

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

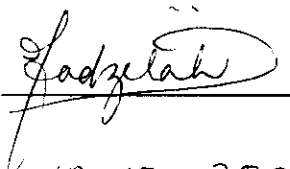
seperti yang tercatat di muka surat tajuk dan kulit kertas projek
(as it appears on the title page and front cover of project paper)

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

Tandatangan
(Signature) :



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 neural. Halaman hypertexts 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 mengkategorikan 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 hypertexts 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

PERMISSION TO USE	I
ABSTRACT (BAHASA MALAYSIA)	II
ABSTRACT (ENGLISH)	III
ACKNOWLEDGEMENTS	IV
TABLE OF CONTENT	V
LIST OF TABLES	XII
LIST OF FIGURES	XIX
CHAPTER 1: INTRODUCTION	1
1.1 AN OVERVIEW	1
1.2 PROBLEM STATEMENT	3
1.3 OBJECTIVE	4
1.4 SCOPE OF THIS PAPER	4
1.5 OUTPUT	4
CHAPTER 2: LITERATURE REVIEW	5
2.1 AN OVERVIEW	5
2.2 STATISTICAL APPROACH	5
2.3 ONTOLOGY	6
2.4 HYPERLINK VECTOR VOTING METHOD	6
2.5 LASER	9
2.6 VECTOR SPACE RETRIEVAL MODEL	10
2.7 BAYESIAN METHOD	10

2.8	MULTILAYER PERCEPTRON	11
2.8.1	The Multilayer Perceptron	11
2.8.1.1	Differentiable Activation Functions	11
2.8.1.2	Multilayer Network Structure	11
2.8.1.3	Representation Power of MLP	13
2.8.2	Backpropagation Learning Algorithm	14
2.8.2.1	Back-propagation training algorithm	15
2.8.3	Derivation of the Backpropagation Algorithm	16
2.8.4	Momentum in back propagation	19
CHAPTER 3: METHODOLOGY		20
3.1	AN OVERVIEW	20
3.2	PHASE 1: SPECIFICATION DETERMINATION	20
3.3	PHASE 2: BASELINE DETERMINATION	20
3.4	PHASE 3: WORD PARSING SOFTWARE DEVELOPMENT AND DATA PREPROCESSING	21
3.4.1	Format Setting	21
3.4.2	Keyword Setting.	22
3.4.3	Data Preprocessing	22
3.5	PHASE 4: TRAINING AND DETERMINATION OF BEST CONFIGURATION WITH NEURAL CONNECTION SOFTWARE	23
3.6	PHASE 5: DEVELOP MODEL SIMULATION SOFTWARE	24
3.6.1	Variable Notation	26
	Neural network archetecture	26
3.6.3	Algorithm of MultiLayer Perceptron Training	27
3.6.4	Network Setting algorithm.	28
3.6.4.1	Determination of Hidden unit	29
3.6.4.2	Determination of Learning Rate.	29
3.6.4.3	Determination of Learning momentum.	29
3.6.4.4	Determination of activation function	29
3.6.4.5	Determination of maximum update.	30
3.6.4.6	Determination of Stopping Criteria	30
CHAPTER 4: RESULT		31
4.1	AN OVERVIEW	31

4.2	NEURAL CONNECTION'S TRAINING RESULT WITH COMPOUND KEYWORD	32
4.2.1	Determining Hidden unit	32
4.2.2	Second Hidden layer determination:	33
4.2.3	Determination of Learning Rate	33
4.2.4	Determination of Momentum	34
4.2.5	Determination of Activation Function	34
4.2.6	Determination of Maximum (Max) Update	35
4.2.6.1	With Sigmoid	36
4.2.6.2	With Tanh	36
4.2.7	Determination of Stopping Criteria	37
4.3	NEURAL CONNECTION TRAINING RESULT WITHOUT COMPOUND KEYWORD	39
4.3.1	Determining Hidden unit	39
4.3.2	Second layer Hidden layer determination:	39
4.3.3	Determination of Learning Rate	40
4.3.4	Determination of Momentum	41
4.3.5	Determination of Activation Function	42
4.3.6	Determination of Max Update	43
4.3.7	Determination of Stopping Criteria	44
4.3.7.1	RMS Error Stopping criteria Determination Result Summary with Maximum Max-Update: (32000, 32000, 32000, 32000)	45
4.3.7.2	RMS Error Stopping Criteria Determination Result Summary for Default Max Update: (500,500,500,4000)	45
4.4	OPTIMUM NETWORK SETTING FOR NEURAL CONNECTION	46
4.5	MLP.JAVA TEST RESULT FOR REV 13 (8:1:1 RATIO DATA)	47
4.5.1	Determination of Hidden Unit	47
4.5.2	Determination of Learning Rate	49
4.5.3	Determination of Learning Momentum	50
4.5.4	Determination of Activation Function	51
4.5.5	Determination of Maximum Update	52
4.5.6	Determination of Stopping Criteria Accuracy	54
4.5.7	Determination of Stopping Criteria RMS Error	55
4.6	MLP.JAVA TEST RESULT FOR REV 13 6:2:2 RATIO DATA	56
4.6.1	Determination of Hidden Unit	56

