

37814

Universiti Teknologi Mara

**Enhancement of Rules-Application-Order (RAO)
Stemming Algorithm Based On the First
Character of Malay Word**

Edatul Muliana binti Ghazalli

**Thesis submitted in fulfillment of the requirements for
Bachelor of Science (Hons) Information Technology
Faculty of Information Technology And
Quantitative Science**

November 2005

DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries, which have been duly acknowledge.

NOVEMBER 21, 2005

EDATUL MULIANA BINTI GHAZALI

2003657552

ABSTRACT

Stemming is important thing to improve retrieval effectiveness. Stemming is used to reduce the size of indexing file for relevancy of document retrieval. Stemming is technique to truncate the word into the root word that will reduce vocabulary size and improve recall. The Malay affixes consist of four different types such as prefix, prefix-suffix, suffix and infix. An effective and powerful of Malay stemmer is it just not to move the suffixes rules only but it must remove all four types of affixes. Without removing all the affixes, the stem cannot be effectively used to index of Malay documents. So in order to get the best order of morphological rule for effective and powerful stemmer the researcher has to find out the best order of morphological rule to stem Malay words based on first character for each alphabet. This project involves the use of two combinations simultaneously. The words that could not stem correctly by the first combination of best order which is primary will shift to alternative combination of best order of morphological rule. The results of experiment B, which is enhance project is better than experiment A, which is Rules-Application-Order (RAO) by Fatimah (1995) because that algorithm has successfully stemmed all word begin with alphabet "A" until "Z" that extracted from Quran documents..

CONTENTS

		Page
DECLARATION		ii
ACKNOWLEDGEMENT		iii
ABSTRACT		iv
CONTENTS		v
LIST OF TABLES		viii
LIST OF FIGURE		ix
CHAPTER 1	INTRODUCTION	
1.0	Research Background	1
1.1	Problem Description	2
1.2	Project Aim	3
1.3	Project Objective	3
1.4	Project Scope	4
1.5	Project Significance	4
1.6	Summary	5
CHAPTER 2	LITERATURE REVIEW	
2.0	Introduction	6
2.1	Stemming Algorithm	7
2.2	Stemming Algorithm for Malay Words	7
	2.2.1 Asim Stemmer	8
	2.2.2 Rules-Application-Order(RAO) Stemming Algorithm	9
2.3	Malay Affixes	9
	2.3.1 Prefix	10
	2.3.2 Prefix-Suffix Pair	11
	2.3.3 Suffix	11

	2.3.4	Infix	12
2.4		Stemming Algorithm in Other Languages	13
	2.4.1	English Language Stemmers	13
		2.4.1.1 Dawson Stemmer	14
		2.4.1.2 Porter Stemmer	14
	2.4.2	Slovene Language Stemmers	15
	2.4.3	French Language Stemmers	16
	2.4.4	Arabic Language Stemmers	16
2.5		Summary	17
CHAPTER 3		RESEARCH METHODOLOGY	
3.0		Project Methodology Overview	18
3.1		Data Acquisition	19
	3.1.1	Database Collection	19
	3.1.2	Dictionary	21
	3.1.3	Stop Word Checking	22
3.2		System Design and Architecture	23
3.3		System Coding	26
3.4		User Interface Design	28
3.5		Testing and Debugging	29
3.6		Documentation	29
3.7		System Requirement	29
	3.7.1	Software Requirement	29
	3.7.2	Hardware Requirement	30
3.8		Summary	30
CHAPTER 4		RESULT AND ANALYSIS DATA	
4.0		Introduction	31
4.1		Comparison with Rules-Application-Order (RAO)	31