Newsgroup Exploration with WEBSOM Method and Browsing Interface

Timo Honkela, Samuel Kaski, Krista Lagus, and Teuvo Kohonen

Helsinki University of Technology
Faculty of Information Technology
Laboratory of Computer and Information Science **Report A32**

Otaniemi 1996

Newsgroup Exploration with WEBSOM Method and Browsing Interface

Timo Honkela, Samuel Kaski, Krista Lagus, and Teuvo Kohonen

Helsinki University of Technology
Faculty of Information Technology
Laboratory of Computer and Information Science
Rakentajanaukio 2 C, SF-02150 Espoo, FINLAND

Report A32
January 1996

ISBN 951-22-2949-8
ISSN 0783-7445
TKK OFFSET

Newsgroup Exploration with WEBSOM Method and Browsing Interface

Timo Honkela, Samuel Kaski, Krista Lagus, and Teuvo Kohonen

Helsinki University of Technology
Neural Networks Research Centre
Rakentajanaukio 2 C
SF-02150 Espoo, FINLAND

Abstract — The current availability of large collections of full-text documents in electronic form emphasizes the need for intelligent information retrieval techniques. Especially in the rapidly growing World Wide Web it is important to have methods for exploring miscellaneous document collections automatically. In the report, we introduce the WEBSOM method for this task. Self-Organizing Maps (SOMs) are used to position encoded documents onto a map that provides a general view into the text collection. The general view visualizes similarity relations between the documents on a map display, which can be utilized in exploring the material rather than having to rely on traditional search expressions. Similar documents become mapped close to each other. The potential of the WEBSOM method is demonstrated in a case study where articles from the Usenet newsgroup "comp.ai.neural-nets" are organized. The map is available for exploration at the WWW address http://websom.hut.fi/websom/

Contents

1 Introduction

Exploration of document collections may be supported by organizing the documents into taxonomies or hierarchies, a task that librarians have carried out throughout the past centuries. Also in the World Wide Web it is quite common to have large, manually ordered collections of hypertext links[1]. However, while the amount of the available textual information increases progressively, automatic methods for its management become necessary.

Efficient search engines have been developed to aid in the information retrieval task. Also in the WWW keyword search is in active use[2]. The basic problem with traditional search methods is the difficulty to devise suitable search expressions, which would neither leave out relevant documents, nor produce long listings of irrelevant hits. Even with a rather clear idea of the desired information it may be difficult to come up with all the suitable key terms and search expressions. Thus, a method of encoding the information based on, e.g., semantically homogeneous word categories rather than individual words would be helpful.

An even harder problem, for which search methods are usually not even expected to offer much support, is encountered when there exists only a vague idea of the object of interest. The same holds true if the area of interest resides at the outer edges of one's current knowledge.

The Self-Organizing Map (SOM) (Kohonen, 1982; Kohonen, 1995) is a means for automatically arranging high-dimensional statistical data so that alike inputs are in general mapped close to each other. The resulting map avails itself readily to visualization, and thus the distance relationships between different data items (such as texts) can be illustrated in a familiar and intuitive manner.

The SOM may be used to order document collections, but to form maps that display relations between document contents a suitable method must be devised for encoding the documents. The relations between the text contents need to be expressed explicitly.

If the words are first organized into categories on a word category map, then an encoding of the documents can be achieved that explicitly expresses the similarity of the word meanings. The encoded documents may then be organized with the SOM to produce a document map. The visualized document map provides a general view to the information contained in the document landscape, where changes between topics are generally smooth and no strict borders exist.

2 WEBSOM Method

The problem addressed by the WEBSOM method is to automatically order, or organize, arbitrary free-form textual document collections to enable their easier browsing and exploration.

Before ordering the documents they must be encoded; this is a crucial step since the ordering depends on the chosen encoding scheme. In principle, a document might be encoded as a

[1] Yahoo (http://www.yahoo.com/) can be mentioned as one example.

[2] As examples of search engines, one may mention, e.g., Altavista (http://www.altavista.digital.com), and Lycos (http://www.lycos.com). The approach presented in this report can be considered to be complementary to these traditional ones.

histogram of its words, whereby for computational reasons the order of the words is neglected. The computational burden would still, however, be orders of magnitude too large with the vast vocabularies used for automatic full-text analysis. An additional problem with the word histograms is that each word, irrespective of its meaning, contributes equally to the histogram. In a useful full-text analysis method synonymous expressions, however, should be encoded similarly.

Since it is not currently feasible to incorporate references to real-life experience of word meanings to a text analysis method, the remaining alternative is to use the statistics of the contexts of words to provide information on their relatedness. It has turned out that the size of the word histograms can be reduced to a fraction with the so-called "self-organizing semantic maps" (Ritter and Kohonen, 1989; Ritter and Kohonen, 1990). At the same time the semantic similarity of the words can be taken into account in encoding the documents.

The basic processing architecture of the WEBSOM method is presented in Fig. 1, and the details are described in the following sections.

2.1 Self-Organizing Map

The Self-Organizing Map (SOM) (Kohonen, 1982; Kohonen, 1995) is a general unsupervised learning method for ordering high-dimensional statistical data so that alike inputs are in general mapped close to each other.

