MALAY TEXT DOCUMENT RETRIEVAL SYSTEM USING THESAURUS APPROACH BASED ON USER QUERY

RAPIZAL ABD. TALIB

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

First and foremost I would like to thank Almighty God, Allah S.W.T for His Mercy and Compassion that enabled me to complete this thesis, Alhamdulillah.

Next, I would like to thank my lecturer cum supervisor, Assoc. Prof. Dr. Zainab Abu Bakar, for giving me valuable guidance without which this thesis could not been completed successfully and within the time frame allocated. She also gave me valuable insight and comments that helped me enormously in the preparation of this thesis. Special thanks is also due to Dean of Information technology and Quantitative Science, who permitted me the usage of the computer facilities in the faculty.

Finally, thank you to my friends and those who helped and gave support directly or indirectly for their encouragement and tolerance in finishing my project.

ABSTRACT

Information Technology has enabled information to be accessed widely via a computer that is linked to the internet. Due to this type of prevalence and advancement in technology, there is an increase interest in searching Malay document to enable scholars and researchers to access the database online. Conflation methods have been successfully used to identify word variants from English and French databases. Many conflation methods such as stemming method have been applied to Malay document retrieval system. Another way of conflating related terms is with a thesaurus, which lists synonymous terms, and sometimes the relationship between them. This study has evaluated and identified the effectiveness of thesaurus in Malay document retrieval system using translated Quran documents, one set of queries and a list of relevance judgement.

TABLE OF CONTENTS

CHAPTER I

INTRODUCTION

1.1 BACKGROUND

The study of information retrieval is 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 systems is to provide the users to perform searching effectively and efficiently. An important facility in any text retrieval systems is term conflation, the ability to obtain word matches (Ekmekcioglu et al. 1996). Words like 'fikir' and 'berfikir' are conflated to a root word 'fikir'. Unfortunately, conflation method such as stemming method is unusable to conflate words such as 'fikir' and 'tilik'. These words can only be conflated by thesaurus that can handle synonymic and also morphological relationship. For the project purpose, the use of thesaurus in Malay document retrieval will be studied and implemented.

1.2 PPROBLEM DESCRIPTION

Despite of implementing stemming method to Quran test collection by Zainab (1999), the result is still unsatisfied. One of the factors is the weakness of stemming method