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

© Tan Hong Keat, 2001. All rights reserved
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**

seerti 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) : [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)

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

2.8.1 The Multilayer Perceptron

2.8.1.1 Differentiable Activation Functions
2.8.1.2 Multilayer Network Structure
2.8.1.3 Representation Power of MLP

2.8.2 Backpropagation Learning Algorithm

2.8.2.1 Back-propagation training algorithm

2.8.3 Derivation of the Backpropagation Algorithm

2.8.4 Momentum in back propagation

CHAPTER 3: METHODOLOGY

3.1 AN OVERVIEW

3.2 PHASE 1: SPECIFICATION DETERMINATION

3.3 PHASE 2: BASELINE DETERMINATION

3.4 PHASE 3: WORD PARSING SOFTWARE DEVELOPMENT AND DATA PREPROCESSING

3.4.1 Format Setting

3.4.2 Keyword Setting.

3.4.3 Data Preprocessing

3.5 PHASE 4: TRAINING AND DETERMINATION OF BEST CONFIGURATION WITH NEURAL CONNECTION SOFTWARE

3.6 PHASE 5: DEVELOP MODEL SIMULATION SOFTWARE

3.6.1 Variable Notation

Neural network architecture

3.6.3 Algorithm of MultiLayer Perceptron Training

3.6.4 Network Setting algorithm.

3.6.4.1 Determination of Hidden unit

3.6.4.2 Determination of Learning Rate.

3.6.4.3 Determination of Learning momentum.

3.6.4.4 Determination of activation function

3.6.4.5 Determination of maximum update.

3.6.4.6 Determination of Stopping Criteria

CHAPTER 4: RESULT

4.1 AN OVERVIEW
4.2 NEURAL CONNECTION'S TRAINING RESULT WITH COMPOUND KEYWORD
4.2.1 Determining Hidden unit
4.2.2 Second Hidden layer determination:
4.2.3 Determination of Learning Rate
4.2.4 Determination of Momentum
4.2.5 Determination of Activation Function
4.2.6 Determination of Maximum (Max) Update
  4.2.6.1 With Sigmoid
  4.2.6.2 With Tanh
4.2.7 Determination of Stopping Criteria

4.3 NEURAL CONNECTION TRAINING RESULT WITHOUT COMPOUND KEYWORD
4.3.1 Determining Hidden unit
4.3.2 Second layer Hidden layer determination:
4.3.3 Determination of Learning Rate
4.3.4 Determination of Momentum
4.3.5 Determination of Activation Function
4.3.6 Determination of Max Update
4.3.7 Determination of Stopping Criteria
  4.3.7.1 RMS Error Stopping criteria Determination Result Summary with Maximum Max-Update: (32000, 32000, 32000, 32000)
  4.3.7.2 RMS Error Stopping Criteria Determination Result Summary for Default Max Update: (500,500,500,4000)

4.4 OPTIMUM NETWORK SETTING FOR NEURAL CONNECTION

4.5 MLP.JAVA TEST RESULT FOR REV 13 (8:1:1 RATIO DATA)
4.5.1 Determination of Hidden Unit
4.5.2 Determination of Learning Rate
4.5.3 Determination of Learning Momentum
4.5.4 Determination of Activation Function
4.5.5 Determination of Maximum Update
4.5.6 Determination of Stopping Criteria Accuracy
4.5.7 Determination of Stopping Criteria RMS Error

4.6 MLP.JAVA TEST RESULT FOR REV 13 6:2:2 RATIO DATA
4.6.1 Determination of Hidden Unit
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
### APPENDIX B: NEURAL NETWORK TRAINING RESULT WITH COMPOUND KEYWORD

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.1</td>
<td>DETERMINING HIDDEN UNIT</td>
</tr>
<tr>
<td>9.1.1</td>
<td>First layer Hidden layer determination:</td>
</tr>
<tr>
<td>9.1.1.1</td>
<td>First hidden layer result summary</td>
</tr>
<tr>
<td>9.1.2</td>
<td>Second layer Hidden layer determination:</td>
</tr>
<tr>
<td>9.1.2.1</td>
<td>Second hidden layer result summary</td>
</tr>
<tr>
<td>9.2</td>
<td>DETERMINATION OF LEARNING RATE</td>
</tr>
<tr>
<td>9.2.1</td>
<td>Learning Rate Result Summary</td>
</tr>
<tr>
<td>9.3</td>
<td>DETERMINATION OF MOMENTUM</td>
</tr>
<tr>
<td>9.3.1</td>
<td>Learning Momentum Summary</td>
</tr>
<tr>
<td>9.4</td>
<td>DETERMINATION OF ACTIVATION FUNCTION</td>
</tr>
<tr>
<td>9.5</td>
<td>DETERMINATION OF MAXIMUM (MAX) UPDATE</td>
</tr>
<tr>
<td>9.5.1</td>
<td>Activation function is sigmoid:</td>
</tr>
<tr>
<td>9.5.1.1</td>
<td>Maximum Update Summary</td>
</tr>
<tr>
<td>9.5.2</td>
<td>Activation function is Tanh:</td>
</tr>
<tr>
<td>9.5.2.1</td>
<td>Maximum Update Summary</td>
</tr>
<tr>
<td>9.6</td>
<td>DETERMINATION OF STOPPING CRITERIA</td>
</tr>
<tr>
<td>9.6.1</td>
<td>RMS Stopping Criteria Summary</td>
</tr>
</tbody>
</table>

