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

Bismillahirahmanirrahim

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