RELEVANCY OF THE SEARCH RESULT USING AI TECHNIQUES

A Master Project submitted to the Graduate School in partial
Fulfillment of the requirements for the degree
Master of Science (Information Technology)
Universiti Utara Malaysia

By
Tan Hong Keat

Sekolah Siswazah
(Graduate School)
Universiti Utara Malaysia

PERAKUAN KERJA KERTAS PROJEK
(Certification of Project Paper)

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

TAN HONG KEAT

calon untuk Ijazah
(candidate for the degree of) Sarjana Sains (Teknologi Maklumat)

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

RELEVANCY OF THE SEARCH RESULT USING AI TECHNIQUES

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 *(Name of Supervisor)* : Puan Fadzilah bt. Siraj

Tandatangan *(Signature)* : Fadzilah

Tarikh *(Date)* : 10-10-2001

PERMISSION TO USE

In presenting this thesis in partial fulfillment of the requirements for a post graduate degree from the 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 purposes may be granted by my supervisor 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 Graduate School
Universiti Utara Malaysia
06010 UUM Sintok
Kedah Darul Aman

ABSTRACT (BAHASA MALAYSIA)

Objektif utamanya alat bantu carian ialah untuk meningkatkan nisbah ketepatan dokument berbanding bilangan dockument yang diperolehi. Projek ini bertujuan untuk membangunkan satu alat bantu pintar bagi menentukan kerelevanan suatu hasil carian berbanding kata kunci pengguna "*Northern Light Power Search*" enjin digunakan bagi kata kunci tertentu yang berkaiting dengan rangkaian nerual. Halaman hyperteks dan bidang komputer sains adalah fokus bagi projek ini. Satu model pintar yang boleh mengklaskan samada relevan atau tidak relevan kepada kata kunci tertentu telah dibangunkan. Pembangunan perisian dibahagikan kepada dua bahagian. Bahagian pertama menumpukan terhadap pembangunan perisian bagi mengira dan menkategorikan kata kunci dalam format pra-definisi. Bahagian kedua pula, menumpukan terhadap pembangunan perisian bagi model rangkaian neural dengan keupayaan "*auto-determining*". Bahasa pengaturacaraan Java digunakan untuk membangunkan perisian tersebut. Perceptron berlapis digunakan sebagai mewakili rangkaian neural yang telah diimplimentasikan dalam project ini. Lapisan generik dan notasi rumus rangkaian neural diperolehi dari model klasikal. Sebelum perisian Perceptron Berlapis dibangunkan, data dari perisian kaunter kata kunci hyperteks telah digunakan dalam "*Neural Connection*" untuk mengenalpasti hasil pengujianan terbaik yang dapat diperolehi. Hasil yang diperolehi daripada Neural Connection mencapai lebih daripada 96%. Walau bagaimanapun, hasil yang diperolehi daripada perisian yang telah dibangunkan berkurangan sebanyak 15%. Ini mungkin disebabkan perisian yang telah diimplementasi menggunakan fungsi aktivasi tidak linear, di lapis tersembunyi dan lapisan output.

ABSTRACT (ENGLISH)

For any search tools, the main objective is to improve the ratio of the number of hits to number of retrievals. The objective of this project is to develop an intelligent tool for determining the relevancy of the search result, with respect to users' keyword. Northern Light Power Search engine was used with specific keywords that are related to neural network. Hypertext and computer science domain was the focus of this study. An intelligent model that can categorize search result as relevant or irrelevant to the keyword specified was developed. This software development was divided into two parts, the first part concentrated on software development to count and categorized keyword in pre-defined format. The second part focus on software development for neural network model with auto-determining capability. Java programming language was used as the programming language to develop the software. Multilayer Perceptron was utilizes as the neural network model implemented in this study. Generic layer and notation of neural network formula were derived from classical model. Prior to the Multilayer Perceptron software development, the data from hypertext keyword counter software was used in "Neural Connection" to confirm the best result that could be achieved. The result from "Neural Connection" has achieved more than 96%. However, the results produced by the developed software decreased by 15%. This may due to the fact that the developed software used non-linear activation function at hidden as well as the output layer.

