Universiti Teknologi MARA

Keyword Indexing for Text Documents Using Signature Files

ABDUL HAKIM BIN A.GAFA
2005614250

Thesis submitted in fulfillment of the requirement for
Bachelor of Computer Science (Hons.)
Faculty of Information Technology and Quantitative Science

November 2008

ACKNOWLEDGEMENT

Alhamdulillah, in the name of Allah the Most Compassionate, the Most Merciful and Most Gracious, praise to Allah, the one and only, for giving me patience, strength and ability to complete this report.

Firstly, I would like to thank Mrs. Prasanna Ramakrisnan my supervisor for the guidance and patience in assisting me in completing my research; I would like to thank Assoc. Prof. Syed Ahmad Syeikh Aljunid my Final Year Project Coordinator for his constant help, guidance, ideas and reassurance.

I would like to express my greatest gratitude for my family who didn't stop motivating me all the times. My friends helped me a lot during the preparation of this research.

Upon complete of this research, I've learnt so many things regarding the knowledge of Information Technology and I would love to learn to improve myself in my future undertakings.

ABSTRACT

Keyword Indexing for Text Documents Using Signature File

BY

Abdul Hakim Bin A.Gafa

November 2008

Information retrieval is the first step in developing retrieval systems for text document in collections. Signature File is popular and effective in searching and retrieving processes (Zobel and Moffat, 2006) other than Inverted Files. This project explores the potential and limitation of prototype text search engines using Signature Files on Malaysian Text Documents. Malaysian Text Documents is an official text report of proceedings and debates in parliament which is documented in Malay Language and maintained by House of Parliament. These document are categorizes into House of Commons and House of Lords. Currently, searching and retrieving information from text document in Malay Language are done manually. These process are tedious, very time consuming and inefficient. Text search engine prototype using signature file can speed up the process of searching and retrieving information from Malaysian text documents. The main of this project is to compare the effectiveness of searching Text documents between using Signature files algorithm and Inverted files algorithm. In order to achieve the main objective, the Signature Files algorithm for indexing methods needs to be understood and implemented. A text search engine prototype for Malay Text Document will developed as a tools to evaluate the effectiveness of searching Text Documents using Signature Files and Inverted Files.

Table of Contents

CHAPTER 1

INTRODUCTION

CHAPTER 2

LITERATURE REVIEW

CHAPTER 1

INTRODUCTION

This chapter will discuss the background and rationale for the study. This chapter emphasis on the research background that significantly relates to the problem statement, research scope, research objectives and significant of text search engine for Text Documents. It also gives details of the significant use of information technology and information systems in Information retrieval.

1.1 Introduction

Information retrieval is fast becoming the dominant form of information access over taking traditional database style searching. According to Baeza-Yates, Ribeirto-Neto (Baeza-Yates and Ribiero-Neto, 1999), information retrieval deals with the representation storage, organization and access to information items. In requesting information from text search engines, user must first translate the information into query. This translation yields a set of keywords or index terms which summarizes the description of user information needs.