4.6.2	Determination of Learning Rate	57
4.6.3	Determination of Learning Momentum	59
4.6.4	Determination of Activation Function	60
4.6.5	Determination of Maximum Update	61
4.6.6	Determination of Stopping Criteria Accuracy	63
4.6.7	Determination of Stopping Criteria RMS Error	64
CHAPTER 5: DISCUSSION		65
5.1	AN OVERVIEW	65
5.2	HYPertext COUNTER	65
5.2.1	Flexibility of neural connection software	66
5.2.2	Long training time	66
5.2.3	Large combination of formats	67
5.3	MULTI-LAYER PERCEPTRON WITH AUTO-DETERMINATION FOR NETWORK SETTING.	67
5.4	CONSTRAINT OF THE PROJECT	69
5.4.1	Many complication mathematics formula	69
5.4.2	Justification for relevant site	69
5.4.3	Percentage relevant side to low.	70
5.4.4	More file type format determination	70
CHAPTER 6: CONCLUSION		71
6.1	AN OVERVIEW	71
6.2	RECOMMENDATION OF FUTURE WORK	73
6.2.1	Justification of relevant site	73
6.2.2	Collect more relevant site	73
6.2.3	Develop software to categories format in different file type	73
6.2.4	Reduction of network setting determination	74
6.2.5	More intelligent stopping criteria	74
6.2.6	Reduction of input unit	74
REFERENCE		75
APPENDIX A: KEYWORD COUNTED		77

8.1	ACCURACY OF SITE SEARCH	77
8.2	FORMAT SETTING OR KEYWORD COUNT.	95
8.3	BY USING COMPOUND WORD	96
8.3.1	Data Selecting for Compound Word Count	108
8.4	BY USING KEYWORD	112
8.4.1	Data Selecting for keyword Count	123
8.5	STATISTICAL STUDY FOR INPUT DATA SET.	128
8.5.1	Rev 1.1 (Compound Keyword) All sites' Input Data Statistics	129
8.5.2	Rev 1.2 (Keyword) All sites' Input Data Statistics	130
8.5.3	Rev 1.1 (Compound Keyword) 100 sites' Input Data Statistics	131
8.5.4	Rev 1.2 (Keyword) 100 sites' Input Data Statistics	132
APPENDIX B: NEURAL NETWORK TRAINING RESULT WITH COMPOUND KEYWORD		134
9.1	DETERMINING HIDDEN UNIT	134
9.1.1	First layer Hidden layer determination:	135
9.1.1.1	First hidden layer result summary	137
9.1.2	Second layer Hidden layer determination:	137
9.1.2.1	Second hidden layer result summary	138
9.2	DETERMINATION OF LEARNING RATE	139
9.2.1	Learning Rate Result Summary	140
9.3	DETERMINATION OF MOMENTUM	141
9.3.1	Learning Momentum Summary	142
9.4	DETERMINATION OF ACTIVATION FUNCTION	143
9.5	DETERMINATION OF MAXIMUM (MAX) UPDATE	144
9.5.1	Activation function is sigmoid:	144
9.5.1.1	Maximum Update Summary	145
9.5.2	Activation function is Tanh:	146
9.5.2.1	Maximum Update Summary	146
9.6	DETERMINATION OF STOPPING CRITERIA	148
9.6.1	RMS Stopping Criteria Summary	150
APPENDIX C: NEURAL NETWORK TRAINING RESULT WITHOUT COMPOUND KEYWORD		151

