Document Clustering using Hierarchical Unsupervised Neural Networks

A thesis Submitted to the
Department of Computer Science and Automatic Control
In Partial fulfillment of the requirements
For the degree of
Doctor of Philosophy

By
Eng. Mahmoud Farouk Hussin Mohamed

THE SUPERVISORS

Prof. Dr. Magdy Hussein Nagi
Prof. Dr. Mohamed Kamel
Prof. Dr. Soheir Ahmed Bassiouny

Alexandria University - 2004
TO MY FAMILY
C.V.

Name : Eng. Mahmoud Farouk Hussin Mohamed
Date of Birth : January 7\textsuperscript{th}, 1971
Place of Birth : Alexandria, Egypt.

Home Address :
4 Abd Elwahab Youssif Street,
Kabbary, Alexandria,
Egypt.

Educational Record :


- B.SC. : 1993, Department of Computer Science and Automatic Control, Faculty of Engineering, Alexandria University.

- M.SC. : 1998, Department of Computer Science and Automatic Control, Faculty of Engineering, Alexandria University.
THE SUPERVISORS

Prof. Dr. Magdy Hussein Nagi
Professor :
Department of Computer and Systems Engineering,
Faculty of Engineering,
Alexandria University.

Prof. Dr. Mohamed Kamel
Professor :
Department of Electrical and Computer Engineering,
University of Waterloo.

Prof. Dr. Soheir Ahmed Bassiouny
Professor :
Department of Computer and Systems Engineering,
Faculty of Engineering,
Alexandria University.
ABSTRACT

Recently, there has been a considerable increase in the availability of full-text document collections in electronic form. This has created a need for tools and techniques that assist users in organizing these collections. Specifically, there is a great interest to provide a solution for information categorization. This is especially true for web-based documents. Among the main methods for categorization is document clustering. Document clustering attempts to organize objects into groups, such that objects within a group are more similar to each compared to objects belonging to different groups. Generally, any clustering technique can be divided into two stages. The first stage is the data representation model. The second stage is the clustering algorithm that produces the clusters based on the input data representation.

In this thesis, we propose a promising method for data representation. The approach utilizes phrases rather than individual words as document features for document clustering. Hierarchical phrase grammar was used to extract frequently occurring phrases. These phrases, combined with words form the features representing the documents.

In addition, in this thesis we propose two novel clustering methods based on unsupervised neural network. These methods are the Hierarchical SOMART (HSOMART) and Two Level-SOMART (TL-SOMART). Both of these methods are based on the use of two successful models of unsupervised neural networks, namely, the Self-Organizing Map (SOM) and Adaptive Resonance Theory (ART). These models have both demonstrated promising results in the task of document clustering. These approaches are well suited for textual input, being capable of identifying structure of high dimensionality within a body of natural language text. These method are also capable of successfully handling data that contains noise.

HSOMART method is built up from a hierarchically organized combined SOM and ART neural network with layered architecture where each layer consists of a number of independent SOMs or ARTs. The key idea of the HSOMART is based on, combining the fast learning capability of SOM to generate compact clusters with the accuracy of the clusters produced by ART.

On the other hand, in case of the TL-SOMART, the SOM is used as a dimension reduction method in the first stage. This is achieved by mapping a high-dimensional data space based on
words or phrases into low-dimensional space based on clusters produced by multiple SOM. The ART in the second stage is used, similar to the HSOMART, to produce the final clusters using a reduced vector space.

The experimental results using the REUTERS corpus, are presented. Results show significant improvement of the suggested data representation and clustering methods evaluated by the entropy as well as the F-measure. It also show that clustering using the phrase based features combined with words achieved a better quality than clustering using words only, and demonstrate an improvement in the clustering performance using HSOMART and TL-SOMART in both quality and time execution.
ACKNOWLEDGMENT

“Thanks to god, who helped me in achieving this work”

I wish to express my deepest gratitude to Prof. Dr. Magdy Nagi, Department of Computer and Systems Engineering, Faculty of Engineering, Alexandria University, for his kind supervision, continuous help, encouragement and valuable instructions during the period of this work.

