12 Self-Organizing Maps of Large Document Collections

Timo Honkela, Krista Lagus and Samuel Kaski

Abstract

All applications presented in the previous chapters applied self-organizing maps to reducing quantitative, numeric data. This chapter shows how textual information can be treated in a similar way and how self-organizing maps can help in more effective retrieval of information than current search engines. The use of WEBSOM is a novel method for organizing collections of text documents into maps, for browsing and exploring links on the World Wide Web, or for organization of electronic messages or files. Timo Honkela and the team at the Neural Network Center at HUT provide several examples of the use of WEBSOM and many more are available on their website (*http://nodulus.hut.fi/websom/*).

12.1 Introduction

A well-organized library provides several catalogues of the material stored: one by author and one by subject. By looking at the subject catalogue a user can gain an idea of what kinds of books the library has on a particular subject. An ordered collection of links in the World Wide Web is a computerized counterpart of a library. Web directories or ready-made categorization of web links are useful for quick access to information by users. They are also useful if the domain is not very well known, or if the user has only a limited idea of the contents of a domain.

Although a library catalogue may provide invaluable help to find a relevant book or article it may nevertheless be time-consuming because a particular topic may not necessarily fit in the predetermined categories. Searching for information has become easier with the availability of electronic full text databases that have the capability for performing complete text searches. Thus, one can look for documents that contain a specific search expression, for example, a string or a combination of strings.

Used as stand-alone tools, keyword searches are obvious limited. Depending on the size and specificity of the database, and of the quality of the chosen keywords, the search may return hardly any documents or an overwhelming number of them.

In this chapter we describe how SOM can be used to combine both an overall organized view of the document collection and the capability of performing

detailed searches. The SOM organizes the documents automatically into maps, where nearby locations usually contain similar documents. The document collection can be explored with the aid of the map view, and content-directed search results provide ideal starting points for exploration of related topics. One may provide keywords or even a whole article for which the system then finds the closest "relatives" on the map. A starting point is always found – there are no null-results or huge outputs.

There are many alternative ways of using the SOM for creating document maps. There are many ways that SOM can be used in information retrieval and textual data mining in general. In this chapter we present only a few approaches including the WEBSOM (Honkela et al., 1996; Kaski et al., 1996; Kohonen et al., 1996; Lagus et al., 1996) [12.01].

12.2 WEBSOM for Document Map Applications

The WEBSOM is a novel method for organizing collections of text documents into maps, and a browsing interface for exploring the maps. The maps are created automatically using the SOM algorithm. With suitable preprocessing, any kinds of text documents can be processed. The name of the system stems from the massive amounts of potentially useful documents that there are available in electronic form in the World Wide Web, including the home pages and Usenet newsgroup articles. The current implementation of the WEBSOM also includes a browsing tool that allows the users to reach the document maps and the organized collection of documents. The browsing tool effectively consists of a set of Web pages. The methodology used to create these examples will be presented in Section 12.3.

By virtue of the SOM algorithm, the documents are positioned on a two-dimensional grid so that nearby locations contain related documents. When the SOM results are transformed into HTML pages, this document collection can be easily explored using a WWW-based browsing environment. During browsing the user may zoom in on any map area by clicking on the map image to view the underlying document space in more detail. The WEBSOM browsing interface is implemented as a set of HTML documents that can be viewed using any graphical WWW browser.

12.2.1 Map of Newsgroup Articles

The Usenet discussion groups on the Internet are somewhat like public bulletin boards or perhaps even more like unmoderated "Letters to the Editor". There are huge numbers of groups, each devoted to a different topic. In the groups people discuss topics of their own interest, ask for information or advice, or offer information, pictures or services. The messages in the groups (usually called articles) are colloquial, mostly rather poorly written, short documents that often contain little topical information. It is not easy to organize them properly. In this sense they form a kind of worst-case scenario for demonstrating organization of text collections or documents.

We retrieved from the Usenet groups a collection of 4600 full-text documents containing approximately 1,200,000 words and organized these with the WEBSOM method. The collection we selected consisted of all the articles that appeared during the latter half of 1995 in the Usenet newsgroup "comp.ai.neural-nets". These 4600 documents were organized on a map of size 24 by 32 nodes. After the map was formed new articles could be added to the map without re-computing it; the goodness of the result naturally depends on how much the topics have changed. At the end of March 1997 the map contained some 12,000 documents.