10.1 DETERMINING HIDDEN UNIT	151
10.1.1 First layer Hidden layer determination:	152
10.1.1.1 First hidden layer result summary	153
10.1.2 Second layer Hidden layer determination:	154
10.1.2.1 Summary	155
10.2 DETERMINATION OF LEARNING RATE	156
10.2.1 Learning Rate Result Summary	157
10.3 DETERMINATION OF MOMENTUM	158
10.3.1 Summary	159
10.4 DETERMINATION OF ACTIVATION FUNCTION	160
10.5 DETERMINATION OF MAX UPDATE	161
10.5.1 Summary	164
10.6 DETERMINATION OF STOPPING CRITERIA	164
10.6.1 RMS Result Summary with Max Update: (32000, 32000, 32000, 32000)	167
10.6.2 RMS Result Summary for Max Update: (500,500,500,4000)	169
APPENDIX D: “HYTXXCNT.JAVA” FILE ALGORITHM	170
11.1 IMPORT JAVA FUNCTION:	170
11.2 CONSTANT AND VARIABLE	170
11.3 METHODS	171
11.3.1 Algorithm of main method	172
11.3.2 Algorithm for getConfig method	173
11.3.3 algorithm for getFormat method	174
11.3.4 Algorithm for GetKeyword method	175
11.3.5 Algorithm for ResetSearchCount	176
11.3.6 Algorithm for DetermineFileType Method	177
11.3.7 Algorithm for HypertxtCounter method	178
11.3.8 Algorithm for TextCOunter Method	179
11.3.9 Algorithm for getFormatIndex method	180
11.3.10 Algorithm for getKeywordIndex method	181
11.4 “HYTXXCNT.JAVA” SOURCE CODE	182

APPENDIX E: MLP.JAVA SOFTWARE	189
12.1 MLP.JAVA SOURCE CODE	192
12.2 MLP.JAVA RESULT RAW DATA	215
12.2.1 Re12_811.out Raw Result	215
12.2.2 Re13_811.out raw result	217
12.2.3 Re13_622.out raw result	225

LIST OF TABLES

Table 3.1: Format setting.	21
Table 3.2: Neural Connection learning setting.	23
Table 4.1: Network setting to determine hidden unit.	32
Table 4.2: Result summary for determining first hidden unit.	32
Table 4.3: Second hidden layer nodes	33
Table 4.4: Network setting to determine learning rate.	33
Table 4.5: Learning rate determination result	33
Table 4.6: Network setting to determine learning momentum.	34
Table 4.7: Learning momentum determination result	34
Table 4.8: Network setting for activation function determination	34
Table 4.9: Activation function determination result.	35
Table 4.10: Network setting to determine maximum update	35
Table 4.11: Maximum update determination result, with sigmoid.	36
Table 4.12: Maximum update determination result, with tanh	36
Table 4.13: Network setting to determine stopping criteria.	37
Table 4.14: Stopping criteria accuracy result	37
Table 4.15: RMS error determination result.	38
Table 4.16: Hidden unit determination result	39
Table 4.17: First hidden layer's hidden unit determination result.	39
Table 4.18: Second layer hidden unit determination result	40
Table 4.19: Network setting to determine learning rate	40
Table 4.20: Learning rate determination result	40

Table 4.21: Network setting to determine learning momentum	41
Table 4.22: Learning momentum determination result	41
Table 4.23: Network setting to determine activation function.	42
Table 4.24: Activation function determination result	42
Table 4.25: Network setting to determine maximum update	43
Table 4.26: Maximum update determination result	43
Table 4.27: Network setting to determine stopping criteria	44
Table 4.28: Accuracy stopping criteria determination result	44
Table 4.29: RMS Error Stopping criteria Determination Result Summary with Maximum Max-Update: (32000, 32000, 32000, 32000)	45
Table 4.30: RMS Error Stopping Criteria Determination Result Summary for Default Max Update: (500,500,500,4000)	45
Table 4.31: Optimum Network setting for neural connection	46
Table 4.32: Best result obtained by neural connection.	46
Table 4.33: Initial Network setting to determine Hidden Unit	47
Table 4.34: Hidden Unit testing Result Summary	47
Table 4.35: Network setting Result After determined Hidden Unit	48
Table 4.36: Best Average and standard Deviation for accuracy and RMS error	48
Table 4.37: Cross Tabulation Table for Best testing and training result	48
Table 4.38: Learning Rate testing Result Summary	49
Table 4.39: Network setting Result after Learning Rate is determined	49
Table 4.40: Best Average and standard Deviation for accuracy and RMS error	50
Table 4.41: Cross Tabulation Table for Best testing and training result	50
Table 4.42: Learning Momentum testing Result Summary	50
Table 4.43: Network setting Result after determined Learning Momentum	50

