TO ENHANCE EXISTING MALAY STEMMING ALGORITHM STARTING WITH THE LETTER 'D'

MOHD NAZRIL HAFEZ B MOHD SUPANDI

THESIS SUBMITTED IN PARTIAL FULFILLMENT FOR THE DEGREE
BACHELOR OF SCIENCE

FACULTY OF INFORMATION TECHNOLOGY
AND QUANTITATIVE SCIENCES
UNIVERSITI TEKNOLOGI MARA
SHAH ALAM

2000

ACKNOWLEDGEMENT

Praise be to Allah (SWT). His loves and peace be upon Nabi Muhammad (SAW) and his family and companions. Thank to Allah (SWT) for giving me the time and strength to finish writing this thesis.

I would like to express my sincere gratefulness and gratitude to my supportive supervisor, PM. Dr. Zainab Abu Bakar for her invaluable guidance, encouragement, and advice during the course of this thesis.

My thanks also go to all my friends for their co-operation where I spent most my time, for sharing happy and unforgettable moments together.

Last but not least, I would like to express my gratitude to my beloved family for their encouragement, patience, support, financial support and sacrifice they have given me during the course of this thesis.

ABSTRACT

This thesis concerns a Malay language documents retrieval system. Stemming algorithm, database Quran translated documents and electronic root dictionaries are used in order to complete this study. The performance of a Malay stemming algorithm is tested based on words that beginning with 'd', using 4 experiments. First, use the original set of data collections. Second, adding a new words in the dictionary. Other than that we modify the total value for 'a', 'k' and 'm'dictionary in header file "dcvarnew.h". Third, the modification into the program is adding the affixes rule format in "rule.txt". Forth, add a new code to differentiate the use of affix rule of "*di+an*" and "*di+kan*". The main objective is to minimize the unstemming, understemming, overstemming, spelling exception and other problems that occurred when 'd' word stemmed. It is achieved the objective when the best order of rule to used to stem the words that beginning with 'd' is met. In which it involves the use of two combinations simultaneously such as the pair combination of 1234 as primary combination and 2341 as the secondary. First, all the words will used the 1234 combination, and if the program encountered that the words can not be solved correctly, the combination will be shifted to the secondary combination that is 2341. These experiments can serves as a benchmark for future research in Malay language. Furthermore, it can help those who are interested to know about certain subject matters from the Al-Quran where the document retrieval system will automatically retrieve all relevant documents in response to the users' queries.

CONTENTS

CHAPTER I

INTRODUCTION

1.1 Background

A stemming algorithm is a computational procedure, which has the ability to reduce all words with the same root to a common form, usually by discarding each word of its derivational and inflectional suffixes (Lovins 1968). Stemming is also carried out in various languages and will be discussed in Chapter II.

As for the Malay stemming algorithm, it is more effective and powerful than one being developed by Asim (1993). The new stemming approach, termed as the Rules-Application-Order (RAO) approach, is the first of its kind to be introduced to cater for the stemming process of Malay word.

There are very good reasons and interests on doing this project. First, we can study on stemming algorithm that is used by Fatimah (1995) in her information retrieval project. Second we can study to the existing algorithm and make modification to overcome the stemming errors such as overstemming,