FINAL YEAR PROJECT REPORT

SIGNATURE FILES INDEXING TECHNIQUE ON MALAY TEXT

AHMAD AZUAN BIN MUHAMMAD KAMIL
2006838004
BACHELOR OF COMPUTER SCIENCE (HONs.)
COMPUTER SCIENCE DEPARTMENT
FACULTY OF COMPUTER AND MATHEMATICAL SCIENCES
SEMESTER DECEMBER 2008 - APRIL 2009

Acknowledgements

Alhamdulillah, praise be to Allah the Almighty for giving me the strength and will to finish this thesis. I would like to express my full gratitude to everyone who has helped me in doing this project.

First and foremost, a special thanks to my supervisor, Assoc. Prof. Nurazzah Abd. Rahman, for being such a great help for the past year. Here I would also like to apologize for any wrongdoings that came from my part.

Secondly, to my coordinator, Dr. Siti Salwa Salleh, for guiding us CS230 students in this subject, and enduring our attitude of the course of these two semesters.

Also, thank you to my friends for helping me in my time of need, and last but certainly not least, my parents, for taking care of me and raising me, and for giving me moral support.

Thank you very much.

Abstract

Information retrieval is the study of determining and retrieving document from a collection in order to satisfy the user's need, usually expressed in natural language. A search engine is an invaluable tool in the context of information retrieval, because it enables the handling of large numbers of data, and provides people with a tool to access various kinds of information at the tip of their fingers. In information retrieval, a text operation called indexing is applied to the documents needed to be retrieved, in order to aid with the retrieval process by making it easier to search through the documents available. Signature files is an indexing technique which transforms a document into a form represented by superimposed bit strings called signatures. A hadith is a record of the traditions or sayings of Prophet Muhammad S.A.W, generally revered by Muslims, and considered as a second source of religious guidance after the al-Quran. There are many hadith search engine available on the web, but most of it are in English or Arab, and only few are in Malay. This study proposes the use of signature files to develop a search engine prototype for Malay hadith text documents. The efficiency of the prototype is evaluated by testing it using a number of queries. The uses of this prototype will aid people in the society by providing them with understanding of various knowledge, which hopefully will lead them to a better life.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

1.1 Background

In recent years, the advancement in technology, along with the advent of mass storage devices have considerably changed information retrieval. Information retrieval (IR) is a subfield of computer science that deals with the automated storage and retrieval of information (not data) from a collection of written documents. The retrieved documents aim at satisfying a user information need usually expressed in natural language (Baeza-Yates and Ribeiro-Neto, 1999). IR can be further divided into further sub-categories, such as data retrieval, text retrieval, and image retrieval.

One of the most commonly used applications of IR is search engine. In computing, search engine is an invaluable tool to have, whether in a website, software, or even online games. As the number of documents grew, so does the need for a way to retrieve a particular document quickly. Without search engines, retrieving particular information from a large collection will be a tedious and tiring process.

Search engines operate algorithmically or a mixture of algorithmic and user input, which, in the context of IR, are called 'queries'. Various techniques and algorithms are available, some of which work better in certain situations and/or conditions than others.