{'ATGATGT'}$	210
Figure 6.8.	Frequent suffix trie structure using proposed frequent max substring technique	212
Figure 6.9.	Comparison of number of indexing terms extracted from two approaches at given frequency threshold value = 2	215
Figure 6.10.	Comparison of number of indexing terms extracted from two approaches at given frequency threshold value = 3	215
Figure 6.11.	Comparison of number of indexing terms extracted from two approaches at given frequency threshold value = 4	216
Figure 6.12.	Comparison of number of indexing terms extracted from two approaches at given frequency threshold value = 5	216

Figure 6.13. Comparison of number of indexing terms extracted from two approaches at given frequency threshold value = 6	217
Figure 6.14. Comparison of number of indexing terms extracted from two approaches at given frequency threshold value = 7	217
Figure 6.15. Comparison of number of indexing terms extracted from two approaches at given frequency threshold value = 8	218
Figure 6.16. Comparison of number of indexing terms extracted from two approaches at given frequency threshold value = 9	218
Figure 6.17. Comparison of number of indexing terms extracted from two approaches at given frequency threshold value = 10	219
Figure 6.18. Reduction rate of number of indexing term enumerations using frequent max substring technique and Vilo's technique when compared with conventional suffix trie algorithm	220

List of Tables

Table 2.1.	Types of Thai characters	33
Table 2.2.	Thai stopwords list from morphology	48
Table 2.3.	Advantages and disadvantages of Thai text indexing techniques	68
Table 3.1.	Comparison of suffix trie, suffix tree and suffix array	92
Table 3.2.	All frequent substrings with number of occurrences	99
Table 3.3.	Basic statistics for Thai text collection	126
Table 3.4.	All test queries consisting of four phrase queries and four single word queries	141
Table 3.5.	Precision and recall values of word inverted index technique	142
Table 3.6.	Precision and recall values of 3-gram inverted index technique	142
Table 3.7.	Precision and recall values of Vilo’s technique	143
Table 3.8.	Precision and recall values of suffix array technique	143
Table 3.9.	Precision and recall values of frequent max substring technique	144
Table 3.10.	Average precision and recall values of five indexing techniques	144
Table 4.1.	Thai part-of-speech as tagset for ORCHID	159
Table 4.2.	Number of insignificant indexing terms and meaningful indexing terms extracted with hybrid method	163
Table 4.3.	Comparison of number of indexing terms extracted from hybrid method and word inverted index technique	165
Table 4.4.	Precision and recall values of word inverted index technique	169
Table 4.5.	Precision and recall values of hybrid method	169
Table 4.6.	Average precision and recall values of two indexing techniques	170
Table 5.1.	Clustering results of using SOM and frequent max substring technique	186
Table 5.2.	Clustering results of using hierarchical clustering approach	187
Table 6.1.	Number of indexing terms extracted from frequent max substring technique	198
Table 6.2.	Number of indexing terms extracted from bi-gram based indexing	200

Table 6.3.	Comparison of retrieval time used by bi-gram based indexing and frequent max substring technique	203
Table 6.4.	Number of frequent max substrings and frequent substrings	213