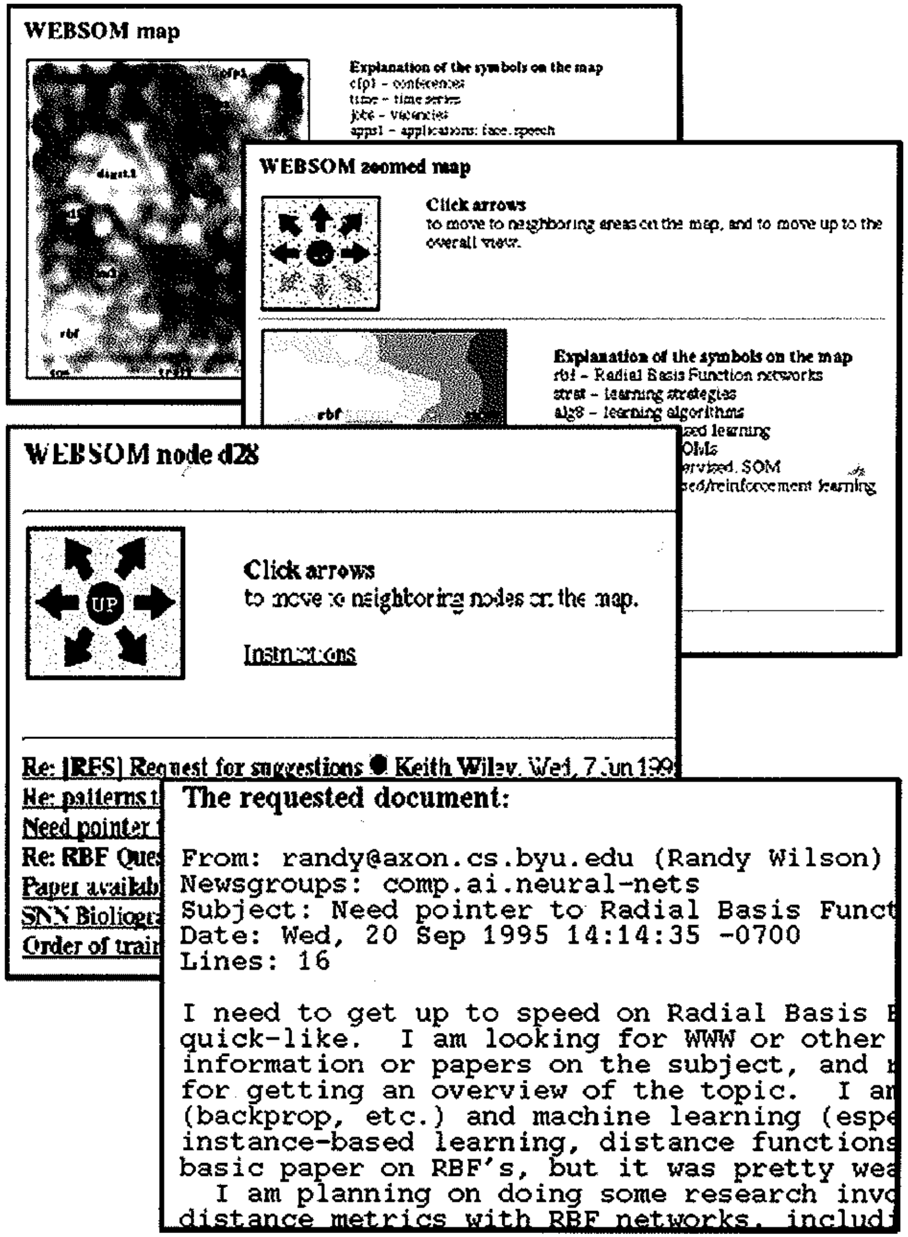
The WEBSOM browsing interface provides several views of this document collection, at different resolutions. A view of the whole map offers a general overview of the whole document collection (the top view in Figure 12.1, better visible in Figure 12.2). The small dots in the display denote the nodes on the map, and the similarity of the contents of the nearby documents in different parts of the map is indicated by the shades of gray. White denotes a relatively coherent area, whereas dark regions denote major shifts in the content of the documents. This toplevel display may be on to a zoomed map view, to a specific map node, and finally to a single document. These four view levels are shown in Figure 12.1 in the order of increasing detail. The first two levels provide a graphical look at the map display, first at the overall level view and then by providing a closer look at the selected area. As one goes deeper into the details by moving to the next level the contents of an individual node are revealed, and finally a document is seen (the view at the bottom of Figure 12.1).