ACKNOWLEDGEMENTS

This section specially dedicated to Pn. Fadzilah Siraj, my supervisor for this project. Thanks for the help, advice, and guidance.

Besides, I would like to thank my lecturer En Zambri Saad, who provided my the knowledge of Java programming. I would like to thank Universiti Utara Malaysia for giving me the opportunity to further study as part-time student.

Finally, thanks to my family member, especially my wife, who always supports with a helping hand.

TABLE OF CONTENT

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 An Overview

In the past, the main sources of information are newspaper, magazine, journal, book, etc. Presently, the World Wide Web (WWW) has become a global source of information in all areas of users' interest. The sources are ranging from commerce to science (Oliveira, Resende & Lehmann, 1999). For this reason engine is become popular in WWW.

As the popularity of the Internet and WWW grows, people begin to experience the pressure of information explosion. Hunting for information on the web becomes more important than ever before (Li & Rafsky, 1996).

Egyhazy, Plunkett and Thompson have identified four generations of information retrieval tools that assists people in searching the WWW. The first generation provided access to references to the end documents rather that to the documents themselves, and indexing and searching were thus applied to document surrogates, such as title or abstracts. These tools require human effort to collect, arrange, code and annotate the various resources. The primary benefit of this tool is providing users with easy browsing capability. The second generation of tool attempts to collect and index resources as an automated function. It reduces the amount of human effort. The third generation deals with WWW search engine, such

The contents of the thesis is for internal user only

Reference

Boyan,J., Freitag,D. & Joachims,T. (1996). A Machine Learning Architecture for Optimizing Web Search Engines. [online] Available: http:// WWW.cs.cmu.edu/afs/cs.cmu.edu/project/reinforcement/paper/boyan.laser.ps (21-March,2001)

Egyhazy,C.J., Plunkett,T.K. & Thompson,D.M. (1998). Intelligent Web Search Agent. [online] Available: http://csgrad.cs.vt.edu/~tplunket/article.html (20-March, 2001)

Lewis,D.D. (1991). Learning in Intelligent Information Retrieval. [online] Available: http://cora.whizbang.com/cgi-bin/details.cgi?id=16220&from=hier_html.cgi (21-March, 2001)

Li,Y. & Rafsky,L. (1996). Beyond Relevance Ranking: Hyperlink Vector Voting. [online] Available: http://la.lti.cs.cmu.edu/callan/Workshops/nir97/li/ps.gz (21-March,2001)

Mueller, M.E.(1999). Information Retrieval in The World Wide Web. [online] Available: http://mir.cl-ki.uni-osnabrueck.de/~martin/onlinepubs/paam-99/node2.html (6-March,2001)

Oliveira,C., Resende,L.G.V. & Lehmann,R. (1999). Interactive Query Expansion in a Meta-search Engine. [online] Available: http://ipanema.ime.eb.br/~de9/RelTec/1999/Rt037-99.PDF (6-March,2001)

Smeaton,A.F. & Crimmins,F. (1997). Relevance Feedback and Query Expansion for Searching the Web: A Model for Searching a Digital Library. In Oliveira,C., Resende,L.G.V. & Lehmann,R. (1999). Interactive Query Expension in a Meta-search Engine. [online] Available: http://ipanema.ime.eb.br/~de9/RelTec/1999/Rt037-99.PDF (6-March,2001)

Nikolaev, N. (2001a). CIS311 Nueral Networks. [online] Available: http://homepages.gold.ac.uk/nikolaev/cis311.htm

Nikolaev, N. (2001b). Multilayer Perceptrons – 1.The Multilayer Perceptron. [online] Available: http://homepages.gold.ac.uk/nikolaev/311multi.htm. In Nikolaev, N. (2001). CIS311 Nueral Networks. [online] Available: http://homepages.gold.ac.uk/nikolaev/cis311.htm