UNIVERSITI TEKNOLOGI MARA

RETRIEVING MALAY HANSARD DOCUMENTS USING TOPIC DISCOVERY

NURUL AIN BINTI MOHD FADZIL THANI

Thesis submitted in fulfillment of the requirements for
Bachelor of Computer Science (Hons)
Faculty of Computer and Mathematical Sciences

January 2012

ACKNOWLEDGEMENTS

In the name of Allah, Most Gracious, Most Merciful. Praise to Allah, the One and only, for giving me the strength and ability to complete this project. Firstly, I would like to send my gratitude and appreciation to my supervisor, Prof. Dr. Zainab Abu Bakar, and also my lecturer, Puan Haslizatul Fairuz Mohamed Hanum, as both lecturers are there to give me valuable advice, unwavering encouragement and patience throughout this project. I would like to thank them for giving me the opportunity to work under their guidance, which has been the most memorable experience. I also deeply like to take this opportunity to send gratitude to my coordinator of CSP650 Project course, En. Abdul Rahman Mohamad Gobil, for tutoring and giving me valuable advice regarding to this course. Not to be forgotten, my former coordinator, Dr. Sharifalillah Nordin, and former supervisor, En. Mohd Fauzi Saman, who give me a lot of knowledge and experience of this course.

My heartfelt gratitude and love goes to my mother Puan
my father Encik Mohd Fadzil Thani Abu Bakar, my siblings whom have helped me so much throughout the years, through tears and laughter. Without all of you, I probably could not make it this far.

Finally, I want to give a great accomplishment to all my friends who accompany me throughout my learning years and encourage me in completing this project, especially my CS2306c group members. You guys are the greatest.

The project was made possible by the effort of many people who provided invaluable information, references material (past researchers) and collaborative support. Thank you very much, may ALLAH bless all of you.

TABLE OF CONTENTS

ABSTRACT

Due to difficulty bring by the overloaded of digitized collection, Information Retrieval rapidly concerns in improving task such as discovering relevant documents. The thesis is performed to improve the issues produced by the lack of keyword-based search for document in indexing and queries, and the shortage of sources on topic discovery for Malay language research. Thus, this thesis uses a topic discovery algorithm, which is Latent Dirichlet Allocation, in indexing to construct a conceptual-based search and selects Malay Hansard document as a data-set that represent Malay language document. The objectives of this thesis are to identify highest frequency words on Malay Hansard document using Word Frequency method, to index the data-set based on word suggested by Latent Dirichlet Allocation method, and to develop a retrieval prototype for this document using conceptual-based search. In this research, the result of highest frequency word from Word Frequency method is indexed as the keyword and acts as a baseline that represents the keyword-based search. While, the result of word suggested by Latent Dirichlet Allocation is indexed as a group of related keywords and it represents the conceptual-based search. As the result, from the indexing of conceptual-based, the retrieval prototype system is able to identify keyword that also related to search query word.

CHAPTER 1 : RESEARCH OVERVIEW

1.0 Introduction

This chapter focuses on the introduction of the thesis entitled "Retrieving Malay Hansard Document using Topic Discovery". This introduction section includes the background, problem statements, objectives, scopes, significance and outline of the thesis.

1.1 Background of Study

Nowadays, much collective knowledge kept on be digitized and stored in electronic forms, such as electronic publications, email, CD-ROMs, and the World Wide Web. This collection usually produced large collections of documents from assorted sources, such as web pages, news articles, scientific articles, digital libraries and e-mail messages. By reason of this overloaded of digitized collection, it brings difficulty to people in finding and discovering their relevant documents. This problem is not new for the study of *information retrieval.*

Information retrieval, which is one of the influential fields in Information Technology, is concerned with locating relevant document based on query provided by user (Mohd Pouzi and Tengku Mohd, 2005). Since 1940s, the problem of information storage and retrieval has increasingly catch attention. Since then, with the advent of computer, numerous thoughts have been given to support the usage of it in order to provide quick and intelligent retrieval systems. Yet, as the accuracy and speed in accessing information becomes more demanding, the problem of effective retrieval remains unsolved (Baeza-Yates and Ribeiro-Neto, 1999).

Lots of search engine represent documents indexing and queries by keywords alone, and base the comparison on the number of similar words in both of them. The greater similar words of the query and document have found in them, the higher the document is ranked. According to Salton and McGill (1983) in the book entitled "Introduction of Modern Information Retrieval", the performance of the earlier stated