In a typical browsing session, the user may start from the overall map view and proceed to examine further a specific area, perhaps later gradually wandering to nearby areas containing related information. After finding a particularly interesting map node the user can use it as a "document bin" which can be bookmarked and checked regularly to see if interesting new articles have arrived.

Let us now examine a specific area in an organized map. As shown in Figure 12.2, WEBSOM positioned articles related to financial issues and fuzzy logic in the middle of the right edge of the map. A closer look at the contents of a few nodes (Figure 12.3) shows that a continuum can be found between discussions of economic applications on the one hand and fuzzy set theory and neural networks methodology on the other. The strictly economy-related node, the titles of which are shown in box 5 in Figure 12.3, contains articles related to financial applications of neural networks, specifically bankruptcy predictions. Moving up on the map, the nodes begin to have articles regarding fuzzy logic in addition to economic issues (inset 3). Finally, in node 2 there are no more articles on economic issues, and fuzzy neural nets seem to have taken over the discussion.

The example shown in Figure 12.3 demonstrates the principles according to which WEBSOM organizes documents. In general the map tries to model the "document space", the space in which the contents of the documents in the document collection can be represented so that nearby locations would represent similar documents. The reason why the articles on neuro-fuzzy systems and economic applications appear right next to each other, and even overlap on the map, is that many of the articles mention both subjects. There is thus an association between these two topics which might a priori seem independent.

In addition to explorations, WEBSOM may be used for content-directed



WEBSOM map

Explanation of the symbols on the map
cfp1 - conferences
time - time series
jobs - vacancies
appl - applications: face, speech

WEBSOM zoomed map

Click arrows
to move to neighboring areas on the map, and to move up to the overall view.

Explanation of the symbols on the map
rbf - Radial Basis Function networks
strat - learning strategies
alg8 - learning algorithms
zed learning
OMs
ervised SOM
sed/reinforcement learning

WEBSOM node d28

Click arrows
to move to neighboring nodes on the map.

Instructions

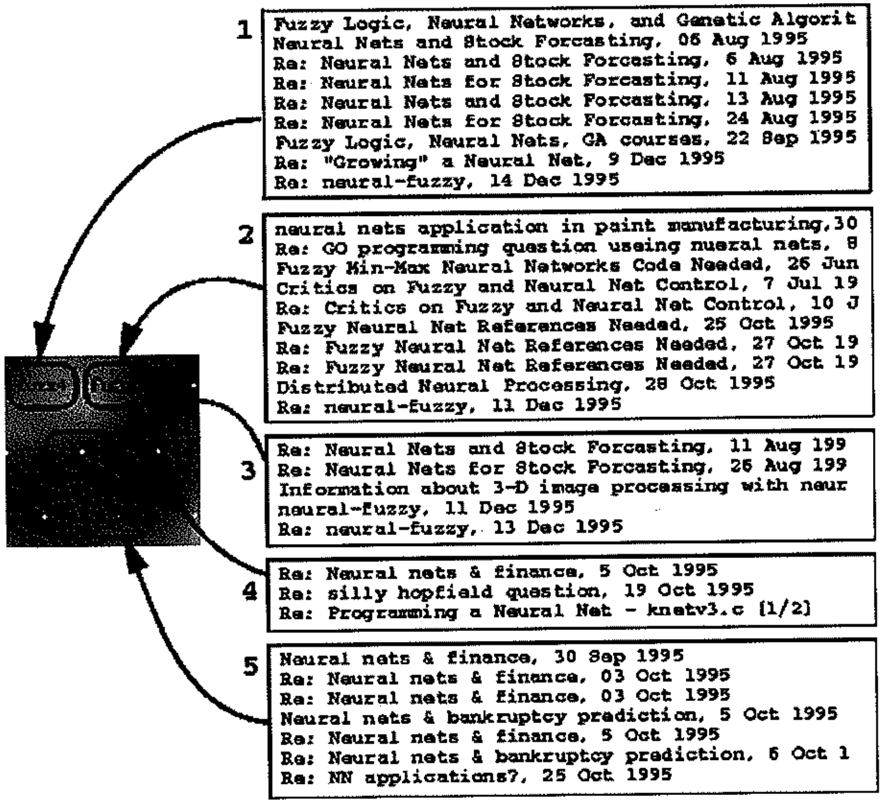
Re: [RFS] Request for suggestions ● Keith Wiley, Wed, 7 Jun 199
Re: patterns t
Need pointer t
Re: RBF Que
Paper availab
SNN Biologi
Order of trai