APPENDIX C: NEURAL NETWORK TRAINING RESULT WITHOUT COMPOUND KEYWORD

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
</tr>
</thead>
<tbody>
<tr>
<td>9.1</td>
<td>DETERMINING HIDDEN UNIT</td>
</tr>
<tr>
<td>9.1.1</td>
<td>First layer Hidden layer determination:</td>
</tr>
<tr>
<td>9.1.1.1</td>
<td>First hidden layer result summary</td>
</tr>
<tr>
<td>9.1.2</td>
<td>Second layer Hidden layer determination:</td>
</tr>
<tr>
<td>9.1.2.1</td>
<td>Second hidden layer result summary</td>
</tr>
<tr>
<td>9.2</td>
<td>DETERMINATION OF LEARNING RATE</td>
</tr>
<tr>
<td>9.2.1</td>
<td>Learning Rate Result Summary</td>
</tr>
<tr>
<td>9.3</td>
<td>DETERMINATION OF MOMENTUM</td>
</tr>
<tr>
<td>9.3.1</td>
<td>Learning Momentum Summary</td>
</tr>
<tr>
<td>9.4</td>
<td>DETERMINATION OF ACTIVATION FUNCTION</td>
</tr>
<tr>
<td>9.5</td>
<td>DETERMINATION OF MAXIMUM (MAX) UPDATE</td>
</tr>
<tr>
<td>9.5.1</td>
<td>Activation function is sigmoid:</td>
</tr>
<tr>
<td>9.5.1.1</td>
<td>Maximum Update Summary</td>
</tr>
<tr>
<td>9.5.2</td>
<td>Activation function is Tanh:</td>
</tr>
<tr>
<td>9.5.2.1</td>
<td>Maximum Update Summary</td>
</tr>
<tr>
<td>9.6</td>
<td>DETERMINATION OF STOPPING CRITERIA</td>
</tr>
<tr>
<td>9.6.1</td>
<td>RMS Stopping Criteria Summary</td>
</tr>
</tbody>
</table>
10.1 DETERMINING HIDDEN UNIT
   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: "HYTXTCNT.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 forgetConfig 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 HypertextCounter 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 "HYTXTCNT.JAVA" SOURCE CODE 182
APPENDIX E: MLP.JAVA SOFTWARE

12.1 MLP.JAVA SOURCE CODE

12.2 MLP.JAVA RESULT RAW DATA
   12.2.1 Re12_811.out Raw Result
   12.2.2 Re13_811.out raw result
   12.2.3 Re13_622.out raw result
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>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4.21</td>
<td>Network setting to determine learning momentum</td>
<td>41</td>
</tr>
<tr>
<td>4.22</td>
<td>Learning momentum determination result</td>
<td>41</td>
</tr>
<tr>
<td>4.23</td>
<td>Network setting to determine activation function</td>
<td>42</td>
</tr>
<tr>
<td>4.24</td>
<td>Activation function determination result</td>
<td>42</td>
</tr>
<tr>
<td>4.25</td>
<td>Network setting to determine maximum update</td>
<td>43</td>
</tr>
<tr>
<td>4.26</td>
<td>Maximum update determination result</td>
<td>43</td>
</tr>
<tr>
<td>4.27</td>
<td>Network setting to determine stopping criteria</td>
<td>44</td>
</tr>
<tr>
<td>4.28</td>
<td>Accuracy stopping criteria determination result</td>
<td>44</td>
</tr>
<tr>
<td>4.29</td>
<td>RMS Error Stopping criteria Determination Result Summary with Maximum Max-Update: (32000, 32000, 32000, 32000)</td>
<td>45</td>
</tr>
<tr>
<td>4.30</td>
<td>RMS Error Stopping Criteria Determination Result Summary for Default Max Update: (500,500,500,4000)</td>
<td>45</td>
</tr>
<tr>
<td>4.31</td>
<td>Optimum Network setting for neural connection</td>
<td>46</td>
</tr>
<tr>
<td>4.32</td>
<td>Best result obtained by neural connection.</td>
<td>46</td>
</tr>
<tr>
<td>4.33</td>
<td>Initial Network setting to determine Hidden Unit</td>
<td>47</td>
</tr>
<tr>
<td>4.34</td>
<td>Hidden Unit testing Result Summary</td>
<td>47</td>
</tr>
<tr>
<td>4.35</td>
<td>Network setting Result After determined Hidden Unit</td>
<td>48</td>
</tr>
<tr>
<td>4.36</td>
<td>Best Average and standard Deviation for accuracy and RMS error</td>
<td>48</td>
</tr>
<tr>
<td>4.37</td>
<td>Cross Tabulation Table for Best testing and training result</td>
<td>48</td>
</tr>
<tr>
<td>4.38</td>
<td>Learning Rate testing Result Summary</td>
<td>49</td>
</tr>
<tr>
<td>4.39</td>
<td>Network setting Result after Learning Rate is determined</td>
<td>49</td>
</tr>
<tr>
<td>4.40</td>
<td>Best Average and standard Deviation for accuracy and RMS error</td>
<td>50</td>
</tr>
<tr>
<td>4.41</td>
<td>Cross Tabulation Table for Best testing and training result</td>
<td>50</td>
</tr>
<tr>
<td>4.42</td>
<td>Learning Momentum testing Result Summary</td>
<td>50</td>
</tr>
<tr>
<td>4.43</td>
<td>Network setting Result after determined Learning Momentum</td>
<td>50</td>
</tr>
</tbody>
</table>
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: HypertxtCounter() 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 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


