APPROVAL

Content Words Extraction for Malay Text Document

By

NURFARAHIDAYU BINTI SAMSHUDIN

Thesis is submitted in fulfilment of the requirement for

Bachelor of Science (Hons) Intelligent System Engineering

Faculty of Science Computer and Mathematics

Approved by:

Pn Haslizatul Fairuz Binti Mohamed Hanum

Thesis Supervisor

Date: 28 April 2011

ACKNOWLEDGEMENT

I would like to thank to Allah Almighty for giving me his blessing in completing my CSP 650 prototype development and report successfully in the time that has been given. Without His consent, I would never have completed anything.

I would like to take this opportunity to convey my appreciation to my supervisor Pn Haslizatul Fairuz Binti Mohamed Hanum for her guidance, encouragement, comment, ideas, and support in order to make me able to produce this work better. Last but not lease, I would like to thank thesis coordinator, and without his guidance I would never accomplish my final year project.

I also would like to thank to my family for their support, financial and encouragement to make me able to complete the prototype development and report completely. Lastly thanks to all my friends especially Mohammad Arsyad bin Mohd Yakop for their information sharing and co operations. May Allah S.W.T bless them.

Thank You.

ABSTRACT

Information is growing rapidly; anyone is able to get the information easily without any restriction especially using the World Wide Web. However, cause of too many information, sometimes readers cannot get the important value of that information. Therefore, it will leads to wrong information and waste of time on reading. The research proposes an algorithm that will automatically extract the Malay documents to improve access to information. Content words extraction techniques is explored and used as possible content and value for the text document. In the process of development the prototype, Bigram technique is introduce to assists on searching the related word of content word. As a result, the prototype will display all related sentences with content words.

TABLE CONTENTS

CONTENT PAGE

CHAPTER ONE: RESEARCH OVERVIEW

CHAPTER TWO: LITERATURE REVIEW

CHAPTER 1

INTRODUCTION

1.0 Introduction

The emergence of the Internet and the availability of very large amounts of documents online that contain valuable information, have caused the need for tools to assist the users to extract the relevant information from the bundle of information without having to read them all, and also to retrieve it in a fast and effective way. However, the problem with textual information is that it is not designed to be handled by computers. Unlike the tabular information typically stored in databases today, documents have limited internal structure. Furthermore, the important information they contain is not explicit but is implicit, hidden in the text. The general objective by using a Natural Language Processing is to minimize the time a user spends in the steps leading to understanding the content of a document or a collection of documents. Natural Language Processing (NLP) is one of the Artificial Intelligence techniques that always involves with text utilization. The application that utilize with NLP is Information Retrieval (Decker, n.d).

Information Retrieval (IR) and document browsing is the process of matching user's query against collections of unstructured documents at the aim of finding the documents that match user's information needs or interests. The common document representation is the term-document vectors in which term matching is applied. In order to match documents to the user's query learning-based IR system use machine learning technique to identify and extract patterns in the documents (Kozima, 1993).