The requested document:

```
From: randy@axon.cs.byu.edu (Randy Wilson)
Newsgroups: comp.ai.neural-nets
Subject: Need pointer to Radial Basis Funct
Date: Wed, 20 Sep 1995 14:14:35 -0700
Lines: 16

I need to get up to speed on Radial Basis B
quick-like.  I am looking for WWW or other
information or papers on the subject, and r
for getting an overview of the topic.  I am
(backprop, etc.) and machine learning (espe
instance-based learning, distance functions
basic paper on RBF's, but it was pretty wea
  I am planning on doing some research invo
distance metrics with RBF networks, includi
```

Figure 12.1. The four different view levels of the WEBSOM browsing interface: the whole map, the zoomed map, the map node, and the document view, presented in the order of increasing detail. Moving between the levels or to neighboring areas on the same level is done by mouse clicks on the images or on the document links. Once an interesting area has been found on the map, exploring the related documents in the neighboring areas is simple. This can be contrasted with traditional information retrieval techniques where the users often cannot know whether there is a considerable number of relevant documents just "outside" their search results.



1
Fuzzy Logic, Neural Networks, and Genetic Algorit
Neural Nets and Stock Forcasting, 06 Aug 1995
Re: Neural Nets and Stock Forcasting, 6 Aug 1995
Re: Neural Nets for Stock Forcasting, 11 Aug 1995
Re: Neural Nets and Stock Forcasting, 13 Aug 1995
Re: Neural Nets for Stock Forcasting, 24 Aug 1995
Fuzzy Logic, Neural Nets, GA courses, 22 Sep 1995
Re: "Growing" a Neural Net, 9 Dec 1995
Re: neural-fuzzy, 14 Dec 1995

2
neural nets application in paint manufacturing,30
Re: GO programming question useing nueral nets, 8
Fuzzy Min-Max Neural Networks Code Needed, 26 Jun
Critics on Fuzzy and Neural Net Control, 7 Jul 19
Re: Critics on Fuzzy and Neural Net Control, 10 J
Fuzzy Neural Net References Needed, 25 Oct 1995
Re: Fuzzy Neural Net References Needed, 27 Oct 19
Re: Fuzzy Neural Net References Needed, 27 Oct 19
Distributed Neural Processing, 28 Oct 1995
Re: neural-fuzzy, 11 Dec 1995

3
Re: Neural Nets and Stock Forcasting, 11 Aug 199
Re: Neural Nets for Stock Forcasting, 26 Aug 199
Information about 3-D image processing with neur
neural-fuzzy, 11 Dec 1995
Re: neural-fuzzy, 13 Dec 1995

4
Re: Neural nets & finance, 5 Oct 1995
Re: silly hopfield question, 19 Oct 1995
Re: Programming a Neural Net - knetv3.c [1/2]

5
Neural nets & finance, 30 Sep 1995
Re: Neural nets & finance, 03 Oct 1995
Re: Neural nets & finance, 03 Oct 1995
Neural nets & bankruptcy prediction, 5 Oct 1995
Re: Neural nets & finance, 5 Oct 1995
Re: Neural nets & bankruptcy prediction, 6 Oct 1
Re: NN applications?, 25 Oct 1995

Figure 12.2. Schematic illustration of a WEBSOM document map used as an information filtering tool. The circle denotes the user's interest area. The symbols inside the circle denote documents that would be selected by the system automatically. Those documents could, for example, be instances of interesting electronic mail or articles from a news supplier. The visual map display can also be used to aid in noticing and checking for related documents in nearby areas.

document search. Any new document, for example a free-text query or an otherwise interesting document, may be mapped onto the document display precisely like any of the previous articles. The position of the new document on the document map provides a starting point for exploring related documents in nearby areas. Furthermore, relevant areas on the document map can be used as "mailboxes" into which target information is automatically gathered (Figure 12.2).

