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

31st 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: 31st 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, TMAT development.

TABLE OF CONTENTS