STEMMING ALGORITHM IN SEARCHING MALAY TEXT

A Thesis submitted to Faculty of Information Technology in
partial fullfilment of the requirements for the degree
Master of Science (Information Technology),
Universiti Utara Malaysia

by

RIZAUDDIN BIN SAIAN

JABATAN HAL EHWAL AKADEMIK
(*Department of Academic Affairs*)
Universiti Utara Malaysia

PERAKUAN KERJA KERTAS PROJEK
(Certificate of Project Paper)

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

RIZAUDDIN BIN SAIAN

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

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

STEMMING ALGORITHM IN SEARCHING MALAY TEXT

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 filed is covered by the project paper).

Nama Penyelia Utama
(Name of Main Supervisor): **PROF. DR. HJH. KU RUHANA KU MAHAMUD**

Tandatangan
(Signature) : 

Tarikh
(Date) : 12/10/04.

PERMISSION TO USE

In presenting this thesis in partial fulfilment of the requirements for a post graduate degree from Universiti Utara Malaysia, I agree that Universiti 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(s) or, in their absence, by the Dean of Faculty of Information Technology. 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 Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

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

Dean of Faculty of Information Technology
Universiti Utara Malaysia
06010 UUM Sintok
Kedah Darul Aman

ABSTRACT

Stemming is one of the processes that can be used to improve performance of a search engine. It reduces the variant word forms to common forms. This project evaluates the retrieval effectiveness of stemming algorithm in searching and retrieving relevant Malay Web pages based on user natural query words. The retrieved Web pages are weighted and ranked using Inverse Document Frequency function. The retrieval effectiveness is measured using standard recall and precision. Experiments performed show that searching with stemming improves retrieval effectiveness when compared to searching without stemming algorithm.

ABSTRAK

Algorithma "Stemming" adalah suatu daripada proses yang boleh digunakan untuk meningkatkan prestasi penggunaan enjin pencarian. Ia dapat menurunkan sesuatu perkataan kepada kata dasar. Projek ini bertujuan untuk menilai sejauh mana berkesannya algorithm ini di dalam membina enjin pencari untuk laman Web berbahasa Melayu. Laman Web yang digunakan adalah dikira pemberatnya dan diperingkatkan menggunakan fungsi frekuensi dokumen tersongsang (inverse document frequency). Keberkesanan capaian adalah diukur menggunakan ukuran "recall" dan "precision" yang piawai. Eksperimen yang dijalankan menunjukkan bahawa penggunaan algorithma "stemming" di dalam enjin pencari meningkatkan keberkesanan capaian.

ACKNOWLEDGEMENT

First of all, I would like to express my gratitude to my supervisor, Prof. Dr. Ku Ruhana Ku Mahamud for her constant and careful guidance, advice and help throughout the project. I really appreciate her constant efforts on my behalf. Last but not least, this project owes much to the enormous kindness of my wife and children for contributing the amount of time that I should have been with them.

TABLE OF CONTENTS

LIST OF TABLE

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

World Wide Web (WWW) is a system of Internet servers that uses HyperText Transfer Protocol (HTTP) to transfer specially formatted documents. The documents are formatted in a language called HyperText MarkUp Language (HTML) that supports links to other documents, as well as graphics, audio, and video files. User can jump from one document to another simply by clicking on the hyperlinks.

With rapid growth of the Internet, the WWW has become one of the most important resources for obtaining information. The importance and popularity of the WWW reflects increasing sophistication of Web sites and their growing interconnectivity. Hence, it is practically impossible to thoroughly explore Web pages manually by the users. As a result, a site search engine is a necessary tool for searching information from a Web site. An efficient search engine will enable the user to search a Web site as a whole. It will make it easier for users to find information on a Web site and gives it a more professional appearance.

A Web page is a document created with HTML that is part of a group of hypertext documents – a way of presenting information in which text, sounds, images, and actions are linked together – available on the WWW. Collectively these

The contents of the thesis is for internal user only

REFERENCES

Ahmad, F., Yusoff, M. & Sembok, T. M. T. (1996). Experiments with a Stemming Algorithm for Malay Words. *Journal of the American Society for Information Science*, *47*(12), 909-918.

Cercone, N. (1978). Morphological Analysis and Lexicon Design for Natural Language Processing. *Computers and Humanities*, *11*, 199-209.

Ekmekçioglu, F. Çuna, Lynch, Michael F. & Willett, Peter (1996). Stemming and N-gram matching for term conflation in Turkish texts. Information Research, 1(1). Available at: http://informationr.net/ir/2-2/paper13.html.

