SEARCH ENGINE SYSTEM FOR THE HOLY QURAN

A thesis submitted to the Graduate School in partial fulfillment
of the requirements for the degree of Masters of Science
(Intelligent System)
University Utara Malaysia

By

NADIR SALEH NABOUS

Matric No: 88572

KOLEJ SASTERA DAN SAINS
(College of Arts and Sciences)
Universiti Utara Malaysia

PERAKUAN KERJA KERTAS PROJEK
(Certificate of Project Paper)

Saya, yang bertandatangan, memperakukan bahawa
(I, the undersigned, certify that)

NADIR SALEH NABOUS

calon untuk Ijazah
(candidate for the degree of) **MSc. (Intelligent System)**

telah mengemukakan kertas projek yang bertajuk
(has presented his/her project paper of the following title)

SEARCH ENGINE SYSTEM FOR THE HOLY QURAN

seperti yang tercatat di muka surat tajuk dan kulit kertas projek
(as it appears on the title page and front cover of project paper)

bahawa kertas projek tersebut boleh diterima dari segi bentuk serta kandungan dan meliputi bidang ilmu dengan memuaskan.
(that the project paper acceptable in form and content, and that a satisfactory knowledge of the field is covered by the project paper).

Nama Penyelia Utama
(Name of Main Supervisor): **ASSOC. PROF. DR. NORITA MD NORWAWI**

Tandatangan
(Signature) : _____

Tarikh
(Date) : 25/6/08

PERMISSION TO USE

In presenting this thesis of the requirements for a Master of Science in Intelligent System (MSc. IS) from University Utara Malaysia, I agree that the University library may make it freely available for inspection. I further agree that permission for copying of this thesis in any manner, in whole or in part, for scholarly purposes may be granted by my supervisor or in their absence, by the Dean of Graduate School. It is understood that any copying or publication or use of this thesis or parts thereof for financial gain shall not be allowed without my written permission. It is also understood that due recognition shall be given to me and to University Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Request for permission to copy or make other use of materials in this thesis, in whole or in part, should be addressed to:

Dean of Graduate School

University Utara Malaysia

06010 Sintok

Kedah Darul Aman

ABSTRACT

People seeking to explore the Holy Quran as a book of Guidance sometimes find it difficult to locate exactly the portion that relates to their quest. Thus there is a need for a search engine to assist users in locating the verses of their interest based on search keyword input. There exists already search engine software for Al-Quran. However it is common for both Arab natives and learners of Arabic to commit Arabic spelling mistakes especially with the use of vernacular languages or Arabic dialect words. This study proposed a query correction mechanism that will convert a vernacular word used in Arabic dialects into word in Al Rasm Al Othmani, the Arabic text as in the Al-Quran. Nadir's algorithm proposed in this study is an enhancement of the ISRI stemming algorithm and implemented specifically for searching Al Rasm Al Othmani Arabic text in the Holy Quran. The number of search improves even when the keyword has slight spelling mistakes or with the use words from the Arabic dialects. The prototype has also been validated by three experts with Islamic religious study background.

ACKNOWLEDGEMENTS

In the Name of Allah, the Most Gracious and the Most Merciful

It is my pleasure to acknowledge the immense contribution of some people who have assisted me one way or the other towards the successful completion of this project.

First of all, I give thanks to the Allah for his guidance and mercy throughout my life and all that has being achieved due to his will. Peace and Blessing to his last Prophet Muhammad (S.A.W.), his household and his companions. My sincere gratitude to my late father who has shown me the path between right and wrong (May his soul rest in peace), also worth mentioning is my mother (I seek Allah's protection for her life) for her continuous support and my family as a whole.

Secondly, my grateful thanks go to my supervisor, Assoc. Prof. Dr. Norita Md. Norwawi who had given her full support and contributed immensely towards the completion of this project. She has actually spent a lot of her time to give me the necessary advice, providing valuable information and editing errors to ensure that the best effort has been given in the completion and achievement of this project.

Lastly, I recognize the efforts of all my friends, Staff of Faculty of Information Technology, University Utara Malaysia and those who contributed directly or indirectly towards the completion of this project. Thanks to all.

