TO IMPROVE STEMMING ALGORITHM ON MALAY WORDS BEGIN WITH ALPHABET B

By

Norasiah Binti Ismail

Final Project Paper Submitted in Partial Fulfilment for the Degree of Bachelor Of Science (Hons.) in Information Technology, Faculty of Information Technology and Quantitative Sciences
Universiti Teknologi MARA

April 2000

ACKNOWLEDGEMENT

Firstly, praise be to Allah (SWT) for giving me 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 guidance, encouragement and advice during the course of this thesis.

My thanks also to all the lecturers that are involved in this project. I would like to thanks my entire friends especially to my housemates for their support, advice and consideration in finishing this project.

Last but nor least, I would like to express my gratitude to my beloved family for their encouragement, patience, support and financial support and sacrifice they have given me during the course of this thesis.

TO IMPROVE STEMMING ALGORITHM ON MALAY WORDS BEGIN WITH ALPHABET B

By

NORASIAH BINTI ISMAIL

April 2001

This thesis concerns a Malay language document retrieval system. Stemming algorithm, Malay Quran translated documents and root dictionaries are used in order to complete this study. The performance on words beginning with letter 'b' of Malay stemming algorithm are tested using 5 experiments. First experiment is use the original set of data collections. In second experiment, affixes rule are added in rule format in file "rule.txt". Third experiments are modifying the total value for 's' dictionary in header file "dcvarnew.h". For fourth experiment, a new word is adding in the dictionary and modifies Malay Quran translated. In fifth experiment, the total value for 'a' dictionary in header file "dcvarnew.h" is modifying. The main objective of these experiments is to minimize the unstemming, understemming, overstemming, spelling exception and other problems that occurred when 'b' words are stemmed. The objective is achieved when the best order of the rules to use to stem the words that beginning with 'b' is met. This involves the use of two combinations simultaneously such as the pair combination of prefix-suffix-prefix suffix-infix as primary combinations and prefix suffix-suffix-prefix-infix as the secondary. First, all the words used the prefix-suffix-prefix suffix-infix combination, and if the program encountered that the words can not be solved correctly, combination will be shifted to the secondary combination that is prefix suffix-suffix-prefix-infix combination. These experiments can serves as a benchmark for future research in Malay language in finding the best approach to stem words that begin with other rest of alphabets.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Background

A stemming algorithm is a computational procedure, which reduces all words with the same root to a common form by stripping from each word its derivational and inflection suffixes according to Lovins (1968). Popovic & Willett (1992) says that it is one of the well-known conflation algorithms that are used to identify morphological variants.

As for Malay stemming algorithm by Fatimah (1995), it is more effective and powerful than one being developed by Asim (1993). The new stemming approach known as Rules-Application-Order (RAO) approach is the first of its kind to be introduced to cater the stemming process of Malay words by Fatimah (1995).

An information retrieval system may retrieve written texts, spoken utterances or graph and images. Text is basic medium for formal communications between human beings. It consists of notes, memoranda, magazines, newspaper and so on. A system that ultimately provides the user with full document texts is called "document retrieval system" according to Fatimah (1995). Document retrieval system constitutes one class of information retrieval (IR) system and it is considered as the main focus of interest in IR.

The IR can be defined broadly as the study of how to determine and retrieve from a corpus of stored information the portions, which are responsive to the particular