TO STUDY THE PERFORMANCE OF STEMMING ALGORITHM ON MALAY WORDS BEGINNING WITH THE LETTER "S"

ROHANA BINTI JANTAN

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

ACKNOWLEDGEMENTS

First and for most, I would like to thank Allah SWT for giving me the time and strength to finish writing this thesis.

I would like to express my deep gratitude to my supervisor, PM Dr. Zainab Abu Bakar because of her invaluable guidance, cooperation, attention and advice towards completing this project. I also like to give my special thanks to my examiner Puan Rohani Mohd Zaid.

My thanks also to the lecturers and to all FTMSK staff for their cooperation in finishing this project.

Last but not least, I would like to thank my friends for their support, criticism and opinion. Finally, I sincerely appreciate advice, support and motivation from my husband, Zamri Lajim, my family, lecturers and all parties involved.

ABSTRACT

This thesis concerns the study of Malay stemming algorithm for the word beginning with the letter "S". This algorithm is used in the Malay language document that is used is the Quran translated document. A Malay stemming algorithm known as Rules-Application–Order (RAO) is applied in the experiment. In the experiments dictionaries of Malay root words and combination of morphological rules also used. The performance of the Malay stemming algorithm is evaluated by applying to the "S" word by removing different combination of prefixes. The "S" words or the resulted stemmed words are checked for their existences in the dictionaries. If these words do exist, the following stemming processes stop. These words are then analyzed. In the analysis, the percentage of each combination is compared to find the best prefixes combination. The result shows that there is still problem of overstemming, understemming and unstemming of word. For a total of unique 411 "S" words there are 0.73% overstemming, 0.73% understemming and 2.68% unstemmed words. Therefore, the algorithm must be modified in order to increase the performance of the stemming algorithm for Malay words.

CONTENTS

CHAPTER I

INTRODUCTION

1.1 BACKGROUND

Information Technology has made possible all information to be captured, transmitted from one person to the other, stored into database, retrieved the data if needed, manipulated or displayed via a computer that is linked to the Internet. Information can be classified as text, graphic or sound. Information that is stored in the database can be used in text retrieval systems.

Information Retrieval (IR) is defined as a study to determine and retrieve from a corpus of stored information the portions that are responsive to particular information needs (Tengku Sembok 1989). IR has some similarities with other areas of information processing such as management information system and database management system.

The main function of IR is to enable end-users with tools to perform a search more effective and efficient. For that, a knowledge-based approach and computational techniques are used to encode the expertise possessed by a trained intermediately to give a good result in information retrieval.