Nadir Saleh Nabous
College of Arts and Sciences
Faculty of Information Technology
University Utara Malaysia
June 2008

I ask for Allah's guidance for what is acceptable to Him

TABLE OF CONTENTS

CHAPTER 3: METHODOLOGY

CHAPTER 4: FINDINGS AND RESULTS

CHAPTER 5: CONCLUSION

APPENDICES

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

BP	Broken Plurals
IR	Information Retrieval
IS	Information System
ISRI	Information Science Research Institute's
IT	Information Technology
NB	Naive Bayes
PBUH	Peace Be Upon Him
SESQ	Search Engine System for Holy Quran
SP_TREC	Suffix-Prefix Text Retrieval Conference
SP_WAL	Suffix-Prefix with Alef-Lam
SP_WOAL	Suffix-Prefix without Alef-Lam
SPS_TREC	Suffix-Prefix-Suffix Text Retrieval Conference
SPS_WAL	Suffix-Prefix-Suffix with Alef-Lam
SPS_WOAL	Suffix-Prefix-Suffix without Alef-Lam
TREC	Text Retrieval Conference
UUM	University Utara Malaysia
WWW	World Wide Web

CHAPTER ONE

INTRODUCTION

This study is organized in five chapters starting with a general overview in the first chapter, the background review of the project that outline the desire and motivation behind the project and the area or domain on which the project is based on. The chapter further describes the problem statement, the objective to be accomplished and its significant which will be derived from the project, the scope and finally outlined the way other chapters will be organized.

1.1 Background

The Holy Quran is a book sent down from Allah almighty through his messenger Muhammad Peace be Upon Him (PBUH) to mankind as guidance in their activities and to offer solution to their problems. According to Wikipeadia (2007), The Qur'ān (Arabic: القـرآن al-qur'ān, literally "the recitation"; also sometimes transliterated as Qur'an, Koran, or Al-Qur'an). It is the central religious text of Islam. Muslims believe that the Qur'an is the book of divine guidance and direction for mankind, and consider the text (in the original Arabic) to be the final divine revelation of Allah. Islam holds that the Qur'an was revealed to Muhammad by the angel Gabriel over a period of 23 years. Muslims regard the Qur'ān as the culmination of a series of divine messages that started with those revealed to Adam, who is regarded in Islam as the first prophet, and continued with the Suhuf-i-Ibrahim (Scrolls of Abraham), the Tawrat (Torah), the Zabur (Psalms) and the Injil aforementioned books are not explicitly included in the Qur'an, but are recognized in the Qur'ān.

The contents of the thesis is for internal user only

REFERENCE

Abduelbaset G., Massimo P, Anne De R. and Jeff Reynolds (2000).Identifying Broken Plurals in Unvowelised Arabic Text. In Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. Barcelona. 246-253

Applegate, L. M. .Rigor and Relevance in MIS Research.Introduction,. MIS Quarterly (23:1), March 1999, pp. 1-2.

Bakar, Z. A. and Rahman, N. A. 2003. Evaluating the effectiveness of thesaurus and stemming methods in retrieving Malay translated Al-Quran documents. In Digital Libraries: Technology and Management of Indigenous Knowledge for Global Access vol. 2911. Springer- Verlag, 653–662.

Beesley, K. 1996. "Arabic Finite-State Morphological Analysis and Generation." COLING-96, 1996.

Boudelaa , Sami; Gaskell, M. Gareth (2002). "A reexamination of the default system for Arabic plurals." Psychology Press Ltd, vol. 17, pp. 321-343, 2002.

Chris Paice. (1996). A Method for the Evaluation of Stemming Algorithms Based on Error Counting. Journal of the American Society for Information Science, 47, 632-649.

Denning, P.J.(1997). A New Social Contract for Research, In Proceeding Communications of the ACM (40:2), February 1997, pp. 132-134.

Frakes, W.(1992). Stemming algorithms. In Information Retrieval: Data Structures and Algorithms, W. Frakes and R. Baeza-Yates, Eds. Prentice-Hall, Englewood Cliffs, NJ, Chapter 8,131–160.

Goweder, Abduelbaset and De Roeck, Anne (2001)."Assessment of a Significant Arabic Corpus." ACL 2001. Arabic language Processing. pp. 73-79, 2001.

Greengrass, M., Robertson, A. M., Robyn, S., and Willett, P. Processing morphological variants in searches of Latin text. Information research news, 6 (4), pp. 2-5, 1996.

Hayder K. Al Ameed, Shaikha O. Al Ketbi, Amna A. Al Kaabi, Khadija S. Al Shebli, Naila F. Al Shamsi, Noura H. Al Nuaimi, Sh aikha S. Al Muhairi (2005). Arabic Light Stemmer: Anew Enhanced Approach

HELM, S. Closer than you think. Medicine Cornpzst. 1, 1 (1983).

Hevner, A. (2004). Design Science in Information Systems Research, Information Systems and Decision Sciences, College of Business Administration, University of South Florida.

Hull, D. (1996). Stemming algorithms a case study for detailed evaluation. JASIS, 47 (1), pp.70-84,

Jinxi X, Alexander F., Ralph W.,(2002). Empirical Studies in Strategies for Arabic Retrieval. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. Pp 269 - 274

Johansson, J. M., March, S. T., and Naumann, J. D. .Modeling Network Latency and Parallel Processing in Distributed Database Design,. Decision Sciences Journal (34:4), Fall 2003.

