IDENTIFYING TOPIC FOR INDEXING MALAY TEXT DOCUMENTS

BY

MUHAMMAD ALHAFIZ B HAMZAH

BACHELOR OF COMPUTER SCIENCE (Hons)

THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENT FOR THE DEGREE OF BACHELOR OF COMPUTER SCIENCE(Hons)

FACULTY OF COMPUTER AND MATHEMATICAL SCIENCES

UNIVERSITI TEKNOLOGI MARA

MAY 2011

Acknowledgement

Assalamualaikum w.b.t

"In the name of Allah, the most Gracious and most Merciful"

First and foremost, I would like to extend my deepest appreciation and gratitude to my dedicated supervisor, Puan Haslizatul Fairuz, for her guidance, encouragement, advices, ideas, and support throught the duration of this project.

Special thanks to my family for their concern, encourage and support throughtout my life. They are one of the most valuable gift in my life. I would like to thanks to Puan Latiffah Adam and others lecturers for their advices and guidance`s.

Last but not least, I would like to express my sincere gratitude to those who had been involved in contributing their time, effort and support in making this research successful. I am most fortunate to have the advice and guidance of many talented people, whose knowledge have enhanced this project in so many ways.

Thank you,

Wassalam.

Abstract

The number of document increase for time to time. Malay document is one of example documents that increase from day to day. The Malay document includes the newspaper, articles, journals and so on. However not all document have their own special topic. It is hard for user to determine the exact topic that they want. Based on the word frequency, the topic can be determined by looking at the words that most frequently used in the document. Thus, this prototype will be implementing using Parliament document Hansard. It will calculate the 50 words that frequently occurred on every document. It will calculate the probability of the word occurred before and after the word frequently use. Furthermore using the frequently used word, it can generate the indexing file. As the conclusion, the prototype that will be developed is capable to give the appropriate output to user.

Keywords : Word frequency, word probability, topic indexing

Table of Contents

CHAPTER 1

INTRODUCTION

1.0 Background

Nowadays, the total number of document increase from day to day. It can be classified by different type of document that can be produce such as journal, report, blog, news and so on. There are also documents the needs standard format such as meeting documents, report, and meeting minutes.

The Malay document currently frequently use in many document. The Malay document also includes report, news, meeting document, articles and many more. Malay document have its own characteristic and it is unique.

Normally many articles or news on the internet are using pdf format. The PDF or Portable Document Format is used for representing two-dimensional document. It is an independent software application. The pdf document includes text, fonts, images and 2D vector graphics.

As the user of Malay document such as journal, articles, newspapers and so on, there is a problem that it is hardly to find a topic of the articles during searching. Normally it will use not related topic for particular document. Mostly, the searching cannot give an accurate result because of there are many topics in one documents. The solution will achieved through counting word frequency and topic based indexing technique.