12.2.2 Map of Conference Abstracts

Another example of the application of WEBSOM is one involving the abstracts of articles accepted for presentation in the Workshop on Self-Organizing Maps held in Helsinki in 1997 (WSOM'97). These abstracts were organized using a simplified form of the WEBSOM method. The resulting map offers a visual overview of the

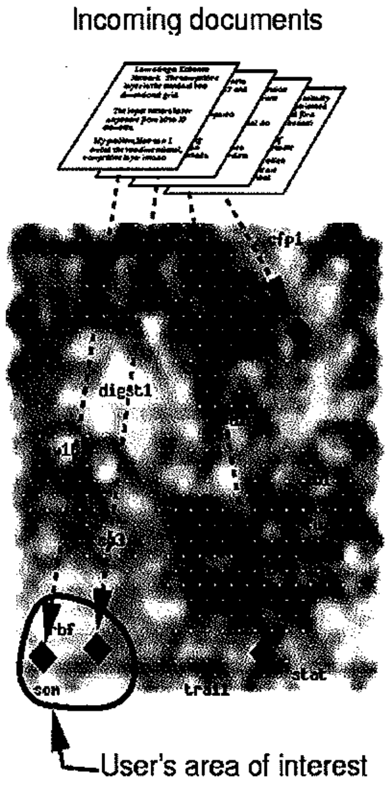


Figure 12.3. An area of the map of neural network articles that contains articles on economic applications of neural networks and neuro-fuzzy systems in neighboring map nodes.

workshop contents. The interactive version of the map can be reached via the WWW address *http://websom.hut.fi/websom/*.

The document collection contained 58 short abstracts, with on average 106 words per abstract, excluding articles ("a", "the", "an") and including the title of the abstract. An abstract of a scientific article may typically introduce the main components of the methodology used, discuss the relations of the work to earlier studies, mention performed simulations and comparisons, and portray an application area of the method or perhaps even several potential applications. Thus, each article may have common subjects of discussion with many other articles. In other words, most of the articles will be related to many different kinds of articles and therefore the actual local dimensionality of the document space may be considerably greater in the case of scientific abstracts than with most Usenet newsgroup articles.

When organizing the articles all of the relations cannot be taken into account. The method must concentrate on visualizing the most salient ones. Furthermore, such grounds for organization may be different in different areas of the document space.

The map of the conference abstracts is shown in Figure 12.4, labeled by manually selected words. After viewing the node contents, many rather interesting small clusters seem to emerge, for example the text processing, finance, and speech recognition areas. In addition, a more extensive area of related theoretical issues can be found under and around the "topology" and "organization" labels. A number of papers discussing a specific application of SOM to the analysis of data sets or to monitoring processes, as well as articles on SOM-based tools, can be found near the visualization label.

Exploration of the map leads to the impression that some areas were organized more based on similarity in theoretical issues (e.g., an area where the topology preservation and organization measures are discussed), whereas in other areas the common factors between abstracts were more often related to the application area, such as in the text-processing area. In still other map areas the abstracts seemed to have both kind of commonalities.

Speech recognition is one example; three of the four abstracts in three nearby nodes mention speech recognition, whereas the fourth abstract seems to be drawn to one of the nodes because it discusses the same theoretical issues as the other abstracts in the same node (i.e., Gaussian mixtures).