Table 4.44: Best Average and standard Deviation for accuracy and RMS error	51
Table 4.45: Cross Tabulation Table for Best testing and training result	51
Table 4.46: Activation Function testing Result Summary	51
Table 4.47: Network setting Result After determined Activation Function	52
Table 4.48: Best Average and standard Deviation for accuracy and RMS error	52
Table 4.49: Cross Tabulation Table for Best testing and training result	52
Table 4.50: Maximum Update testing Result Summary	52
Table 4.51: Network setting Result after determined Maximum Update	53
Table 4.52: Best Average and standard Deviation for accuracy and RMS error	53
Table 4.53: Cross Tabulation Table for Best testing and training result	53
Table 4.54: Stopping Criteria Accuracy testing Result Summary	54
Table 4.55: Network setting Result after determined Stopping Criteria Accuracy	54
Table 4.56: Best Average and standard Deviation for accuracy and RMS error	54
Table 4.57: Cross Tabulation Table for Best testing and training result	54
Table 4.58: RMS Error testing Result Summary	55
Table 4.59: Network setting Result after determined Stopping Criteria RMS Error	55
Table 4.60: Best Average and standard Deviation for accuracy and RMS error	55
Table 4.61: Cross Tabulation Table for Best testing and training result	56
Table 4.62: Hidden Unit testing Result Summary	56
Table 4.63: Network setting Result after determined best Hidden Unit	57
Table 4.64: Best Average and standard Deviation for accuracy and RMS error	57
Table 4.65: Cross Tabulation Table for Best testing and training result	57
Table 4.66: Learning Rate Testing Result Summary	57
Table 4.67: Network setting Result after determined Learning Rate	58

Table 4.68: Best Average and standard Deviation for accuracy and RMS error	58
Table 4.69: Cross Tabulation Table for Best testing and training result	58
Table 4.70: Learning Momentum testing Result Summary	59
Table 4.71: Network setting Result after determined Learning Momentum	59
Table 4.72: Best Average and standard Deviation for accuracy and RMS error	59
Table 4.73: Cross Tabulation Table for Best testing and training result	60
Table 4.74: Activation Function testing Result Summary	60
Table 4.75: Network setting to determine Activation Function	60
Table 4.76: Best Average and standard Deviation for accuracy and RMS error	61
Table 4.77: Cross Tabulation Table for Best testing and training result	61
Table 4.78: Network setting to determine Maximum Update	61
Table 4.79: Maximum Update testing Result Summary	61
Table 4.80: Network setting Result after determined Maximum Update	62
Table 4.81: Best Average and standard Deviation for accuracy and RMS error	62
Table 4.82: Cross Tabulation Table for Best testing and training result	62
Table 4.83: Stopping Criteria Accuracy testing Result Summary	63
Table 4.84: Network setting Result after determined Stopping Criteria Accuracy	63
Table 4.85: Best Average and standard Deviation for accuracy and RMS error	63
Table 4.86: Cross Tabulation Table for Best testing and training result	63
Table 4.87: RMS error testing Result Summary	64
Table 4.88: Network setting Result after determined Stopping Criteria RMS error	64
Table 4.89: Best Average and standard Deviation for accuracy and RMS error	64
Table 4.90: Cross Tabulation Table for Best testing and training result	64
Table 8.1: Page, URL, Title and relevancy.	77

Table 8.2:	Summary of Site Relevancy	94
Table 8.3:	Format Setting for Keyword count.	95
Table 8.4:	Compound Keyword Count table.	96
Table 8.5:	Compound Keyword Count Data selection output.	108
Table 8.6:	Keyword Count table.	112
Table 8.7:	Keyword Count Data selection output.	124
Table 8.8:	305 SitesCompound Keyword Count Input Statistics	129
Table 8.9:	305 Sites Keywords Count Input Statistics	130
Table 8.10:	100 Sites Compound Keyword Count Input Statistics	131
Table 8.11:	100 Sites Keywords Count Input Statistics	132
Table 9.1:	Format setting for Compound word count.	134
Table 9.2:	Configuration Setting For First Hidden Layer Node Study.	134
Table 9.3:	Preliminary Study Of First Hidden Layer Node.	135
Table 9.4:	Result Of First Hidden Layer Study	136
Table 9.5:	Summary of first hidden layer node study.	137
Table 9.6:	Preliminary study of second hidden layer node.	137
Table 9.7:	Second hidden layer node study	138
Table 9.8:	Result summary for second hidden layer study.	138
Table 9.9:	Configuration Setting For Learning Rate Coefficient Study	139
Table 9.10:	Preliminary of Learning Rate.	139
Table 9.11:	Result of Learning Rate study	139
Table 9.12:	Summary of Learning Rate Study.	140
Table 9.13:	Configuration Study For Learning Momentum Study.	141
Table 9.14:	Preliminary Study for Learning Momentum.	141
Table 9.15:	Result for Learning Momentum study.	141

