

University Teknologi MARA

**USING TEXT MINING  
FOR  
INFORMATION EXTRACTION**

**Saliza binti Ramly  
2005616868**

Thesis submitted in fulfillment of the requirements for:

**Bachelor of Science (Hons) Information System Engineering  
Faculty of Information System and  
Quantitative Sciences**

31<sup>st</sup> May 2007

## **DECLARATION**

I certify that this thesis and the research to which it refers are the product of my own work and that any ideas or quotation from the work of other people, published or otherwise are fully acknowledged in accordance with the standard referring practices of the discipline

.....

**Saliza binti Ramly**

2005616868

**Date: 31<sup>st</sup> May 2007**

## **ABSTRACT**

The growth 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. An e-mail is composed of date, e-mail address, subject, body of the e-mail, and so on. It is possible for the body to include pictures, sounds, and programs, but usually the body is mainly composed of textual data. Thus, it is possible to use text mining techniques in order to analyze e-mails. The research focuses on the email of students in Faculty of Information Technology and Quantitative Sciences (FTMSK). There are three objectives of the research that have been achieved. The survey was conducted to achieve the first objective. The second objective was achieved through content analysis and website observation. Researcher was identified the basic techniques that usually used and tabulate it in form of table. A number of organizations that have been done some development on text miner as their commercial product also have been identified. Finally, the third objective of the research was achieved through the development of a tool using text mining techniques. Furthermore, the Prototyping Methodology is chosen in order to develop the system. The researcher identified appropriate techniques from the past researches and existing text mining tool. As a result, categorization, clustering and summarization techniques was selected and applied for Text Mining Application Tool, TMAP development.

# TABLE OF CONTENTS

<b>TITLE</b>	
<b>APPROVAL</b>	
<b>DECLARATION</b>	
<b>ACKNOWLEDGEMENT</b>	iii
<b>ABSTRACT</b>	iv
<b>TABLE OF CONTENTS</b>	v
<b>LIST OF FIGURES</b>	ix
<b>LIST OF TABLES</b>	x
<b>LIST OF APPENDICES</b>	xi
<b>LIST OF ABBREVIATIONS</b>	xii
<b>CHAPTER 1: INTRODUCTION</b>	
1.0 Introduction	1
1.1 Overview Of Research	1
1.2 Problem Statement	3
1.3 Objectives Of The Project	3
1.4 Scope Of The Project	4
1.5 Significance Of The Project	4
1.6 Limitation of Project	5
1.7 Report Overview	5
1.8 Summary	6
<b>CHAPTER 2: LITERATURE REVIEWS</b>	
2.0 Introduction	7

2.1	Information	8
2.2	Text	9
2.3	Information Retrieval	11
2.4	Text Mining	14
2.5	Web Text Mining	16
2.6	Information Extraction	18
2.7	Text Mining Techniques	19
2.7.1	Summarization	20
2.7.2	Categorization	21
2.7.3	Concept Linkage	23
2.7.4	Clustering	24
2.7.5	Information Visualization	26
2.8	Benefits Of Text Mining	28
2.9	Text Mining Applications	30
2.10	Existing Commercial Software	33
2.11	Email	34
2.12	Summary	35

## **CHAPTER 3: METHODOLOGY AND APPROACHES**

3.0	Introduction	36
3.1	Methodology	38
3.1.1	Conceptual Study	38
3.1.2	Exploratory Study	38
3.1.3	Presentation Of Application	39
3.1.3.1	Produce Documentation	39
3.1.3.2	Design And Build	40
3.1.3.3	Prototype Methodology	40
3.1.3.3.1	Develop Abstract Specification	41