EVALUATION OF SEARCH RESULT OF DOCUMENT SEARCH BASED GA (DSEGA)

Kamal Norfarid Kamarudin

Master of Science Intelligent System (IS)

EVALUATION OF SEARCH RESULT OF DOCUMENT SEARCH BASED GA (DSEGA)

A thesis submitted to the Faculty of Information Technology in partial fulfillment of the requirements for the degree
Master of Science (Intelligent System)
Universiti Utara Malaysia

By

Kamal Norfarid Kamarudin

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)

KAMAL NORFARID KAMARUDIN

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)

EVALUATION OF SEARCH RESULT OF DOCUMENT SEARCH BASED GA (DSEGA)

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): **MR. AZMAN YASIN**

Tandatangan
(Signature) :

Tarikh
(Date) : 29/6/04.

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a postgraduate degree from Universiti 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 purpose may be granted by my supervisor(s) or, in their absence by the Dean of the 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 Universiti Utara Malaysia for any scholarly use which may be made of any material from my thesis.

Requests 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

ABSTRAK

(BAHASA MELAYU)

Semenjak tahun 1940 an, masalah di dalam penyimpanan capaian maklumat telah menarik perhatian ramai. Capaian maklumat telah menjadi bertambah sukar. Pengenalan komputer dikatakan mampu menyelesaikan masalah ini dengan membangunkan sistem capaian maklumat. Walaupun pengumpulan dan penyimpanan maklumat menjadi lebih mudah, capaian kepada maklumat yang sesuai menjadi semakin sukar. Algoritma Genetik adalah teknik optimum yang apabila diberikan sesuatu matlamat atau fungsi fitness, ia akan mencari penyelesaian titik optimal. Penyelesaian akan dicari dengan kaedah terus and kaedah carian dipinjam dari idea ivolusi semulajadi. Secara amnya, Algoritma Genetik sangat efektif di dalam mencari penyelesaian terhadap permasalahan yang kompleks dan pelbagai demensi. Sistem DSeGA adalah sistem carian pintar di Fakulti Teknologi Maklumat, Universiti Utara Malaysia. Sistem in dibangunkan dengan menggunakan teknik capaian maklumat dan Algoritma Genetik. Sistem ini tidak pernah diuji dengan menggunakan koleksi data yang standard. Matlamat utama projek ini adalah untuk menguji system DSeGA dengan menggunakan tiga koleksi ujian standard (CACM, CRANFIELD,TIME). Projek ini memberikan penilaian terhadapat keputusan carian sistem DSeGA. Secara kesimpulannya, sistem DSeGA perlu dianalisis semula untuk mempertingkatkan tahap prestasi sistem.

ABSTRACT

(ENGLISH)

Since the 1940s the problem of information storage and retrieval has attracted increasing attention. Retrieve document becoming ever more difficult. With the advent of computers, a great deal of thought has been given to using them to provide rapid and intelligent retrieval systems. Although it has become easier to collect and store information in document collections, it has become increasingly difficult to retrieve relevant information from these large document collections. Genetic algorithms describe a set of optimization techniques that, given a goal or fitness function, are used to search a space for optimal points. The space is searched in a directed, stochastic manner, and the method of searching borrows some ideas from evolution. In practice, genetic algorithms have proven very effective in searching through complex, highly nonlinear, multidimensional search spaces. DSeGA system is an intelligent search agent toolkit at Faculty of Information Technology of Universiti Utara Malaysia. It is composed by a series of module that using information retrieval method and genetic algorithm. This toolkit does not tested by any standard test data collection. The aim of this research is to test DSeGA system with three standard data collection (Cranfield, CACM and TIME). The finding of this research is an evaluation of DSeGA system search result. It was discovered that DSEGA system cannot performed the way that the system should be. The conclusion of this research is DSeGA system need to be investigated to enhance the system performance.

ACKNOWLEDGEMENTS

Praise to Allah S.W.T, whom granted me the ability, guidance and strength to complete this project. My sincere appreciation to my lovely mother, Pn. Faridah Abd. Ghani, my father En. Kamarudin Muhd Yasin and family for all of their support, devotion and encouragement. Their love and concern is the energy of all my effort.

I am grateful to my supervisor, En. Azman Yasin for his continuous assistant and guidance during the research, preparations and until the completions of this project. Also to En Azizi Ab. Aziz, thank you for your comment and advice.

My heartfelt gratitude to Norhasyimah for her sweetest support, patient and understanding over the entire period of my study. To Hasnatul Nazuha, many thanks for your endless encouragement and help. Your support has meant more to me than I can express.

Equally deserving of this recognition are Wan Faridah Hanum, Hazlinda Ghazali, Shariful Hafizi, Muhamad Arif, Iswandi and Fakhrul, thank you for the support that you all have been given me towards the completion of this project.

TABLE OF CONTENT

LIST OF TABLE

LIST OF FIGURE

CHAPTER ONE

INTRODUCTION

Rapid advances in science and technology in the last three decades have leads us to call our society in information society. More information is generated about topics than ever before. In this complicated society, we often need relevant information to carry out the tasks at hand and to make intelligent decisions. From a large amount of data it is difficult to find actually needed data at a given time, and to distinguish relevant from extraneous data. The research area called information retrieval (IR) was established in the early 1960s to develop computer-aided effective processes of searching and extracting specific information.

