Application of the Recommendation Architecture Model for Text Mining

Uditha Ratnayake B.Sc. (Eng.) (Hons)

This thesis is presented for the degree of Doctor of Philosophy of Murdoch University

October 2003
Declaration

I declare that this thesis is my own account of my research and contains as its main content work which has not previously been submitted for a degree at any tertiary education institution.

Uditha Ratnayake

October 2003
Acknowledgement

I am very grateful to Prof. Tamás (Tom) Gedeon and Dr. Graham Mann, my principal supervisors, for their constant support and inspiring guidance. Tom’s expertise and encouragement throughout the research process and the development of the thesis were invaluable. Graham’s constructive feedback, enthusiasm and insight made a huge difference to the progress of the thesis. I also thank Dr. Nalin Wickramarachchi, my supervisor in Sri Lanka, for his advice while I worked in Sri Lanka.

My heartfelt gratitude extends to Andrew Coward for his constructive comments on my work in various phases and for many stimulating discussions. His patience in explaining various concepts of the Recommendation Architecture and the help given for programming the prototype are truly appreciated.

My thanks also extend to my colleagues Alex and Kevin for providing useful advice and moral support during my stay at Murdoch. I am indebted to Madu, my husband, for his continuous support, encouragement and patience. Finally, I would like to thank my mother, father and family members for their encouragement and assistance.
List of Publications

The following publications were derived from this research in applying the Recommendation Architecture for the domain of text mining.

Refereed Journal papers


Refereed Conference Papers


Abstract

The Recommendation Architecture (RA) model is a new connectionist approach simulating some aspects of the human brain. Application of the RA to a real world problem is a novel research problem and has not been previously addressed in literature. Research conducted with simulated data has shown much promise for the Recommendation Architecture model’s ability in pattern discovery and pattern recognition. This thesis investigates the application of the RA model for text mining where pattern discovery and recognition play an important role.

The clustering system of the RA model is examined in detail and a formal notation for representing the fundamental components and algorithms is proposed for clarity of understanding. A software simulation of the clustering system of the RA model is built for empirical studies. In the argument that the RA model is applicable for text mining the following aspects of the model are examined. With its pattern recognition ability the clustering system of the RA is adapted for text classification and text organization. As the core of the RA model is concerned with pattern discovery or identification of associative similarities in input, it is also used to discover unsuspected relationships within the content of documents. How the RA model can be applied to the problems of pattern discovery in text and classification of text is addressed demonstrating results from a series of experiments. The difficulties in applying the RA model to real life data are described and several extensions to the RA model for optimal performance are proposed from the insights obtained from experiments. Furthermore, the RA model can be extended to provide user-friendly interpretation of results. This research shows that with the proposed extensions the
RA model can be successfully applied to the problem of text mining to a large extent. Some limitations exist when the RA model is applied to very noisy data, which are also demonstrated here.
# Table of Contents

**Chapter 1  Introduction**

1.1 A Brief Overview of Four AI Models which Simulate the Localized Learning of the Human Brain

1.1.1 The Evolutionary Selection Circuits Model and the Theory of Neural Group Selection

1.1.2 Evolving Connectionist Systems (ECOS)

1.1.3 CAM Brain (CAM – Cellular Automata Machine)

1.2 The Recommendation Architecture

1.3 Application to Text Mining

1.4 Motivation for Research

1.5 Contributions of this Thesis

1.6 Overview of the Thesis

**Chapter 2  The Recommendation Architecture Model**

2.1 Introduction

2.2 Major Characteristics of the Recommendation Architecture

2.3 Functional Overview of the Recommendation Architecture

2.3.1 A Formal Notation for the Functional Components

2.4 The Clustering System

2.4.1 A Formal Notation for the Basic Operations

2.4.2 Factors which Determine Changes in a Specific Device

2.4.3 Growth of the Clustering System

2.4.4 Overview of the Column Output and the Competitive Function

2.5 Summary

**Chapter 3  Information Access and Text Mining**

3.1 Introduction to Information Access Systems
3.2 Advances in Information Retrieval and Filtering .................................................36
  3.2.1 Advances in Retrieval and Filtering Systems Based on Classical Models ......37
  3.2.2 Latent Semantic Indexing .............................................................................39
  3.2.3 Neural Network Models ...............................................................................40
  3.2.4 Retrieval and Filtering for User Requirements .............................................40
3.3 Text Categorization, Clustering and Classification .............................................41
  3.3.1 Text Categorization .......................................................................................42
  3.3.2 Clustering Algorithms ..................................................................................45
  3.3.3 Classification Techniques .............................................................................47
3.4 Text Mining .........................................................................................................48
  3.4.1 Neural Network Models .................................................................................48
  3.4.2 Recommendation Architecture for Text Mining .........................................52
3.5 Conclusion ............................................................................................................54

