Verifying the Proximity Hypothesis for Self-organizing Maps

Chienting Lin, Hsinchun Chen, and Jay F. Nunamaker
Department of Management Information Systems, McClelland Hall 430
University of Arizona, Tucson, Arizona 85721, {linc, hchen, nunamaker}@BPA.Arizona.EDU

Abstract

The Kohonen Self-Organizing Map (SOM) is an unsupervised learning technique for summarizing high-dimensional data so that similar inputs are, in general, mapped close to each other. When applied to textual data, SOM has been shown to be able to group together related concepts in a data collection. This article presents research in which we sought to validate this property of SOM, called the Proximity Hypothesis, through a user evaluation study. Built upon our previous research in automatic concept generation and classification, we demonstrated that the Kohonen SOM was able to perform concept clustering effectively, based on its concept precision and recall scores judged by human experts. We believe this research has established the Kohonen SOM algorithm as an intuitively appealing and promising neural network based textual classification technique for addressing part of the long-standing "information overload" problem.

1 Introduction

With the sudden emergence and proliferation of the Internet services, the information overload problem has become more pressing than ever. Researchers in the field of information and knowledge management have started to seek assistance from the information retrieval and artificial intelligence communities, who have much to offer concerning advanced information indexing, searching, and classification techniques. Previous research has strongly suggested the Kohonen SOM algorithm [7] as an ideal candidate for classifying textual documents. The Kohonen SOM [1] provides an intuitively appealing organization of input data. Documents are classified according to their content and conceptual regions are formed and named on a two-dimensional grid. Kohonen SOM output also exhibits two distinctive characteristics that are appealing for cognitive and visual reasons: first, the related topics/regions are clustered closely (the Proximity Hypothesis); second, larger regions represent more important issues in the data collections (the Size Hypothesis). The graphical display of SOM maps prompted us to experiment with these features. In this study, we intended to validate the Proximity Hypothesis through a user study. The Proximity Hypothesis, if verified, has significant implications for designing an effective and graphically appealing human-computer interface for textual analysis.

In the next section, we summarize the techniques (statistical or neural network based) for document classification. Section 3 presents a framework we developed for applying the Kohonen SOM algorithm in document and concept clustering. Section 4 describes our hypothesis, experimental procedures and results. Conclusions and future directions are summarized in section 5.

2 Document Clustering Techniques

Classification of textual documents requires grouping similar concepts/terms by category or topic. Two approaches to cluster analysis exist: the statistical approach and the neural network approach. In this section, we provide only a brief summary of the conventional statistical approach and a more detailed review of the newer parallel, neural network approach because our techniques are based on a neural network algorithm.

In the serial, statistical approach, automatic document classification involves determining a document representation structure and method for determining similarities between documents. The hierarchical clustering of documents can be done divisively or agglomeratively [2]. Divisive clustering breaks one complete cluster into smaller pieces. In agglomerative clustering, similarities between individual documents are used as a starting point and a gluing operation is carried out to form larger groups. Stepp [3] described conceptual

0-7695-0001-3/99 $10.00 (c) 1999 IEEE

clustering as the new frontier in artificial intelligence. Algorithms for clustering involve co-occurrence of feature values, discovering conjunctive features among the attributes rather than variations in the value taken by a single attribute, and clumping concepts based upon most commonly occurring relations in the data. Using these techniques, classes of similar objects are basically found by doing pair-wise comparisons among all the data elements. These clustering algorithms are serial in nature in that pair-wise comparisons are made one at a time and the classification structure is created in a serial order.