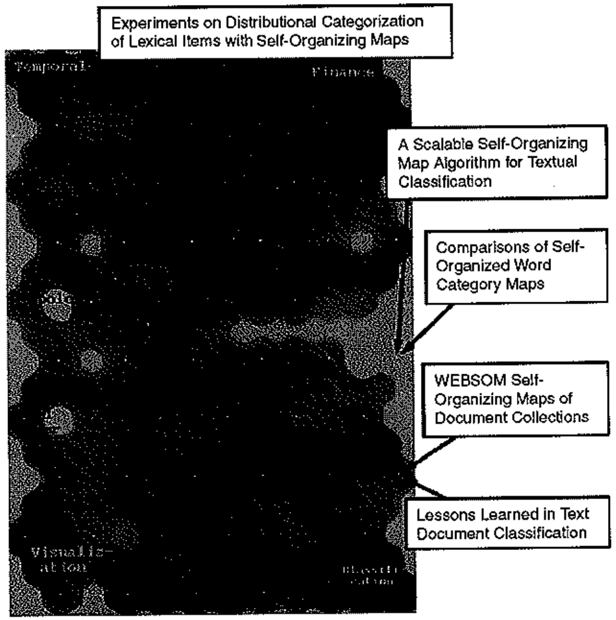


Figure 12.4. The WSOM'97 abstract map. The map has been labeled manually, and the titles of the workshop session "Text and document maps" are shown in the boxes. One of the articles, the one whose title is shown in the topmost box, has found a place quite apart from the other articles in the session. When reading the abstracts the reason becomes evident: the topmost abstract concentrates on linguistic and grammatical aspects and therefore also uses rather different language than the others which have a more practical approach to organizing real-world text and documents.

The labels "Theory" and "Applications" have been included to indicate the perceived grounds of map organization: on the upper left side, abstracts more often seem to be found together because of their theoretical commonalities, whereas lower and right-side areas (except a single node in the lower right corner) appear to be organized mainly according to application considerations.

12.2.3 Other Applications

The idea of using SOM in the exploration of document collections became more widely known in 1990 after the publication of an article by Doszkocs et al. (1990). They quoted Kohonen who said that SOM "is able to represent rather complicated hierarchical relations of high-dimensional spaces in a two-dimensional display." They concluded that document spaces are certainly such high-dimensional spaces. Concrete experiments based on the idea were first published by Lin et al. (1991) and Scholtes (1991), and related approaches have been published, e.g., by Chen et al. (1996) and Merkl (1993).

In the World Wide Web, one obvious application of the document maps is the ordering of home pages, or in fact any available documents. For example, an information provider may organize its material for easier public use. A company may also use the WEBSOM in an Intranet application. Often, the necessary information needed in a particular task can be found inside the organization but the means for finding it are needed. Also electronic mail messages could be automatically positioned on a suitable map according to personal interests. Relevant areas and single nodes on the map can be used as "mailboxes" into which specified information is automatically gathered. The method could also be used to organize official letters, personal files, library collections, and corporate full-text databases. A document map or a collection of maps provides an integral solution.

Administrative or legal documents may be difficult to locate by traditional information retrieval methods because of the specialized terminologies used. For instance, the product developers of a company are likely to express themselves in different terms and paraphrases than the marketing staff. The category-based and redundantly encoded approach of the WEBSOM method is expected to diminish the terminology problem.

12.3 Document Map Creation

12.3.1 Document Encoding

Given the above examples we will now explain the methodology for creating document maps.

The documents must be encoded before they can be organized. The ordering of the document maps depends on the chosen document-encoding scheme. Therefore the encoding should retain the relationships between the contents of the documents, while still being computationally efficient. When aiming at an organization that is based on the topical content of the documents it is useful to discard information that is irrelevant in distinguishing different topic areas. Examples of such information might be exact synonyms and "connector words" used by the authors of the documents.

Perhaps the most straightforward approach would be to limit to encoding only the titles of the articles instead of the full text, hoping that the titles summarize the essentials of the content. However, in most cases this choice is unsatisfactory because the title often gives a very limited or even a misleading view of the contents of an article. Statistical variation is better overcome if more information can be utilized for each article.

A straightforward document encoding is achieved by representing each document as a histogram in some "relevant" vocabulary. In other words, the document is represented by a real-number vector, where each word of the vocabulary corresponds to a component. The value of a component tells how many times the corresponding word appeared in the document, or is a suitable function of this frequency of occurrence. This encoding method is often referred to as the vector space model (Salton and McGill, 1983). Another version of this approach is to calculate occurrences of several consecutive words or letters, *n*-grams, instead of single words. Clearly, if the vocabulary is large, the document vectors are dimensionally too large for practical computations.

For computational reasons the vector space model used is suitable only in situations where the relevant vocabulary is for some reason very small, such as when the document collection is very tightly concentrated around some topic or when the relevant vocabulary is chosen manually. The vector space model may also be used when the document collection is small and the computational burden is thus not so big.

Another problem with vocabulary-based representations, besides the computational burden, is that they do not take into account the synonymy, or in general the fact that some pairs of words are more similar or more related than others. In the vector space model each word is equally distant from any other word. The problem becomes emphasized when the material to be organized has been written by different authors who may favor different choices of words. In a useful full-text analysis method synonymous expressions should therefore be encoded similarly, and in a computationally effective manner.

