

**WORD STEMMING FOR MALAY DOCUMENT BASED
RETRIEVAL SYSTEM USING LATENT SEMANTIC
INDEXING TECHNIQUE**

BY

**MUHD RUZLAN BIN KAMARUZAMAN
BACHELOR OF SCIENCE (Hons) COMPUTER SCIENCE**

**THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENT FOR THE DEGREE OF
BACHELOR OF SCIENCE**

**FACULTY OF SCIENCE COMPUTER AND
MATHEMATICS**

UNIVERSITI TEKNOLOGI MARA

NOV 2010

Acknowledgements

Bismillahirrahmanirrahim

In the name of Allah. The most Gracious and most Merciful

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 Haslizatul Fairuz Binti Mohamed Hanum for his outstanding guidance and valuable advice, who took a keen interest for my work all over the year, both in the implementation part and the writing of this thesis. Here I would also like to apologize for any wrongdoing that came from my part.

Secondly, to my coordinator, En.Fakhrul Hazman Bin Yusoff for guiding us CS2320 students in this subject, and our attitude of the course of these two semesters.

I would like also to thank all the CS230 students for the collaboration that we had all over the year. Finally, my warmest thanks to my parents for their support and enthusiasm during my studies in UiTM.

Thank you very much

Abstract

Documents retrieval in Information Retrieval Systems (IRS) is generally about understanding of information in the documents concern. The more the system able to understand the contents of documents the more effective will be the retrieval outcomes. But understanding of the contents is a very complex task. Conventional IRS applies algorithms that can only approximate the meaning of document contents through keywords approach using Latent Semantic indexing. In information retrieval, a text operation is called indexing is applied to the documents need to be retrieved, in order to aid with the retrieval process by making it easier to search through the documents available. This study purpose the use of Latent Semantic Indexing to develop a search engine prototype for Malay parliament documents. The efficiency of the prototype is evaluated by testing it using a number of queries.

TABLE OF CONTENTS

DECLARATION	ii
ACKNOWLEDGEMENT	iii
ABSTRACT	iv
TABLE OF CONTENTS	v
LIST OF FIGURES	vi
LIST OF ABBREVIATIONS	vii

CHAPTER 1 : INTRODUCTION2
1.1. Introduction2
1.2. Research Background4
1.3. Problem Statement4
1.4. Aim6
1.5. Objectives of the Research6
1.6. Research Scope6
1.7. Significance of the Research7
CHAPTER 2 : LITERATURE REVIEW10
2.1. Introduction10
2.2. Definition of Information Retrieval10
2.2. Information Retrieval System11
2.3. Information Retrieval Framework11
2.3.1 The Retrieval Procedure14
2.4. Indexing15
2.5. Document Retrieval16
2.6. IR model Involved17
2.6.1 Overview of Latent Semantic Indexing17

2.6.2 Use of Latent Semantic Analysis19
2.6.3 Disadvantages of Latent	
Semantic Indexing (LSI)21
2.6.3.1 Storage21
2.6.3.2 Efficiency of LSI22
2.6.4 Primary step in Latent Semantic	
Indexing (LSI)22
2.6.5 A Summary of Latent	
Semantic Indexing (LSI)23
2.6.7 Overview of Singular Value	
Decomposition (SVD)25
2.7 Related Researches27
2.7.1 Thematic Annotation extracting	
Concept out of document27
2.7.2. Research on Multi Document Summarization	
Based on Latent Semantic Indexing27
2.7.3 Word Stemming Algorithms and Retrieval	
Effectiveness in Malay and Arabic	
Documents Retrieval Systems28
2.7.4 Modelling Chinese Documents with	
Topical Word- Character Models29
2.7.5 Measuring the retrieval effectiveness of	
an IR System30
CHAPTER 3 : METHODOLOGY32
3.1. Introduction32
3.2. Malaysian Parliament Document34