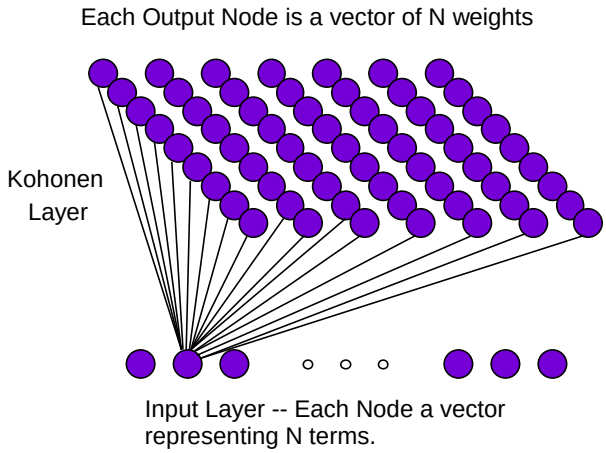

Figure 1: Kohonen SOM topology

The neural network approach, on the other hand, addresses clustering and classification problems by means of a connectionist approach. Algorithms based upon neural networks are parallel in that multiple connections among the nodes allow for independent, parallel comparisons. Neural network techniques can be classified as supervised and unsupervised. In supervised learning, a set of training examples is presented, one by one, to the network. The network then calculates outputs based on its current input. The resulting output is then compared with a desired output for that particular input example. The network weights are then adjusted to reduce the error. In unsupervised learning, network models are first presented with an input vector from the set of possible network inputs. The network learning rule adjusts the weights so that input examples are grouped into classes based on their statistical properties. Doszkocs, Reggia and Lin [4] provide an excellent overview of connectionalist models in information retrieval including artificial networks, spreading activation models, associative networks and parallel distributed processing. Chen [5] provides an up-to-date review of various machine learning techniques, neural networks, and genetic algorithms for intelligent information retrieval applications.

Among unsupervised learning methods, the Kohonen SOM has been strongly suggested as an ideal candidate for clustering of textual documents. Kohonen based his neural network on the associative neural properties of the brain. The network contains two layers of nodes: an input layer and a mapping layer in the shape of a two-dimensional grid. The output layer acts as a distribution layer. The number of nodes in the input layer is equal to the number of features associated with the input. Each node of the mapping layer also has the same number of features as there are input nodes. Thus, the input layer and each node of the mapping layer can be represented as a vector which contains the number of features of the input. The network is fully connected in that every mapping node is connected to every input node. The topology of the Kohonen SOM network is shown in Figure 1.

Several recent studies adapted the SOM approach to textual analysis and classification. Ritter and Kohonen [8] applied the Kohonen SOM to textual analysis in an attempt to detect the logical similarity between words from the statistics of their contexts. Miikkulainen [9] developed DISCERN (Distributed Script processing and Episodic memoRy Network) as a prototype of a subsymbolic natural language processing system based on the Kohonen SOM. Lin, Soergel and Marchionini [10] used the Kohonen SOM for classifying documents for information retrieval. The documents were classified according to their content and conceptual regions were formed and named on a two-dimensional grid. Lin's work first demonstrated the feasibility of using the Kohonen algorithm for classification of textual documents. Orwig and Chen adopted a scalable SOM algorithm to classify electronic brainstorming outputs and Internet homepages [6][7]. The scalability was achieved using the Scalable SOM (SSOM) technique developed by Roussinov and Chen [11]. The SSOM data structure and algorithm took advantage of the sparsity of coordinates in the document input vectors and reduced the SOM computational complexity by several order of magnitude, thus making large-scale textual categorization tasks a possibility. Kaski,

Honkela, Lagus and Kohonen reported WEBSOM [12], an SOM-based text classifier for clustering postings to the Usenet newsgroups. The basic WEBSOM architecture consists of two hierarchically interrelated SOMs. A *word category map* [8] is created first to describe relations of words based on their averaged short contexts. In the second stage, the text of a document is mapped onto the word category map previously created and a histogram of the hits on it is formed. The *document map* is then obtained using the histograms as the fingerprints of the textual documents. WEBSOM also provides automatic labeling as exhibited in Chen's work, but the lack of distinct region boundaries could limit its use.

3 A Framework for Document Classification

The Kohonen SOM algorithm for classifying textual documents requires outputs from automatic indexing or noun-phrase extraction process, which contain index terms for the documents and a list of terms in decreasing order of frequency for the entire collection. Based on the indexing terms identified, each document then is represented by a term vector of 1 or 0. The number of 1s in each document is equal to the number of terms in the document and each vector position corresponds with one unique term.

We chose a 20 by 10 grid map for displaying SOM outputs, based upon what would fit on an output screen. We used a hexagonal neighborhood area which considers six surrounding nodes to be a node's immediate neighborhood. Finally, we used the bubble adjustment method, which is an adjustment of the weights of neighboring nodes based upon the decreasing gain term. In the initial training phase, we used a gain term adjustment of 0.05, and a neighborhood size of 10. In the fine-tuning phase we used a small gain term adjustment of 0.01, and a smaller neighborhood size of 3.