Chapter 4  Software Simulation of the Recommendation Architecture ............56
  4.1 Introduction .......................................................................................................56
  4.2 Overview of the Prototype ................................................................................57
  4.3 Model Experiment .............................................................................................61
    4.3.1 Formation of the Input Space .................................................................61
    4.3.2 Clustering Run .........................................................................................63
    4.3.3 Results and Discussion .............................................................................64
  4.4 Conclusion .........................................................................................................66

Chapter 5  Modelling the Input Space of the RA for Pattern Discovery and
Classification of Text ...............................................................................................67
  5.1 Introduction .......................................................................................................67
  5.2 Why is Feature Selection Necessary? ...............................................................68
  5.3 Feature Selection Methods ...............................................................................69
Chapter 5  

5.3.1 Types of Feature Selection Methods .................................................................70
5.3.2 The Feature Selection Methods Used for the Experiments .........................71
5.4 Unguided Pattern Discovery ...............................................................................75
5.4.1 Experiment 1 - TREC data with feature selection using the Document Frequency Thresholding method .................................................................76
5.4.2 Experiment 2 - News Group data with feature selection using the Document Frequency Thresholding method .................................................................80
5.4.3 Summary........................................................................................................85
5.5 Guided Pattern Discovery ....................................................................................86
5.5.1 Experiment 1 - TREC data with feature selection using the modified Two-step algorithm .................................................................87
5.5.2 Experiment 2 - Newsgroup data with feature selection using the modified Two-step algorithm .................................................................90
5.5.3 Summary........................................................................................................95
5.6 Conclusion......................................................................................................95

Chapter 6  

Chapter 6  Extending the Clustering System of the RA ...........................................97
6.1 Introduction .....................................................................................................97
6.2 Parameter Selection ......................................................................................99
6.3 Increasing Recognition Accuracy .....................................................................101
6.3.1 Problem of Very Specific Columns .............................................................102
6.3.2 Problem of Very Generic Columns .............................................................105
6.3.3 Demonstration of the Solutions for Very Specific Columns and Very Generic Columns (Extensions -I and II) - Experiment 3a .................................................................109
6.4 Increasing Column Sensitivity – Extension-III ..............................................110
6.4.1 Extension for Feature Intensity Recognition ............................................110
6.4.2 Experiments 3a and 3b – Applying the Extended RA to TREC Data .......112
6.4.3 Experiments 4a and 4b - Applying the Extended RA to Newsgroup Data.....114
6.5 Extending the RA for Text Mining .................................................................117

6.5.1 Automatic Column Labelling – Extension-IV .............................................119

6.5.2 Column Labelling for Experiments 3b and 4b ............................................119

6.5.3 Post–Processing the Output .......................................................................124

6.5.4 Searching for Similar Documents .............................................................126

6.5.5 Performance Evaluation ..........................................................................128

6.6 Summary .......................................................................................................130

Chapter 7 Conclusion ......................................................................................132

7.1 Principal Lessons ..........................................................................................132

7.2 Future Directions .........................................................................................139

Appendix A Additional Experimental Results - Newsgroup Data ...............141

Appendix B Affect of Word Stemming on Document Classification ..............148

REFERENCES .................................................................................................156

BIBLIOGRAPHY ..............................................................................................165
List of Figures

Figure 2-1 Overview of the 4 layers of the Recommendation Architecture.............................19
Figure 2-2 A Device ..................................................................................................................19
Figure 2-3 Layers in one column............................................................................................20
Figure 4-1 Object model of the prototype .............................................................................58
Figure 4-2 Frequency of occurrence of the features in each category .................................62
Figure 4-3 Feature distribution in four input vectors from three different categories ..........62
Figure 4-4 Column creation and stabilization performance ....................................................66
Figure 5-1 Feature density of input vectors in relation to frequency of occurrence (TREC data with Frequency Thresholding method) .......................................................................77
Figure 5-2 Feature density of input vectors in relation to frequency of occurrence (Newsgroup data with Frequency Thresholding method) ........................................................................83
Figure 5-3 Feature density of input vectors in relation to frequency of occurrence (TREC data with modified Two-step algorithm) ..................................................................................87
Figure 5-4 Features that were selected for each topic and their frequency of occurrence (TREC data) ..........................................................................................................................88
Figure 5-5 Feature density of input vectors in relation to frequency of occurrence (Newsgroup data with modified Two-step algorithm) ..................................................................................91
Figure 5-6 Features that were selected for each group and their frequency of occurrence (Newsgroup data) ..........................................................................................................................92
Figure 6-1 Number of columns created for a given number of inputs ..................................103
Figure 6-2 Number of columns created for a given number of inputs with Extension-I .......105
Figure 6-3 Part of the output for column 3 showing the relationships between document vectors and gamma layer device numbers ..................................................................................126
Figure A-1 Document vector sizes in terms of feature density (Experiment NG-1) .............142
Figure A-2 Document vector sizes in terms of feature density (Experiment NG-2) .............144
Figure A-3 Document vector sizes in terms of feature density (Experiment NG-3) .............146
Figure B-1 Frequency of document vectors sizes in terms of feature density..................149