There are three important paradigms of research in the area of IR: Probabilistic IR, Knowledge-based IR, and Artificial Intelligence based techniques like neural networks and genetic algorithm (GA). GA is based on the Darwinian principles of natural selection. GA method of searching borrows ideas from evaluation. An implementation of GA contains a population of individuals, fitness function, generations and population. In practice, genetic algorithms have proven very effective in searching through complex, highly nonlinear, multidimensional search spaces (Goldberg, 1988). Several researchers like Martin-Bautista and Vila (1999),

REFERENCES

Blair, D.C. and Maron, M.E. (1985). An evaluation of retrieval effectiveness for a full text document retrieval systems. *Communications of ACM,* 28(3):289-299.

Chen, H., Chung, Y., Ramsey, M. & Yang, C.C. 1998(a). A Smart Itsy Bitsy Spider for the Web. *JASIS* 49(7): 604-618

Chen, H., Shankaranarayanan, G., She, L. & Iyer, A. 1998(b). A Machine Learning Approach to Inductive Query by Examples: An Experiment Using Relevance Feedback, ID3, Genetic Algorithms, and Simulated Annealing. *JASIS* 49(8): 693-705.

Chen, H. and Dhar, V., (1991). Cognitive process as a basis for intelligent retrieval systems design. *Information Processing Management,* 27(5):405-432.

Chen, H., Fan, H., Chau, M., and Zeng, D., (2001), MetaSpider: Meta-Searching and Categorization on the Web. *Journal of the American Society of Information Science*

D.E. Goldberg (1989). *Genetic Algorithms in Search, Optimization and Machine Learning.* Addision-Wesley, Reading, MA.

Du, H and Crestani, H. (2002). *Retrieval Effectiveness of Written and Spoken Queries: An Experiment Evaluation,* University of Strachclyde, UK.

Fan, W., Gordon, M. & Pathak, P. (2000). *Personalization of search engine services for effective retrieval and knowledge management.* ICIS, hlm. 20-34

Gordon, M., (1988). Probabilistic and genetic algorithms for document retrieval. *Communication of the ACM*, 31(10):1208-1218.

Le Roy *et.al.* (2003). The Used of Dynamic Context to Improve Casual Internet Search. *Journal of the ACM,* 21(3):229-253.

Martin-Bautista, M.J., Vila, M. & Larsen, H.L. (1999). A Fuzzy Genetic Algorithm Approach to an Adaptive Information Retrieval Agent. *Journal of the American Society for Information Science.* 50(9):760-771

Nick, Z.Z and Themis, P. (2001). Web Search Using Genetic Algorithm. *IEEE Internet Computing, 2001.*

Oren, R., (2002), Reexamining tf.idf based information retrieval with Genetic Programming, *Proc. of SAICSIT*: 224–234.

Oren, R., (2000), Improving the Effectiveness of Information Retrieval with Genetic Programming, Thesis Report.

Quinlan, J.R. (1979). Discovering rules by induction from large collections of examples. In *Expert System in the Micro-electronic Age,* p168-201, Michie,D., Editor, Edinburgh univ. Press.

Rumelhart, D.E. Hinton, G.E. and J.Williams (1986). Learning internal representations by error propagation. In *Parallel Distributed Processing,* p318-362, MIT Press, Cambridge, MA.

Sarwar *et.al* (2000). Analysis of Recommendation Algorithms for E-Commerce *Journal of the ACM,* 17-20.

Simon, H., (1991). Artificial Intelligence : wkere has it been, and where is it going? *IEEE Transactions on Knowledge and Data Eng:neering,* 3(2):128-136.

Shamrie Sainin (1999). *Agen Carian: Carian Dokumen Berasakan Genetik Algoritma,* Laporan Projek Latihan Industr, Universiti Utara Malaysia.

Weiss, S.M. and Kulikowski, C.A. (1991). *Computer Systems That Learn : Classification and Prediction Method from Statistics, Neural Networks, Machine Learning and Expert Systems.* Morgan Kaufman publishers, Inc., San Mateo, CA.

Yang, J. & Korfhage, R.R. (1993). *Effects of query term weights modification in document retrieval: a study based on a genetic algorithm.* Proceedings of the Second Annual Symposium on Document Analysis and Information Retrieval, Las Vegas. hlm 271-285.

P.J.M. van Laarhoven and E.H.L. Aarts (1988). *Simulated Annealing : Theory and Applications.* D. Reidal Publishing Company, Dordrecht.

Wu, J.L. and Agogino, A.M, (2004), Automatic Key Phrase Extraction with Multi-Objectives Genetic Algorithm, *Proc. of Hawai International of Science*, 2004.

William R. Hersh, Diane L. Elliot, David H. Hickam, Stephanie L. Wolf, and Anna Molnar (1995). Towards new measures of information retrieval evaluation. In *Proceedings of the 18th Annual InternationalACM SIGIR Conference on Research and Development in Information Retrieval*, 164-170

Yuri, K. and Jochen, R.M, (2003), Current Status of Evaluation of Information Retrieval, *Journal of Medical System*, 27(5).