INCORPORATING STEMMING ALGORITHM IN THE MALAY INFORMATION RETRIEVAL THAT EMPLOYS THESAURUS APPROACH

MOHD ROSMADI MOKHTAR

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

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

2001

ACKNOWLEDGEMENTS

بسم الله الرحمن الرحيم
الحمد لله رب العالين
الصلاة والسلام على أشرف الانبياء والمرسلين

Praise to Allah the Almighty for providing me a superior strength throughout this period of study and for all that has bestowed on me. It is with His power the study is completed.

I would like to express a very special gratitude and thankfulness to my supervisor Assoc. Prof. Dr. Zainab Abu Bakar who gave me her reliance and insightful guidance that I have profited very much. She also gave lots of support and her precious time to guide me trough all this time.

My special appreciation to the Dean of the Faculty of Information Technology and Quantitative Sciences, Assoc. Prof. Azizi Ngah Tasir and the Programme Head of Bachelor (Hons.) Information Technology, Puan Kalsom Nasir who had managed a very good bachelor program throughout these years.

To my friend Mohd Zaini Mat Abas and others in the faculty, who are simply too numerous to indicate, thanks for all your time, understanding and support. May Allah will give in return for all.

Last but not least, to my parents, Hj Mokhtar b. Din and
my siblings and lastly to my beautiful nieces and nephews, I would like to forward my appreciation to them for their continuous support during good or bad time, their patience and compassion. Thank you.

والله علم بالصواب
أمين

ABSTRACT

This project incorporates the ROA stemming algorithm with thesaurus approach by Rapizal. It is an opportunity to find out whether combining stemming with thesaurus will improve retrieval effectiveness and efficiency. Advance in information technology has made it possible for a wide range of text-based information to be search and retrieved online, locally or from remote hosts. A wide range of text-based information therefore can be searched and retrieved from online connection anywhere in the world. This type of popularity is due to advancement in technology that is rapidly growing from day to day. There are many Malay word variants that have the same meaning available from Malay words itself. In order to overcome these words variants problems, the development of computational technique that could transform both user's search and database words into a single canonical form is introduces. It is known as conflation methods. One of well-known conflation methods is stemming algorithms, where it is used to identify morphological variants. Stemming algorithms are language dependent. They have proven to be successful to reduce words with the same stem to a common form and are evidenced by the work many researchers. Unfortunately, conflation method is unable to conflate different words that possess the same meaning. These words can only be conflated by a thesaurus that can handle hierarchic, synonymic, and also morphological relationship. To create a thesaurus for a given subject an extensive manual and highly skilled, therefore to solve this problem, another language dependent conflation method, thesaurus is used. Its can build all types of relationship that exist between words. The information retrieval thesaurus typically contains a list of terms, where a term is either a single word or phrase. The relationships between them are also included to assist in coordinating indexing and retrieval. So from this project study it is found that the incorporations of stemming algorithm and thesaurus successfully increase the retrieved and relevant documents using Malay query words but on the other hand reduces its efficiency.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

1.1 Background

The study on information retrieval is on how to determine and retrieve from a mass of prepared information, the part that is relevant to particular information needs (Sembok 1989). The main function of information retrieval system is to provide the users to perform searching effectively and efficiently. In principle, information storage and retrieval is simple. Suppose there is a store of documents and a person formulates a question to the storekeeper to which the answer is a set of documents satisfying the information need expressed by his question. The storekeeper can obtain the set by reading all the documents in the sore, retaining the relevant documents and discarding all the others. Although this solution seems to be perfect, but this it is obviously impracticable.

Information retrieval is a study on how to determine and retrieve from a corpus of stored information; the part that is relevant to particular information needs (van Rijsbergen, 1979). The main function of information retrieval system is to provide the user with tools to perform searching effectively and efficiently.