No words can express my gratitude to Prof. Dr. Mohamed Kamel, Department of Electrical and Computer Engineering, University of Waterloo, for his precise and valuable supervision and suggestion, continuous support, constructive criticism, and unlimited help.

I would like also to offer my faithful thanks and appreciation to Prof. Dr. Soheir Bassiouny, Department of Computer and Systems Engineering, Faculty of Engineering, Alexandria University, for his encouragement, precise supervision and for his continuous and unlimited support. Thanks are also due to my great organization, Arab Academy for Science and Technology, and special thanks to the president of the Academy, his excellency, Dr. Gamal el-Din Moukhtar.

Special thanks are due to my parents, my wife, and my two children for their love, patience, and encouragement. They have brought so much happiness and goodness into my life.

Finally, my warm appreciation to the staff of the Department of Computer and Systems Engineering, Alexandria University and all members of Pattern Analysis and Machine Intelligence (PAMI) group, University of Waterloo.
CONTENTS

C.V. iii
THE SUPERVISORS iv
ABSTRACT v
ACKNOWLEDGEMENT vii
CONTENTS viii
LIST OF PUBLICATIONS xii
LIST OF FIGURES xiii

CHAPTER I

INTRODUCTION 1

1.1 Problem statement 1
1.2 Challenges of document clustering 2
  1.2.1 Data representation 3
  1.2.2 Dimensionality reduction 3
  1.2.3 Clustering Methods 4
1.3 The Proposed Neural Network based document clustering methods 5
  1.3.1 Hierarchical SOMART document clustering (HSOMART) 6
  1.3.2 Two-level SOMART document clustering (TL-SOMART) 6
1.4 Road-map to the thesis 6

CHAPTER II

A REVIEW OF DOCUMENT CLUSTERING PROCESS 9

2.1 Introduction 9
2.2 Document Data Representation 10
  2.2.1 Vector Space Representation 10
    2.2.1.1 What is meant by “term” 10
    2.2.1.2 Weighting Schemes 12
    2.2.1.3 Dimensionality Reduction 14
  2.2.2 N-gram Representation 20
  2.2.3 Suffix Tree Representation 20
  2.2.4 Graph Representation 20
2.3 Document Clustering Methods 21
  2.3.1 Partitioning Clustering 21
2.3.2 Hierarchical Clustering 22
2.3.3 Neural Network based Clustering 24
2.4 Cluster Quality Measures 25
2.5 Requirements for Document Clustering Algorithms 26
   2.5.1. combination of algorithms 26
   2.5.2. Extraction of Informative Features 27
   2.5.3. Reduction of the Dimension of a Document Space 27
   2.5.4 Noise Tolerance 28
   2.5.5 Presentation 28

CHAPTER III
LITERATURE REVIEW 30
3.1 Introduction 30
3.2 Document Data Representation 30
3.3 Document Clustering methods based on unsupervised Neural Network 31
    3.3.1 WEBSOM 34
    3.3.2 Clustering of the Self-Organizing Map 36
    3.3.3 growing hierarchical self-organizing map (GHSOM) 37

CHAPTER IV
NEURAL NETWORK MODELS FOR DOCUMENT CLUSTERING 41
4.1 Introduction 41
4.2 An Overview of ANN 41
   4.2.1. What is a neural network 41
   4.2.2 Network Architecture 41
   4.2.3 Learning 43
      2.2.3.1 Learning Paradigm 44
         2.2.3.1.1 Supervised Learning 44
         2.2.3.1.2 Unsupervised Learning 44
         2.2.3.1.3 Reinforcement Learning 45
      2.2.3.2 Learning Theory 45
   2.2.4 Summary 45
4.3. Self-Organizing Map (SOM) 46
CHAPTER V
PROPOSED NEURAL NETWORK BASED DOCUMENT CLUSTERING