After the training and tuning phases, the SOM

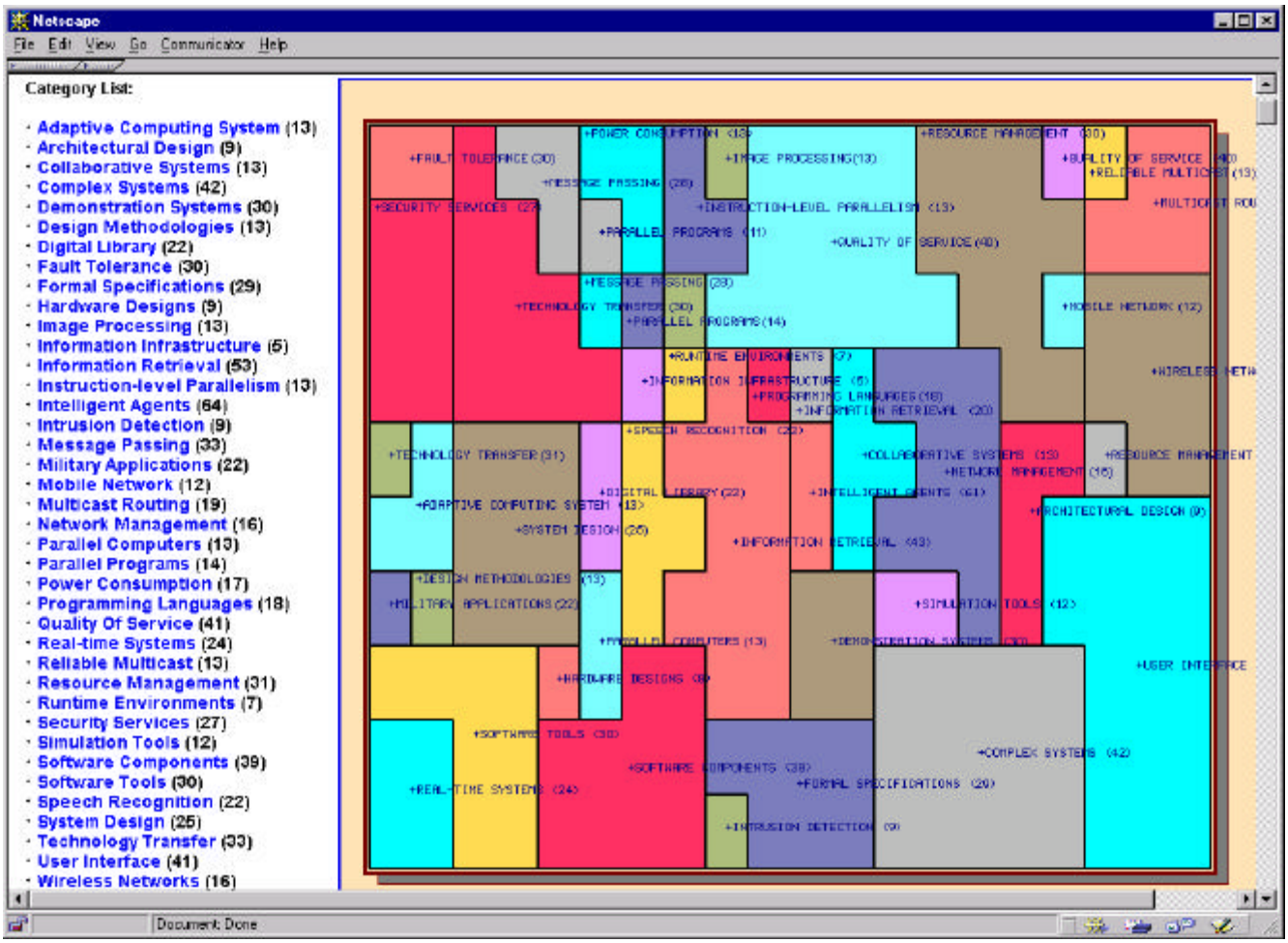

Figure 2: The SOM Map for the ITO collection

visualization consisted of running the same input file against the trained map and reporting the map grid location that is the closest in Euclidean distance to each input. Each document (vector) and each term (represented as a unit vector) were thus mapped to a node and also to a region (of the same nodes) on the map. The nodes were then labeled so nodes with the same labels could form regions. The SOM algorithm we adopted, as opposed to the original Kohonen SOM, is summarized below:

(1) Initialize input nodes, output nodes, and connection weights:

Represent each document (or image) as an input vector of N keywords (or image features) and create a two-dimensional map (grid) of M output nodes (e.g., a 20-by-10 map of 200 nodes). Initialize weights from N input nodes to M output nodes to small random values.

