Universiti Teknologi MARA

Evaluation Of The Effectiveness Of Clustering Algorithm In Retrieving Malay Documents

Aminah bt Mahmood

Thesis submitted in fulfillment of the requirements for
Bachelor of Science (Hons) Information Technology
Faculty of Information Technology And
Quantitative Science

October 2004

ACKNOWLEDGMENT

Alhamdulillah, all praises to the almighty Allah S.W.T for giving me the good health and strength, and with His blessing, finally my project thesis had successfully developed with my best efforts. These efforts include the development, research and testing of the theories and programs to determine their effectiveness. This project would have taken much longer to create, if I did not have the help and criticism of the following people. So here, I wish to acknowledge the excellent support and valued contributions from a number of persons that I have received in the completion of this thesis.

Firstly, I would like to express my sincere and deepest gratitude and thankfulness to my supervisor lecturer, Encik Normaly Kamal Ismail for his suggestions, comments, assistance and generous guidance for improvement during the preparation of this project. I am very grateful to him who reviewed the report, made valuable suggestions and brought an embarrassingly large number of errors and omissions to my attention.

Special thanks to my friends, for their valuable contributions especially to my group 6B classmates whose cooperation, friendship and understanding were crucial to the production of this project. Apart from that, I would also like to address my special appreciation to many lecturers in FTMSK for their cooperation, help and strong support. No words can be portrayed other than thank you and may Allah bless you all.

Finally, as always, my gratitude goes to my family especially thanks to my dad, a special person that always stand two-step behind me, for his never ending encouragement, patience and supplements along with his moral support for my effort. I love you dad!

ABSTRACT

In recent years, we have witnessed a tremendous growth in the volume of text documents available on the Internet, digital libraries, new sources and company-wide intranets. This has led to an increased interest in developing methods that can help users to effectively navigate, summarize and organize this information with the ultimate goal of helping them to find what they are looking for. The main issue in this information age is the efficiency and effectiveness of the retrieval system that can be used by the information provider. A good retrieval system should provide tools to perform searching accurately based on user requirements. Cluster analysis is a technique for multivariate analysis that assigns items to automatically created group based on a calculation of the degrees of association between items and groups. In the information retrieval (IR) field, cluster analysis has been used to create groups of documents with the goal of improving the efficiency and effectiveness of retrieval, or to determine the structure of the literature of a field. The IR community has explored document clustering as an alternative method of organizing retrieval results, but clustering has yet to be deployed on the major search engines. This study has evaluated and identified the effectiveness of clustering algorithm in Malay document retrieval system using Hadith test collections, which consists of Hadith documents, relevant judgments and one set of queries. Three types of experiments are conducted. First experiment use exact match, which is no method, is applied. Second experiment use stemming method. Finally, the last experiment uses combination of stemming and clustering methods.

TABLE OF CONTENTS

CHAPTER 1

INTRODUCTION

This chapter sets forth the rationale, significance and objectives of the study. It includes details of the background, problem statement, research questions and the hypotheses to be tested.

1.1 Background

Information Retrieval is one of the important fields in Information Technology. The study on Information Retrieval is how to determine and retrieve from a mass of prepared information, the part that is relevant to particular information needs (Sembok 1989). Clustering algorithm is one of the methods that used in Information Retrieval. There are also other methods such as stemming, thesaurus construction, N-gram, Malay Spelling Exchange Rule (MASER) or Dynamic Programming. The main function of information retrieval systems is to provide the users with tools to perform searching effectively and efficiently.

This project is mainly about retrieving Malay documents. It will evaluate the effectiveness and efficiency of clustering algorithms in retrieving Malay documents. The documents will be retrieved based on Malay query words entered by the users. To evaluate the performance of the clustering algorithms, this project requires Malay test collections which are Hadith translated documents, a set of natural language query words, a list of stop words, a list of rules and relevant judgment list. The evaluation can employ possible combinations with and without clustering algorithms. Retrieved Malay texts are ranked and compared to the list of relevant judgment.