Figure B-2 Features that were selected for each group and their frequency of occurrence for
the training set...........................................................................................................150

Figure B-3 Features that were selected for each group and their frequency of occurrence for
the test set ..................................................................................................................151
# List of Tables

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>4-1</td>
<td>Precision and Recall for each column by the major category identified</td>
<td>65</td>
</tr>
<tr>
<td>4-2</td>
<td>Column sizes in regular section devices</td>
<td>65</td>
</tr>
<tr>
<td>5-1</td>
<td>Total number of documents acknowledged from each column</td>
<td>78</td>
</tr>
<tr>
<td>5-2</td>
<td>Some frequent words in the documents accepted by each column and the TREC</td>
<td>79</td>
</tr>
<tr>
<td></td>
<td>topics that can correspond to the columns</td>
<td></td>
</tr>
<tr>
<td>5-3</td>
<td>Columns that respond to a set of document vectors from different categories</td>
<td>84</td>
</tr>
<tr>
<td>5-4</td>
<td>Precision and Recall for each column by the major document category identified</td>
<td>90</td>
</tr>
<tr>
<td></td>
<td>(Experiment 1-TREC data)</td>
<td></td>
</tr>
<tr>
<td>5-5</td>
<td>Precision and Recall for each column by the major document category identified</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>(Experiment 2-Newsgroup data)</td>
<td></td>
</tr>
<tr>
<td>6-1</td>
<td>Topic and the percentage of documents in each topic that responded to column 3</td>
<td>107</td>
</tr>
<tr>
<td>6-2</td>
<td>Precision and Recall for each column by the major document category identified</td>
<td>107</td>
</tr>
<tr>
<td></td>
<td>(Experiments 1 and 2)</td>
<td></td>
</tr>
<tr>
<td>6-3</td>
<td>Precision and Recall for each column by the major document category identified</td>
<td>110</td>
</tr>
<tr>
<td></td>
<td>(Experiment 3a)</td>
<td></td>
</tr>
<tr>
<td>6-4</td>
<td>Precision and Recall for each column by the major document category identified</td>
<td>113</td>
</tr>
<tr>
<td></td>
<td>(Experiment 3a)</td>
<td></td>
</tr>
<tr>
<td>6-5</td>
<td>Precision and Recall of each column by the major document category identified</td>
<td>113</td>
</tr>
<tr>
<td></td>
<td>(Experiment 3b)</td>
<td></td>
</tr>
<tr>
<td>6-6</td>
<td>Precision and Recall of each column by the major document category identified</td>
<td>116</td>
</tr>
<tr>
<td></td>
<td>(Experiment 4a)</td>
<td></td>
</tr>
<tr>
<td>6-7</td>
<td>Precision and Recall of each column by the major document category identified</td>
<td>116</td>
</tr>
<tr>
<td></td>
<td>(Experiment 4b)</td>
<td></td>
</tr>
<tr>
<td>6-8</td>
<td>TREC topic labels for the major group discovered by each column and the labels</td>
<td>120</td>
</tr>
<tr>
<td></td>
<td>assigned to the columns by the extended RA system.</td>
<td></td>
</tr>
</tbody>
</table>
Table 6-9 Frequently occurring word pairs ............................................................................122
Table 6-10 Newsgroup names assigned for the major group discovered by each column by the
labels assigned to the columns by the extended RA system ...............................................123
Table 6-11 Column-wise breakdown of document groups for columns 3 and 7 .....................125
Table 6-12 A set of the document vectors by the columns they respond to .........................127
Table A-1 Some frequent words in the documents accepted by each column .......................143
Table A-2 Precision and Recall by the major category/categories acknowledged by each
column .....................................................................................................................................145
Table B-1 Precision and Recall for each column by the major document category identified
(Experiment Stem1)..............................................................................................................152
Table B-2 Precision and Recall for each column by the major document category identified
(Experiment Stem2)..............................................................................................................153
Table B-3 Precision and Recall for each column by the major document category identified
(Experiment Stem3)..............................................................................................................153
Table B-4 Average precision and average recall for six columns ........................................154