Kazem,T., Elkhoury, R. and Coombs, J. (2005). Arabic Stemming Without Root Dictionary. In the Proceedings of the IEEE Computer Society International Conf. on Information Technology: Coding & Computing (ITCC'05)

Khoja, S. (1999). Stemming Arabic Text. Lancaster, U.K., Computing Department, Lancaster University. Retrieved on 24th January 2008 from www.comp.lancs.uk/ computing/users/khoja/stemmer.ps.

Kraaij, W. and Pohlmann R (1996). Viewing stemming as recall enhancement. In Proceedings of ACM SIGIR96. pp.40-48,

Larkey, L. S. and Connell, M. E. (2001) "Arabic information retrieval at UMass in TREC-10." In TREC 2001.Gaithersburg: NIST, 2001.

Leah S. Larkey, Lisa Ballesteros, Margaret E. Connell (2002).Improving Stemming for Arabic Information Retrieval: Light Stemming and Co-occurrence Analysis. In Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval. Pages: 275 - 282

Leah S. Larkey, Fangfang F., Margaret C., Victor L.(2004). Language-specific Models in Multilingual Topic Tracking. In Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. Pp 402-409

Lee, A (1999). Inaugural Editor's Comments. In MIS Quarterly, pp. v-xi, March 1999

Markus, M.L., Majchrzak, A., and Gasser, L., A Design Theory for Systems that Support Emergent Knowledge Processes, MIS Quarterly September, 2002, pp. 179-212.

Mira A., Jelit A., Bobby N. Tahaghoghi S.M. and HUGH E.,(2007). Stemming Indonesian: A Confi x-Stripping Approach. In proceeding ACM Transactions on Asian Language Information Processing, V. 6, 4, 13

Mohamed K, Amine B, Tajje-E. R. (2004) Automatic Arabic Document Categorization Based on the Naïve Bayes Algorithm. In Proceeding of coling 20th Workshop on Computational Approaches to Arabic Script-based Languages, Geneva, August 23rd-7th,

Mohammed A. and Ophir F. (2004). On Arabic Search: Improving the Retrieval Effectiveness via a Light Stemming Approach. In Proceedings of the eleventh international conference on Information and knowledge management, McLean, Virginia, USA. ISBN:1-58113-492-4, pp:340 - 347

Monz, C. and de Rijke, M. (2001). Shallow morphological analysis In monolingual information retrieval for German and Italian. In Cross-language information retrieval and evaluation: Proceedings of the CLEF 2001 workshop, C. Peters, Ed.: Springer Verlag,.

Naglaa Thabet (2003). Stemming the Qur'an. School of English Literature, Language and Linguistics University of Newcastle

Noamany, M. (2001). Personal communications.

Rachidi, T., Bouzoubaa, M., ElMortaji, L., Boussouab, B. and Bensaid, A. (2003a). Arabic user search Query correction and expansion. In a proceeding of COPSTIC'03, Rabat December 11-13

Rachidi, T., Iraqi, O., Bouzoubaa, M., Ben El Khattab, A., El Kourdi, M., Zahi, A. and Bensaid, A. (2003b). Barq: Distributed multilingual Internet search engine with focus on Arabic language. In proceedings of IEEE Conf. on Sys., Man and Cyber., Washington DC, October 5-8, 2003

Robert K, and Bruce W.C. (1992) In proceeding ACM Transactions on Information Systems, Vol 10,2, pp,115-141

Roberto Navigli and Paola Velardi(2003). An Analysis of Ontology-based Query Expansion Strategies. In Proceedings of the 14th European Conference on Machine Learning.

Salton, G.(1988). Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer, Addison-Wesley, Reading, MA, 1988

Silver, M.S., Markus, M.L., and Beath, C.M.(1995)"The Information Technology Interaction Model: A Foundation for the MBA Core Course," MIS Quarterly (19:3), Septembe, pp. 361-390.

Simon, H.A. The Sciences of the Artificial, 3rd Edition, Mit Press, Cambridge, MA, 1996.

Tai, S. Y., Ong, C. S., and Abdullah, N. A. On designing an automated Malaysian stemmer for the Malay language. (poster). In Proceedings of the fifth international workshop on information retrieval with Asian languages, Hong Kong, pp. 207-208, 2000.

Tsichritzis, D (1997). The Dynamics of Innovation, Beyond Calculation: The Next Fifty Years of Computing, Copernicus, 1997, pp. 259-265.

Xiaoli Li and Bing Liu (2003).Learning to Classify Texts Using Positive and Unlabeled Data. In Proceedings of Eighteenth International Joint Conference on Artificial Intelligence (IJCAI-03)