Cross Language Information Retrieval (Malay- Arabic) For Hadith Document Using Stemming and Exact Matching Technique

BY

FARHANA BINTI HASAN

BACHELOR OF COMPUTER SCIENCE (HONS)

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

FACULTY OF COMPUTER AND MATHEMATICAL SCIENCES

UNIVERSITI TEKNOLOGI MARA

NOV 2010

Acknowledgement

Assalamualaikum w.b.t

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

May His blessing be upon the Prophet Muhammad s.a.w

I would like to express my profound gratitude to Allah S.W.T, for He has bestowed me with ideas, strength and opportunity to finish up this project.

My deepest appreciation and gratitude also goes especially to my dedicated supervisor, Assoc. Prof. Nurazzah bt Abdul Rahman, for her invaluable support, encouragement, supervision and useful suggestions throughout this research work. Her moral support and continuous guidance enabled me to complete my work. I am also thankful to Dr. Fahkrul Hazman Mohd Yusoff, my thesis coordinator, for all of his comments and valuable suggestions that he gave to me. I would also like to thanks Mr Normaly Ismail and other lecturers for their advices and guidance's.

My personal gratitude goes to my family, thanks for their love and supporting throughout my life. I also would like to thank to all of my friends, who shared their love and experiences with me. Finally thanked to all who directly or indirectly involved during completing this final project.

Thank you.
Wassalam

Abstract

Classical Information Retrieval (IR) is the sifting out of the documents most relevant to a user's information requirement expressed as a "query", from a large electronic store of documents. A search engine performs IR by retrieving relevant web pages from the internet. Cross Language Information Retrieval (CLIR) allows the user to state their query in one language, and retrieve documents in another. Some CLIR systems use language resources such as bilingual dictionaries to translate the user's original query. Generally, Hadith directory provide facility to search Hadith, but the main problem is translation between Malay to Arabic Hadith document is rarely found and it use Arabic as lingual franca. Thus mean, only people who have master on Arabic or at least have basic Arabic can use that system. As effect from this situation, it will create language barrier for the non-Arabic because only a few people especially Malay people can use this facility. Therefore, Cross Language Information Retrieval (CLIR) is use to overcome this problem. The objectives of this project are to develop a Cross Language Information Retrieval CLIR (Malay-Arabic) search engine for Hadith (Sahih Bukhari & Sahih Muslim) text documents using stemming and exact match and to create a digitized dictionary (Malay-Arabic) with a limited scope. In investigate the retrieval effectiveness by using Recall and Precision formula, there are five experiments are conducted based on the queries on that language (Roslan, 2008).

Keyword: Cross Language Information Retrieval (CLIR), Hadith, Retrieval Effectiveness, Translation.

TABLE OF CONTENTS

CHAPTER 1 - INTRODUCTION

This chapter describes the background of the problem which refers to the rational parts of the research called the problem statement. It starts from the basic introduction of the project, problem statement, objectives, scope and significance of the project.

1.1 Introduction

In principle, this study is about translation between Arabic and Malay language. These two languages belong to different settings and different language families. Arabic is classified as a member of the Semitic family of languages. Malay language is as a member of the Malaysians language family (Youssef Kadri & Jian, 2004).

Malay language is a language widely used in Malaysia. Mean while Arabic is an old language and internationally used especially in the Middle East countries. Many Malays, which are mostly Moslem, are learning Arabic to perform and practice their religion. This is because Arabic is the language for the Moslems' holy book, the Quran.

Arabic alphabets are much more complex than Latin alphabets. It is written right-to-left, and the characters are written continuously in a word (Khirulnizam Abd Rahman et al., 2008). This mean Arabic word is difficult to translate. Since the Malay word is using Latin characters, there is nothing much different in handling the characters to translate.