(2) Present each document or image in order:

Represent each document by a vector of N features and present to the system.

(3) Compute distances to all nodes:

Compute distance d_j between the input and each output node j using

$$d_j = \sum_{i=0}^{N-1} (x_i(t) - w_{ij}(t))^2$$

where $x_i(t)$ is the input to node i at time t and $w_{ij}(t)$ is the weight from input node i to output node j at time period t.

(4) Select winning node j* and update weights to node j* and neighbors:

Select winning node j* as that output node with minimum d_j. Update weights for node j* and its neighbors to reduce their distances (between input nodes and output nodes). (See [16] for the algorithmic detail of neighborhood adjustment.)

(5) Label regions in map:

After the network is trained through repeated presentation of all inputs, submit unit input vectors of single terms to the trained network and assign the winning node the name of input feature. Neighboring nodes which contain the same feature then form a concept or topic region. The resulting map thus represents regions of important terms or image patterns (the more important a concept, the larger a region) and the assignment of similar documents or images to each region.

1	29	USER INTERFACE
2	29	QUALITY OF SERVICE
3	28	TECHNOLOGY TRANSFER
4	25	FAULT TOLERANCE
5	25	RESOURCE MANAGEMENT
6	22	MILITARY APPLICATIONS
7	22	MESSAGE PASSING
8	21	COMPLEX SYSTEMS
9	20	REAL-TIME SYSTEMS
10	20	SOFTWARE TOOLS
11	19	SOFTWARE COMPONENTS
12	17	SECURITY SERVICES
13	17	SYSTEM DESIGN
14	16	REAL TIME
15	16	PROGRAMMING LANGUAGES
16	16	FILE SYSTEM
17	15	SOFTWARE ARCHITECTURE
18	15	REAL-TIME APPLICATIONS
19	14	ANALYSIS TOOLS
20	14	SECURITY MECHANISMS

Table 1: Top 20 most frequently occurring terms in the ITO collection

(6) Apply the above steps recursively for large regions:

For each map region which contains more than k (e.g., 100) documents or images, conduct a recursive procedure to generate another self-organizing map until each region contains no more than k documents or images.

4 User Evaluation

In order to validate the proximity hypothesis for SOM maps, we recently designed and conducted a user study aimed at answering the following questions: Can the SOM really cluster related topics together? Can the results be systematically validated using human beings as judges? Specifically, does the term associations suggested by the SOM match the associations that human subjects expect to see?

4.1 Experimental Hypothesis

To evaluate the term associations produced by the SOM, the SOM maps were compared to maps

generated at random, which provided region associations created by chance without the treatment of the Kohonen SOM algorithm. Concept precision and recall were used as the measurements. The respective null and alternative hypotheses were:

H_0: SOM performs no better than a map generated randomly in terms of precision and recall;

H_1: otherwise (SOM does perform better).

4.2 Test Collections

Two data collections were used in order to compare results across different domains and types of documents. EBS was a set of electronic brainstorming output containing 206 comments. Each comment was one to four lines of textual description regarding the future of groupware. This collection is a good example of a small-sized data collection which focuses on a single topic. ITO was a collection of 586 project summaries for project proposals that have been awarded by the Information Technology Office (ITO) in the Defense Advanced Research Projects Agency (DARPA). These documents contained 3-4 pages of structured textual descriptions about the title, performer, objective, approach, and accomplishments of research projects. While the project summaries tended to be consistent in size, the range of topics within the ITO collection was much greater than within the EBS collection, covering everything from digital libraries to intelligent agents to IP multi-casting. ITO provides a good example of a rich technical domain with a wide range of hardware and software topics.

Figure 3: User Interface for Experiment

4.3 Preprocessing

Automatic indexing and noun-phrasing were applied to the EBS and ITO collections. Respectively: 1,104 concept terms were identified for the EBS collection, while 4,258 terms were identified for ITO. Table 3 shows a list of the top 20 most frequently appearing terms, along with their term frequency, in the ITO collection. For the EBS collection, we used 500 training iterations and 5,000 tuning iterations. The vector size (the number of terms used to form document vectors) was 100. It took 5 minutes to generate the EBS dataset on a mid-sized DEC Alpha 3000/600 server. For the ITO collection, the training and tuning cycles were 1,800 and 4,800, respectively. The vector size was 1,000. It took 20 minutes to generate when computation was done on a DEC 2100 server.