Table 9.16: Summary of Learning Momentum Result.	142
Table 9.17: Configuration Setting For Activation Function Study.	143
Table 9.18: Result of Activation Function Study.	143
Table 9.19: Configuration Setting for Maximum Update Setting	144
Table 9.20: Result of Maximum Update Study for Sigmoid Function.	144
Table 9.21: Summary of Maximum Update Study Result for Sigmoid Function	145
Table 9.22: Result of Maximum Update Study for Tanh Function.	146
Table 9.23: Summary of Maximum Update study Result for Tanh Function.	146
Table 9.24: Configuration Setting for Stopping Criteria.	148
Table 9.25: Preliminary Study of Stopping Criteria Accuracy Percentage.	148
Table 9.26: Preliminary Study of RMS error.	148
Table 9.27: Result of RMS Error study.	149
Table 9.28: Summary of RMS Error Study Result.	150
Table 10.1: Format setting for keyword count	151
Table 10.2: Configuration Setting For First Hidden Layer Node Study.	151
Table 10.3: Preliminary Study Of First Hidden Layer Node.	152
Table 10.4: Result Of First Hidden Layer Study	152
Table 10.5: Summary of first hidden layer node study.	153
Table 10.6: Preliminary study of second hidden layer node.	154
Table 10.7: Second hidden layer node study.	154
Table 10.8: Result summary for second hidden layer study.	155
Table 10.9: Configuration Setting For Learning Rate Coefficient Study	156
Table 10.10: Preliminary of Learning Rate.	156
Table 10.11: Result of Learning Rate study	156
Table 10.12: Summary of Learning Rate Study.	157

Table 10.13: Configuration Study For Learning Momentum Study	158
Table 10.14: Preliminary Study for Learning Momentum.	158
Table 10.15: Result for Learning Momentum study.	158
Table 10.16: Summary of Learning Momentum Result.	159
Table 10.17: Configuration Setting For Activation Function Study.	160
Table 10.18: Result of Activation Function Study.	160
Table 10.19: Configuration Setting for Maximum Update Setting	161
Table 10.20: Preliminary Study of Maximum Update.	161
Table 10.21: Result of Maximum Update Study.	162
Table 10.22: Summary of Maximum Update Study Result.	164
Table 10.23: Configuration Setting for Stopping Criteria.	164
Table 10.24: Preliminary Study of Stopping Criteria Accuracy Percentage.	165
Table 10.25: Preliminary Study of RMS error for Max Update (32000x4).	165
Table 10.26: Result of RMS Error study for Max Update (32000x4).	165
Table 10.27: Summary of RMS Error Study Result for Max Update (32000x4)	167
Table 10.28: Preliminary Study of RMS error for Max Update (500x3,+4000).	167
Table 10.29: Result of RMS Error study for Max Update (500x3+4000).	167
Table 10.30: Summary of RMS Error Study Result for Max Update (500x3 + 4000).	169
Table 12.1: Network Setting File Variable	191

LIST OF FIGURES

Figure 2.1:	Multilayer Perceptron (MLP)	12
Figure 3.1:	Generic multilayer perceptron network architecture notation	26
Figure 11.1:	Main method algorithm	172
Figure 11.2:	getConfig() algorithm	173
Figure 11.3:	GetFormat algorithm	174
Figure 11.4:	GetKeyword() algorithm	175
Figure 11.5:	ResetSearchCount() algorithm	176
Figure 11.6:	DetermineFileType() algorithm	177
Figure 11.7:	HypertextCounter() algorithm	178
Figure 11.8:	TextCounter algorithm	179
Figure 11.9:	getFormatIndex algorithm	180
Figure 11.10:	getKeywordIndex algorithm	181
Figure 11.11:	Hytxtcnt.java Source code	188
Figure 12.1:	File content for "re12_811.nst".	191

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 than 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 Expansion 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 Neural Networks. [online] Available: <http://homepages.gold.ac.uk/nikolaev/cis311.htm>

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