CHAPTER VI
EXPERIMENTAL RESULTS AND ANALYSIS

CHAPTER VII
CONCLUSIONS AND FUTURE RESEARCH

7.1 Conclusions

7.2. Future Research

REFERENCES
## LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>SOM</td>
<td>Self-Organizing Map</td>
</tr>
<tr>
<td>ART</td>
<td>Adaptive Resonance Theory</td>
</tr>
<tr>
<td>HSOMART</td>
<td>Hierarchical SOMART</td>
</tr>
<tr>
<td>TL-SOMART</td>
<td>Two-Level SOMART</td>
</tr>
<tr>
<td>HSOM</td>
<td>Hierarchical SOM</td>
</tr>
<tr>
<td>TL-SOM</td>
<td>Two-Level SOM</td>
</tr>
<tr>
<td>LSI</td>
<td>Latent Semantic Indexing</td>
</tr>
<tr>
<td>PCA</td>
<td>Principle Component Analysis</td>
</tr>
<tr>
<td>MDS</td>
<td>Multi-Dimensional Scaling</td>
</tr>
<tr>
<td>IG</td>
<td>Information Gain</td>
</tr>
<tr>
<td>MI</td>
<td>Mutual Information</td>
</tr>
<tr>
<td>CHI</td>
<td>$X^2$ statistic</td>
</tr>
<tr>
<td>DIG</td>
<td>Document Index Graph</td>
</tr>
<tr>
<td>ANN</td>
<td>Artificial Neural Network</td>
</tr>
<tr>
<td>GHSOM</td>
<td>Growing Hierarchical SOM</td>
</tr>
<tr>
<td>HART</td>
<td>Hierarchical ART</td>
</tr>
<tr>
<td>PART</td>
<td>Projective ART</td>
</tr>
</tbody>
</table>
LIST OF FIGURES

1.1 Document Clustering Task 3
2.1 Dimensionality Reduction Step 16
3.1 Schematic representation of a basic neural network 32
3.2 A taxonomy of feedforward and recurrent/feedback network architectures 33
3.3 The ART neural network 40
3.4 The basic WEBSOM architecture, a) The word category map first learns to represent relations of words based on their averaged contexts. This map is used to form a word histogram of the text to be analyzed. B) The histogram, a “fingerprint” of the document, is then used as input to the second SOM, the document map 45
3.5 First abstraction level is obtained by creating a set of prototype vectors using, e.g., the SOM. Clustering of the SOM creates the second abstraction level. 47
3.6 GHSOM architecture: the GHSOM evolves to a structure of SOMs reflecting the hierarchical structure of the input data. 49
3.7 Architecture of a hierarchical SOM. 50
4.1 The system design 53
4.2 Class entropy and F-measure for 1000 document set, and clustered by SOM technique for different values of phrase importance factor 57
4.3 Class entropy and F-measure for 1000 document set, and clustered by SOM technique for different number of features in both word and phrase representation 59
4.4 Class entropy and F-measure for 1000 document set, and clustered by SOM technique for different number of features in both word and phrase combined with word representation 59
4.5 Architecture of a hierarchical SOMART 61
4.6 Number of epochs needed in both SOM and ART, and the improvement of ART over SOM in the computed time for different number of documents 63
4.7 Logarithmic Execution Time and Actual Execution Time Comparison 65
Between SOM and ART Techniques for different number of documents in the collection set from 0 - 1000

4.8 Execution Time Comparison Between SOM and ART Techniques for number of documents in the collection set, first from 0 to 200, and from 200 to 400

4.9 Execution Time Comparison Between SOM and ART Techniques for number of documents in the collection set, first from 400 to 600, and from 600 to 1000

4.10 Class entropy and F-measure for 1000 document set, and clustered by SOM and ART techniques for different number of clusters

4.11 Two-level SOMART Document Clustering

5.1 HSOM layer architecture used to cluster 1000 documents

5.2 HSOM layer architecture used to cluster 3000 documents

5.3 HSOMART layer architecture used to cluster 1000 documents

5.4 HSOMART layer architecture used to cluster 3000 documents

5.5 Class entropy and F-measure for 1000 document set, using binary representation and clustered by SOM technique

5.6 Class entropy and F-measure for 1000 document set, using term frequency representation and clustered by SOM technique

5.7 Class entropy and F-measure for 1000 document set, using tf/idf representation and clustered by SOM technique

5.8 Class entropy and F-measure for 1000 document set, using binary representation and clustered by HSOM technique

5.9 Class entropy and F-measure for 1000 document set, using term frequency representation and clustered by HSOM technique

