STUDY OF STEMMING ALGORITHM FOR MALAY WORDS WHICH BEGIN WITH ALPHABETS 'M'

MOHD ZAWAWI BIN MOHD YUNUS

THESIS SUBMITTED IN PARTIAL FULFILMENT FOR THE DEGREE OF BACHELOR OF SCIENCE

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

2000

ACKNOWLEDGEMENT

In the name of Allah, the Compassionate, the Merciful, Praise be to Allah, Lord of the Universe, and Peace and Prayers be upon His final Prophet and Messenger.

I would like to express my sincere gratefulness and gratitude to my supervisor, Professor Madya Dr. Zainab Binti Abu Bakar for her invaluable guidance, encouragement and advice during the course of this thesis. She is the one who encouraged me the freedom of individual endeavor essential to the fulfillment of my work and gave very possible assistance in the pursuit of this study.

A pleasure to acknowledge to the lecturers, supporting staffs especially the assistance laboratory of Faculty of Information Technology and Quantitative Science for their cooperation through this research. I would like also to express my gratitude to Shahezlin Shaharuddin, Mohd Nazrul Che Mahmud and Mohd Nazril Hafez Md Supandi for their cooperation, encouragement, contribution and spending time together to discuss about this research.

Last but not least, in collection of many kindness, continuous encouragement for perseverance and patience, my grateful thanks to my family especially my parents, Mohd Yunus Saaban and Not forgetting thanks to my classmates and housemates for their patience, support and encourage me to complete this research. Thanks to all.

ABSTRACT

This research concerns a study of stemming algorithm for Malay words begin with alphabet 'M'. This research involves a Malay stemming approach called Rules-Application-Order (RAO). The performance of this Malay stemming algorithm is tested using the test collection of 1066 words that starts with the letter 'M' that have been extracted from 6236 Malay Quran documents. It also used 24 different combinations of Malay affixes that consist of prefix, prefix-suffix, suffix and infix. The results are obtained from the experiments that use the four rules and it combination. The type of errors found in the stemming algorithm is overstemmed, understemmed, spelling exception and unstemmed. These stemming algorithm problems will be solved by doing five experiments such as analysis the existing algorithm, do correction in the file, adding rules, correct the stemming algorithm and use two combination rules. The results of the experiments will show that the algorithm has successfully stemmed all Malay words begin with alphabet 'M' that extracted from Quran documents.

CONTENTS

CHAPTER I

INTRODUCTION

1.1 Background

Stemming algorithm is a computational procedure, which has the ability to reduce all words with the same root 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.

The Malay-stemming algorithm by Fatimah (1995) is more effective and powerful than one being developed by Asim (1993). This stemming approach, termed as the Rules-Application-Order (RAO), is introduced to cater for the stemming process of Malay words. Fatimah (1995) also designs the dictionary of Malay root words in her project.

This study is performed to solve the errors of the stemming algorithm that occurs in Fatimah (1995). This study is a continuous study done previously by Fazlina (1999). In Fazlina (1999), she just concentrates to find the errors of the Fatimah's (1995) stemming algorithm only. This study also uses Fatimah's method (Fatimah 1995) to overcome the problems exists.