The SOM output for the EBS collection is presented and discussed in [6]. The SOM Map for the ITO collection is shown in Figure 2. The SOM map created contains 45 regions, with 39 unique concepts. An alphabetical list view showing the unique concept regions in a sidebar is provided as an alternative to the two-dimensional grid. For easier map visualization, clicking items from the list will make their corresponding SOM regions blink. The numbers on the map correspond with the number of documents that are classified into a particular concept region. The concept regions can be clicked to view the documents directly. The initial outputs and computational characteristics for the ITO maps are interesting. We observed that many of the larger concept regions appeared to be meaningful and to relate to each other (e.g. DIGITAL LIBRARY and

```
ANALYSIS OF VARIANCE
SOURCE     DF        SS         MS         F         p
FACTOR      1   0.30104    0.30104    39.07    0.000
ERROR      58   0.44690    0.00771
TOTAL      59   0.74794
                                    INDIVIDUAL 95 PCT CI'S FOR MEAN
                                    BASED ON POOLED STDEV
LEVEL      N      MEAN    STDEV  ------+---------+---------+---------+
SOM        30   0.32292  0.08817                          (-----*----)
Random     30   0.18125  0.08738  (----*-----)
                                 ------+---------+---------+---------+
 POOLED STDEV =  0.08778             0.180     0.240     0.300     0.360

TEST OF MU = 0.0000 VS MU N.E. 0.0000
               N      MEAN    STDEV   SE MEAN        T    P VALUE
 SOM-Rand     30    0.1417   0.1303    0.0238     5.95     0.0000

95.0 PERCENT C.I. for MU SOM – Random:        (  0.0930,  0.1903)
```

Table 2: Recall/Precision analysis for the small EBS collection

INFORMATION RETRIEVAL form neighboring regions in the middle of the map, while MOBILE NETWORK and WIRELESS NETWORK are on the top-right region of the map).

4.4 Subjects and Experimental Procedures

The experiment involved 30 human subjects, mainly graduate students from the MIS and ECE departments at the University of Arizona. Subjects chosen possessed prior knowledge and training in collaborative computing and advanced artificial intelligence/software engineering topics so they could evaluate both the EBS and the ITO collections.

We first sampled regions from the SOM maps. For each region sampled, the number of its neighbors and the neighbors' respective region labels were recorded. We then asked the human subjects to select from a list of all region labels the same number of concepts as the SOM had found most relevant to the same concept we sampled. The subjects were asked to perform a total of nine tasks, four for the EBS collection and five for the ITO collection. The user-interface for the experiment was a Java applet, shown in Figure 3, which allowed users to drag and drop terms from the term list to single out concept terms most related to the head concept. Subjects completed the experiment in times ranging from 15 to 30 minutes.

4.5 Experimental Results

The respective performances of the SOM and the random maps were computed using concept recall and precision as measurements using equations in which, X represented terms suggested by either the SOM or the random map, and Y represented terms suggested by the human subjects:

$$\text{Precision} = \frac{(X \cap Y)}{X} \qquad \text{Recall} = \frac{(X \cap Y)}{Y}$$

In our study, since the number of terms suggested by the human subject equaled the number of terms suggested by the SOM maps or the random maps, the concept precision and recall scores were effectively the same.

Tables 2 and 3 contain the results of statistical analysis (both ANOVA and paired *t*-tests) of the comparison of the recall/precision levels of SOM and the random maps as judged by the subjects. For the EBS collection, SOM achieved a respectable

```
ANALYSIS OF VARIANCE
SOURCE     DF        SS         MS         F         p
FACTOR      1   0.61491    0.61491    120.01     0.000
ERROR      58   0.29718    0.00512
TOTAL      59   0.91210
                                    INDIVIDUAL 95 PCT CI'S FOR MEAN
                                    BASED ON POOLED STDEV
LEVEL       N      MEAN     STDEV  -----+---------+---------+---------+-
Som        30   0.26846   0.09030                                  (--*---)
Rand       30   0.06599   0.04576   (--*---)
                                    -----+---------+---------+---------+-
 POOLED STDEV =  0.07158              0.070     0.140     0.210     0.280

TEST OF MU = 0.0000 VS MU N.E. 0.0000
              N      MEAN    STDEV   SE MEAN        T    P VALUE
 Som-Rand    30    0.2025   0.0963    0.0176    11.52     0.0000

95.0 PERCENT C.I. for Som – Random:           (  0.1665,  0.2384)
```