Frakes, W. B. (1992). Stemming Algorithms. In W. B. Frakes and R. Baeza (Ed.), *Information Retrieval, Data Structures and Algorithms.* (pp. 131-160). Prentice Hall.

Frakes, W.B. (1984). Term Conflation for Information Retrieval. In van Rijsbergen, C. J. (Ed.), *Research and Development in Information Retrieval* (pp. 383-390). CUP: Cambridge.

Freud, G.E. & Willett, P. (1982). Online Identification of Word Variants and Arbitrary Truncation Searching Using a String Similarity Measure. *Information Technology Research and Development*, *1*, 177-187.

Hafer, M.A. & Weiss, S.F. (1974). Word Segmentation by Letter Successor Varieties. *Information Storage and Retrieval*, *10*, 371-385.

Harman, D. (1991). How Effective is Suffixing? *Journal of the American Society for Information Science*, *42*(1), 7-15.

Idris, N. & Syed Mustapha, S. M. F. D. (2001, April 23). Stemming for Term Conflation in Malay Texts. *International Conference of Artificial Intelligence, Las Vegas*, p. 1512 – 1517.

Kantrowitz, M., Mohit, B., & Mittal, V. (2000). Stemming and Its Effects on TFIDF Ranking. *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 357-359.

Kobayashi, M. & Takeda, K. (2000, June). Information Retrieval on the Web. *ACM Computing Surveys*, *32*(2), 144-173.

Lawrence, S. & Giles, C. L. (1999). Accessibility of information on the Web. *Nature*

Lennon, M., Peirce, D. S., Tarry, B. D. & Willet, P. (1981). An evaluation for some conflation algorithms for information retrieval. *Journal of Information Science*, *3*, 177-183.

Lovins, J.B. (1968). Development of a Stemming Algorithm. *Mechanical Translation and Computational Linguistics*, *11*, 22-31.

Niedermair, G.T., Thurmair, G. & Buttel, I. (1985). MARS A Retrieval Tool on the Basis of Morphological Analysis. In van Rjsbergen, C. J. (Ed.), *Research and Development in Information Retrieval* (pp. 369-380). CUP: Cambridge.

Paice, C. D. (1990). Another Stemmer. *ACM SIGIR Forum*, *24*(3), 56-61.

Pirkola, A. (2001, May). Morphological Typology of Languages for IR.. *Journal of Documentation*, *57*, 330-348.

Popovic, M. & Willet, P. (1992, June). The Effectiveness of Stemming for Natural-Language Access to Slovene Textual Data. *Journal of the American Society for Information Science*, *43*(5), 384-390.

Porter, M. F. (1980, July). An Algorithm for Suffix Stripping. *Program*, *14*(3), 130-137.

Raben, J. & Lieberman, D.V. (1976). Text comparison: principles and a program. In Jones, A

& Churchouse, R. F. (Eds.), *The computer in literacy and linguistic studies* (pp. 297-308). Cardiff: University of Wales Press.

Savoy, J. (1993, January). Stemming of French Words Based on Grammatical Catagories. *Journal of the American Society for Information Science*, *44*, 1-9.

Stephen, G.A. (1994). String Searching Algorithm. In *Lecture Notes Series on Computing*. Singapore: World Scientific Publishing Co. Pte. Ltd.

Ulmschneider, J.E. & Doszkocs, T. (1983). A Practical Stemming Algorithm for Online Search Assistance. *Online Review*, *7*, 301-318.

Van Rijsbergen, C. J. (1979). *Information Retrieval* (Second Edition). London: Butterworths.

Walker, S. & Jones, R.M. (1987). Improving Subject Retrieval in Online Catalogues. *1. Stemming, Automatic Spelling Correction and Cross-Reference Tables*, British Library Research Paper, London.

Wen Ji-Rong., Nie Jian-Yun. & Zhang Hong-Jiang. (2001, May 1). Clustering User Queries of a Search Engines. *ACM*, pp. 162-168.

Yoshiaki, M. & Keishi, T. (1999, February). Finding Context Paths for Web Pages. *Proceedings of the tenth ACM Conference on Hypertext and Hypermedia: returning to our diverse roots.*

Zainab Abu Bakar & Nurazzah Abd. Rahman (2004). Evaluating the Effectiveness of Conflation Methods in Retreiving Malay Translated Al-Quran Texts and Images. *Conference on Scientific and Social Research, UiTM.*

Zeti Zuryani Mohd Zakuan (2004). Penipuan Kad Kredit di Malaysia. LLM Thesis, Universiti Kebangsaan Malaysia.