12.3.2 Two-level Architecture of the WEBSOM Method

WEBSOM is an explorative full-text information retrieval method that is based on applying the SOM in two processing stages on the document collection.

First, word category maps (also called self-organizing semantic maps; see Ritter and Kohonen, 1989) are formed. They are SOMs that have been organized according to word similarities. A simple way of measuring the similarity of words is to compare their average short contexts. Each word in the sequence of words is first represented by a fixed *n*-dimensional real vector with random-number components. The averaged context vector of a word then consists of the estimate of the expected value of the representations of the words that have co-occurred with it in a text corpus. The word category map is a SOM that has organized the words based on their context vectors.

The map is calibrated after the training process by inputting the word context instances once again to the word category map and labeling the best-matching nodes according to the words. Usually, interrelated words that have similar contexts appear close to each other on the map (Figure 12.5). Each node may become labeled by several symbols, often synonymous or belonging to the same closed category, thus forming "word categories" in the nodes.

To reduce the computational load the words that occur only a few times in the whole text corpus were neglected before the analysis and treated as empty slots. In order to emphasize the subject matter of the articles and to reduce erratic variations caused by the different discussion styles, common words that are not supposed to discriminate any discussion topics were discarded from the vocabulary.

In the second level of analysis the documents are encoded by mapping their text, word by word, onto the word category map, whereby a histogram of the "hits" on it is formed. To speed up the computation, the positions of the word labels on the word category map may be looked up by hash coding. To reduce the sensitivity of the histogram to small variations in the document content, the histograms are "blurred" on the two-dimensional map using a Gaussian convolution kernel. The document map is then formed with the SOM algorithm using the histograms as "fingerprints" of the documents.

The document map has been found to reflect relations between documents; similar documents tend to occur near each other on the map. Not all nodes may be well focused on one subject only, however. While most discussions seem to be confined into rather small areas on the map, the discussions may also overlap.

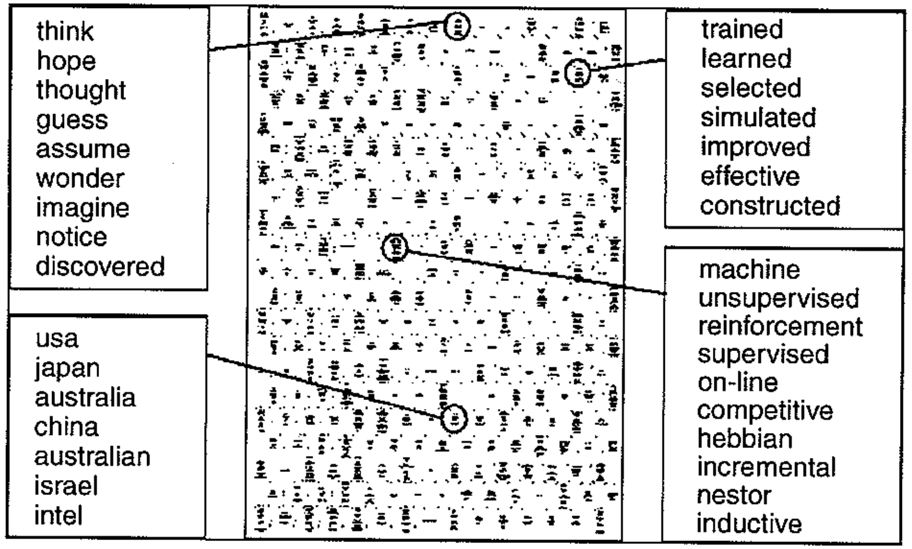


Figure 12.5. A word category map. The contents of four map nodes are shown in the boxes.

12.3.3 Other Approaches

There exist also other document encoding schemes which try to take into account the relations of different words in the encoding. In the latent semantic indexing (LSI) (Deerwester et al., 1990) each word is encoded with a vector that reflects the co-occurrence of the words in the same documents. Words that have occurred often in the same documents will attain similar codes, and the documents will be encoded as weighted sums of the codes of their words. The codes will be formed by a linear method that is based on matrix algebra.

