Universiti Teknologi MARA

Keyword Indexing Using Inverted File On Hansard Documents

Rosnawati Abdul Kudus

Thesis submitted in fulfillment of the requirements for
Bachelor of Science (Hons) Computer Science
Faculty of Information Technology And
Quantitative Sciences

May 2008

ACKNOWLEDGEMENT

I would like to thank my supervisor Mrs. Nur Atiqah Sia Abdullah for her constant help, guidance, ideas and reassurance. Thanks to my parents for their constant prayers and words of encouragement throughout the completion of this thesis.

ABSTRACT

Keyword Indexing Using Inverted File on Hansard Document

BY

Rosnawati Binti Abdul Kudus

May 2008

Information retrieval is the first step in developing retrieval systems for text document in collections. Inverted file is the most popular and effective in searching and retrieving processes (Zobel and Moffat, 2006). This project explores the potential and limitation of prototype text search engines using inverted files on Malaysia Hansard Documents. Malaysia Hansard Document is an official verbatim 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 hansard document are done manually. These process are tedious, very time consuming and inefficient. Text search engine prototype using inverted file can speed up the process of searching and retrieving information from hansard document. The objectives of this study are to develop a text search engine prototype for Malaysia Hansard Documents and to evaluate the prototype for seven speakers' speech text. Scopes of the research are to search and retrieve document up to two words and in Malay language. The methodologies in this study includes preliminary study about the models of text search engines and identify similar studies, analyze indexing techniques, defines data structure for inverted file which includes hash table, linked lists, vector, array and quick sort, collect and preprocessing hansard document, design and develop prototype using java platform, conduct testing to evaluate the accuracy of the prototype tool and analyze findings. From the experiment that has been conducted, the accuracy of search keywords through the prototype and manual check is 100 percents. This finding

TABLE OF CONTENTS

		PAGE
DECLARATION		**i**
ACKNOWLEDGEMENT		**ii**
ABSTRACT		**iii**
TABLE OF CONTENTS		**v**
LIST OF TABLES		**viii**
LIST OF FIGURES		**ix**
CHAPTER 1	**INTRODUCTION**	
1.1	Introduction	1
1.2	Problem Statements	2
1.3	Objectives	2
1.4	Research Questions	2
1.5	Significance of the Study	3
1.6	Limitation of the study	3
1.7	Chapter Organization	3
CHAPTER 2	**LITERATURE REVIEW**	
2.1	Overview	4
2.2	Definition of Information Retrieval	4
2.3	Type of Information Retrieval	5
2.4	Information Retrieval Framework	6
2.5	Related Areas of Information Retrieval	9
2.6	Current Research in Information Retrieval	11
	2.6.1 A Time Machine for Text Search	11
	2.6.2 Indexing Data Spaces	11

CHAPTER 1

INTRODUCTION

This chapter provides the background of the study. It also describes the objectives, the significant of text search engine for Malaysia Hansard Documents, the problems and the limitation that lead to this study.

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.

Text search engine is a key technology in information retrieval. Text search engines are tools for finding documents in a collection such as newspaper articles, academic publications, company reports, research grants applications, manual pages, encyclopedias, parliamentary proceedings, bibliographies, historical records, electronic mails and court transcripts. Besides, text search engines are implemented in most of the web search and searching files in the operating system.

Three main methods for indexing implemented in the text search engines are inverted files, suffix arrays and signature files. From these three methods, the inverted file is the most popular and effective in searching and retrieving process (Zobel and Moffat, 2006).