Table 3: Recall/Precision analysis for the large ITO collection

32.29% precision and recall, versus 18.12% for the random map. For the ITO dataset, SOM had 26.85% precision and recall, versus 6.59% for the random map. In both cases, the *p*-values were less than 0.0000. With such a small *p*-value, we have strong evidence that the mean change is greater than zero. We rejected our null hypothesis and concluded that SOM does make a difference in terms of the ability to cluster similar concepts/regions together. In summary, the results from our evaluation were encouraging. In light of the cognitive demand and the cumbersome nature of classifying/clustering textual documents, we believe this research has established the Kohonen SOM as a promising and visually appealing neural network based textual analysis and mining technique.

5 Conclusions and Future Directions

This research aimed at validating the properties of the Kohonen SOM in the domain of textual classification. In evaluation, we compared SOM's clustering of concept terms with a random map topology and used human beings as judges. More work is needed to expand this research to validate the other hypothesis of the SOM, the Size Hypothesis. The size hypothesis suggests a positive correlation between the size of SOM regions and their relative importance in data collection, as judged by SOM users. Partitions of the EBS collections will be best suited for validating such SOM characteristics. Another possible extension involves evaluation of the multi-layered SOM [7], which uses the divide-and-conquer strategy to provide scalable organization for large-scale textual collections.

6 Acknowledgement

This project is supported by the following grants: NSF/DARPA/NASA Digital Library Initiative, IRI-9411318, 1994-1998 (B. Schatz, H. Chen, et. al, "Building the Interspace: Digital Library Infrastructure for a University Engineering Community") and DARPA Information Management Program, N66001-97-C-8535, 1997-2000 (B. Schatz, H. Chen, "The Interspace Prototype: An Analysis Environment for Semantic Interoperability"). We would also like to thank Dr. Olivia Sheng and Dr. Ron Larsen for their comments and insights that have guided us in our research.

References

[1] T. Kohonen, *Self-Organizing Maps*, Springer-Verlag, Berlin, Heidelberg, 1995.

[2] G. Salton, *Automatic text processing*, Addison Wiley Publishing, MA, 1989.

[3] R. Stepp. Concepts in conceptual clustering, in *Proceedings of the Tenth International Joint Conference on Artificial Intelligence*, Milan, Italy, 1987.

[4] T. Doszkocs, J. Reggia, and X. Lin, Connectionist models and information retrieval, *Annual Review of Information Science and Technology*, 25, 209-260.

[5] H. Chen, Machine learning for information retrieval: Neural networks, symbolic learning, and genetic algorithms, *Journal of the American society for Information Science*, 46, 194-216, 1995.

[6] R. Orwig, H. Chen, and J. F. Nunamaker, A Graphical, Self-Organizing Approach to Classifying Electronic Meeting Output, *Journal of the American Society for Information Science*, 48(2): 157-170, 1997.

[7] H. Chen, C. Schuffels, and R. Orwig, Internet Categorization and Search: A Self-Organizing Approach, *Journal of Visual Communications and Image Representation*, Vol. 7. No. 1. March. pp.88-102, 1996.

[8] H. Ritter and T. Kohonen, Self-organizing semantic maps, *Biological Cybernetics*, 61, 241-254, 1989.

[9] R. Miikkulainen, *Sybsymbolic natural language processing: An integrated model of scripts, lexicons, and memory*, the MIT Press, Cambridge, MA, 1993.

[10] X. Lin, D. Soergel, and G. Marchionini, A self-organizing semantic map for information retrieval, in *Proceedings of the Fifteenth Annual International ACM/SIGIR Conference on R&D in Information Retrieval*, Copenhagen, pp. 37-50, 1992.

[11] D. Roussinov and H. Chen, A Scalable Self-organizing Map Algorithm for Textual Classification: A Neural Network Approach to Thesaurus Generation, *Communication Cognition and Artificial Intelligence*, Spring 1998.

[12] S. Kaski, T. Honkela, K. Lagus, and T. Kohonen, Creating an Order in Digital Libraries with Self-Organizing Maps, in *Proceedings of World Congress on Neural Networks*, San Diego, CA., 1996.