Another related method has been used in the HNC's MatchPlus system (Gallant, 1991). Each word is encoded with a "context vector" and each document is encoded as a weighted sum of the context vectors of its words. The context vectors can be formed by means of linear algebra somewhat like in the LSI (Hecht-Nielsen, 1994), or they can be formed by judging manually how similar each word is to a set of "basis words" (Gallant et al., 1992). Each dimension of the context vectors then represents the manually evaluated similarity of meaning between the word to be encoded and one of the basis words.

To achieve maximal generality of the results we have used only the textual content of the documents for the encoding in the WEBSOM. It would also be possible to incorporate some prior information about the words, such as information obtained using a thesaurus. If the text is written in the form of a hypertext it is possible to incorporate information about the linkages between the documents (Girardin, 1995). Likewise, in scientific texts it is possible to include information about references to other documents.

12.4 Conclusions

Using Self-Organizing Maps for analyzing and visualizing large document collections is a novel approach for information retrieval and data mining. The potential of the method is based on its capability to generate overall views of document collections. The views can be used for exploration, and the map can also be used for associative search. Moreover, the word category maps help in finding a "common language" for different authors and for the users of the information. It is not necessary to use exactly the same words and phrases as is required in keyword-based systems.

The widespread commercial use of the WEBSOM naturally requires system integration in which the new technology is adopted to be used alongside the traditional word processing and information retrieval systems. In an optimistic scenario, the map becomes familiar for all computer users who need to deal with large text document collections. It remains to be seen whether this kind of development leads to inclusion of neural networks technology into operating systems and even commonly used hardware. One could then replace "c" by "s" in the word "computer"...

Springer Finance

Risk-Neutral Valuation: Pricing and Hedging of Financial Derivatives
N.H.Bingham and Rüdiger Kiesel
ISBN 1-85233-001-5 (1998)

Mathematics of Financial Markets
Robert J. Elliott and P. Ekkehard Kopp
ISBN 0-387-98553-0 (1998)

Mathematical Models of Financial Derivatives
Y.-K. Kwok
ISBN 981-3083-25-5 (1998)

Guido Deboeck and Teuvo Kohonen (Eds)

Visual Explorations in Finance

with Self-Organizing Maps

With 129 Figures
including 12 Color Plates

Springer

Guido Deboeck, PhD
3850 North River Street, Arlington, VA 22207, USA

Teuvo Kohonen, PhD
Helsinki University of Technology, Neural Networks Research Centre,
P.O. Box 2200, FIN-02015 HUT, Finland

ISBN 3-540-76266-3 Springer-Verlag Berlin Heidelberg New York

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library
Visual explorations in finance: with self-organizing maps.
- (Springer finance; 3674)
1. Finance 2. Finance - Computer simulation
I. Deboeck, Guido J. II. Kohonen, Teuvo
332
ISBN 3540762663

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress
Deboeck, Guido.
Visual explorations in finance with self-organizing maps/Guido J. Deboeck and Teuvo Kohonen (eds.).
p. cm. - (Springer finance)
Includes bibliographical references and index.
ISBN 3-540-76266-3 (case bound)
1. Finance - Decision making - Data processing. 2. Neural networks (Computer science) 3. Self-organizing systems. I. Kohonen, Teuvo.
II. Title. III. Series. 98-11388
HG4012.5.D43 1998
332'.0285 - dc21 CIP

The use of registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant laws and regulations and therefore free for general use.

The publisher makes no representation, express or implified, with regard to the accuracy of the information contained in this book and cannot accept any legal responsibility or liability for any errors or omissions that may be made.

Typeset by The Midlands Book Typesetting Company, Loughborough
Printed and bound at the Athenæum Press Ltd., Gateshead, Tyne & Wear
12/3830-543210 Printed on acid-free paper

Dedication

Rose-Maelle F. (Haiti)
Tafadzwa G. (Zimbabwe)
Kohinoor A. (Bangladesh)
Alenjie A. (Philippines)
Antony V. (Peru)
Judy Njoki M. (Kenya)
Murtopo (Indonesia)

Any royalties we receive from this book will be applied to the health and education of our foster kids through Childreach Inc., the US member of PLAN International. Childreach Inc., a global, child-focused development organization with counterparts in 12 countries contributes to programs that help over one million children in 40 countries in the world. To become a sponsor or learn more about Childreach visit *www.childreach.org*.