5.10 Class entropy and F-measure for 1000 document set, using tf/idf representation and clustered by HSOM technique

5.11 Class entropy and F-measure improvement using HSOM compared with SOM in both phrase and word data representation

5.12 Class entropy and F-measure for 1000 document set, using binary representation and clustered by SOM, HSOM, and HSOMART techniques

5.13 Class entropy and F-measure for 1000 document set, using term representation
frequency representation and clustered by SOM, HSOM, and HSOMART techniques

5.14 Class entropy and F-measure for 1000 document set, using tf/idf representation and clustered by SOM, HSOM, and HSOMART techniques

5.15 Class entropy and F-measure improvement using HSOMART and HSOM compared with SOM

5.16 Execution time for 1000 document set, using binary representation and clustered by HSOM, and HSOMART techniques

5.17 Execution time for 1000 document set, using TF representation and clustered by HSOM, and HSOMART techniques

5.18 Execution time for 1000 document set, using TF representation and clustered by HSOM, and HSOMART techniques

5.19 Class entropy and F-measure for 1000 document set using binary weighting scheme, and clustered by SOM, HSOM, and HSOMART techniques using both word (W) or phrase (P) data representation

5.20 . Class entropy and F-measure for 1000 document set using TF weighting scheme, and clustered by SOM, HSOM, and HSOMART techniques using both word (W) or phrase (P) data representation

5.21 Class entropy and F-measure for 1000 document set using TF-IDF weighting scheme, and clustered by SOM, HSOM, and HSOMART techniques using both word (W) or phrase (P) data representation

5.22 Class entropy and F-measure for 3000 document set using binary weighting scheme, and clustered by SOM, HSOM, and HSOMART techniques using both word (W) or phrase (P) data representation

5.23 Class entropy and F-measure for 3000 document set using TF weighting scheme, and clustered by SOM, HSOM, and HSOMART techniques using both word (W) or phrase (P) data representation

5.24 Class entropy and F-measure for 3000 document set using TF-IDF weighting scheme, and clustered by SOM, HSOM, and HSOMART techniques using both word (W) or phrase (P) data representation

5.25 Class entropy and F-measure for 1000 document set, using binary representation and clustered by SOM, TSOM, and TSOMART
5.26 Class entropy and F-measure for 1000 document set, using TF representation and clustered by SOM, TSOM, and TSOMART techniques

5.27 Class entropy and F-measure for 1000 document set, using TFIDF representation and clustered by SOM, TSOM, and TSOMART techniques
LIST OF TABLES

2.1 A vector representation of document 1 and document 2. 12
2.2 A vector representation of stemmed document 1 and stemmed document 2. 13
2.3 A binary vector representation of sample document 1 and sample document 2 14
2.4 Three different distance function in agglomerative hierarchical clustering 25
4.1 Resulting grammar from the phrase “white house chief staff howard baker” using the mutual information association measure 56
4.2 Class entropy for 1000 document, and clustered by SOM using 4X4 output map for different values of phrase importance factor 58
4.3 F_measure for 1000 document, and clustered by SOM using 4X4 output map for different values of phrase importance factor 58
4.4 Class entropy and F_measure for 1000 document, and clustered by SOM using 6X6 output map for different number of features in both word and phrase combined with word representation 59
4.5 Class entropy and F_measure for 1000 document, and clustered by SOM using 8X8 output map for different number of features in both word and phrase combined with word representation 60
4.6 Comparison of number of epochs, and computed time of both SOM and ART for different number of documents 63
4.7 Class entropy and F_measure for 1000 document set, clustered by SOM and ART algorithms with different number of clusters 66
5.1 Summary description on data sets 71
5.2 Comparison of the Class entropy and F-measure using binary representation and clustered by SOM technique 75
5.3 Comparison of the Class entropy and F-measure using term frequency representation and clustered by SOM technique 75
5.4 Comparison of the Class entropy and F-measure using tf/idf representation and clustered by SOM technique 76
5.5 Comparison of the Class entropy and F-measure using binary representation 76
representation and clustered by HSOM technique

5.6 Comparison of the Class entropy and F-measure using term frequency representation and clustered by HSOM technique

5.7 Comparison of the Class entropy and F-measure using tf/idf representation and clustered by HSOM technique

5.8 Comparison of the Class entropy using binary representation and clustered by SOM, HSOM, and HSOMART techniques

5.9 Comparison of the F-measure using binary representation and clustered by SOM, HSOM, and HSOMART techniques

5.10 Comparison of the Class entropy using tf representation and clustered by SOM, HSOM, and HSOMART techniques

5.11 Comparison of the F-measure using tf representation and clustered by HSOM, HSOM, and HSOMART techniques

5.12 Comparison of the Class entropy using tf/idf representation and clustered by SOM, HSOM, and HSOMART techniques

5.13 Comparison of the F-measure using tf/idf representation and clustered by SOM, HSOM, and HSOMART techniques

5.14 Comparison of the execution time using binary representation and clustered by HSOM, and HSOMART techniques

5.15 Comparison of the execution time using TF representation and clustered by HSOM, and HSOMART techniques

5.16 Comparison of the execution time using TF/IDF representation and clustered by HSOM, and HSOMART techniques

5.17 Comparison of the Class entropy using binary weighting scheme and clustered by SOM, HSOM, and HSOMART techniques using both word and phrase data representations, for 1000 document set

5.18 Comparison of the F-measure using binary weighting scheme and clustered by SOM, HSOM, and HSOMART techniques using both word and phrase data representations, for 1000 document set

5.19 Comparison of the Class entropy using TF weighting scheme and clustered by SOM, HSOM, and HSOMART techniques using both word and phrase data representations, for 1000 document set

5.20 Comparison of the F-measure using TF weighting scheme and
clustered by SOM, HSOM, and HSOMART techniques using both word and phrase data representations, for 1000 document set

5.21 Comparison of the Class entropy using TF/IDF weighting scheme and clustered by SOM, HSOM, and HSOMART techniques using both word and phrase data representations, for 1000 document set

5.22 Comparison of the F_measure using TF/IDF weighting scheme and clustered by SOM, HSOM, and HSOMART techniques using both word and phrase data representations, for 1000 document set

5.23 Comparison of the Class entropy using binary weighting scheme and clustered by SOM, HSOM, and HSOMART techniques using both word and phrase data representations, for 3000 document set

5.24 Comparison of the F_measure using binary weighting scheme and clustered by SOM, HSOM, and HSOMART techniques using both word and phrase data representations, for 3000 document set

5.25 Comparison of the Class entropy using TF weighting scheme and clustered by SOM, HSOM, and HSOMART techniques using both word and phrase data representations, for 3000 document set

5.26 Comparison of the F_measure using TF weighting scheme and clustered by SOM, HSOM, and HSOMART techniques using both word and phrase data representations, for 1000 document set

5.27 Comparison of the Class entropy using TF/IDF weighting scheme and clustered by SOM, HSOM, and HSOMART techniques using both word and phrase data representations, for 1000 document set

5.28 Comparison of the F_measure using TF/IDF weighting scheme and clustered by SOM, HSOM, and HSOMART techniques using both word and phrase data representations, for 1000 document set

5.29 Comparison of the Class entropy using binary representation and clustered by SOM, TL-SOM, and TL-SOMART techniques

5.30 Comparison of the F_measure using binary representation and clustered by SOM, TL-SOM, and TL-SOMART techniques

5.31 Comparison of the Class entropy using TF representation and clustered by SOM, TL-SOM, and TL-SOMART techniques

5.32 Comparison of the F_measure using TF representation and clustered by
SOM, TL-SOM, and TL-SOMART techniques

5.33 Comparison of the Class entropy using TF/IDF representation and clustered by SOM, TL-SOM, and TL-SOMART techniques

5.34 Comparison of the F_measure using TF/IDF representation and clustered by SOM, TL-SOM, and TL-SOMART techniques
# LIST OF ALGORITHMS

2.1 Single-pass partitioning  
2.2 Iterative partitioning  
2.3 Hierarchical agglomerative clustering  
3.1 HSOM document clustering  
4.1 Grammar Generation and phrase extraction  
4.2 HSOMART document clustering  
4.3 Two-level SOMART Document Clustering