We do not present the details of the application of the SOM algorithm here; they have been described in detail elsewhere, e.g., in the documentation of the SOM_PAK program package (Kohonen et al., 1996).

The WEBSOM method that has been used in the present experiments is also described in (Kaski et al., 1996). A more detailed description of the WWW interface of the WEBSOM is presented in (Lagus et al., 1996). In addition, we have carried out an experiment with a collection of articles from 20 different Usenet newsgroups, whereby partly supervised methods were used to enhance the separability of the different groups (Honkela et al., 1996).

Next we present the details of actual processing in WEBSOM method including preprocessing of the input, formation of the word category map, and, finally, formation of the document map.

2.2 Preprocessing

Before application of the Self-Organizing Map on the document collection we removed some non-textual information (e.g., ASCII drawings and automatically included signatures) from the newsgroup articles. Numerical expressions and special codes were treated with heuristic rules.

To reduce the computational load the words that occurred only a few times (say, less than 50 times) in the whole data base were neglected and treated as empty slots.

In order to emphasize the subject matters of the articles and to reduce erratic variations caused by the different discussion styles, the common words that are not supposed to dis-

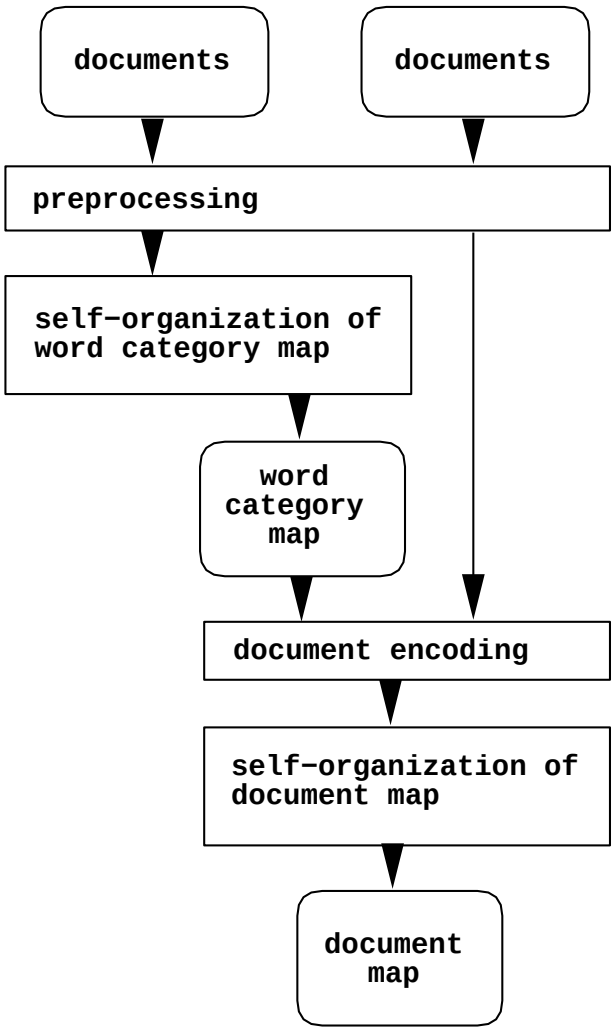


Figure 1: The basic architecture of the Websom method. The document map is organized based on documents encoded with the word category map. Both the maps are produced with the SOM algorithm.

criminate any discussion topics were discarded from the vocabulary. In the actual experiment, 800 common words (types) of the total of 2500 were removed.

2.3 Word category map

The *word category map* is a "self-organizing semantic map" (Ritter and Kohonen, 1989) that describes relations of words based on their averaged short contexts. The ith word in the sequence of words is represented by an n-dimensional real vector x_i with random-number components. The averaged context vector of this word reads

$$X(i) = \begin{bmatrix} \mathrm{E}\{x_{i-1}|x_i\} \\ \varepsilon x_i \\ \mathrm{E}\{x_{i+1}|x_i\} \end{bmatrix} , \tag{1}$$

where E denotes the estimate of the expectation value evaluated over the text corpus, and ε is a small scalar number. Now the $X(i) \in \Re^{3n}$ constitute the input vectors to the word category map. In our experiments $\varepsilon = 0.2$ and $n = 90$. The training set consists of all the $X(i)$ with different x_i.

The SOM is calibrated after the training process by inputting the $X(i)$ once again to the word category map and labeling the best-matching nodes according to symbols corresponding to the x_i parts of the $X(i)$. In this method a node may become labeled by several symbols, often synonymous or forming a closed attribute set. Usually interrelated words that have similar contexts appear close to each other on the map.

On the *word category map* similar words tended to occur in the same or nearby map nodes, forming "word categories" in the nodes. Sample categories are illustrated in Fig. 2. The map was computed using a massively parallel neurocomputer CNAPS, and fine-tuned with the SOM_PAK software (Kohonen et al., 1996).

2.4 Document map

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 reduce the sensitivity of the histogram to small variations in the document content, the histograms are "blurred" using a Gaussian convolution kernel[3]. Such "blurring" is a commonplace method in pattern recognition, and is justified also here, because the map is ordered. The *document map* is then formed with the SOM algorithm using the histograms as "fingerprints" of the documents. To speed up computation, the positions of the word labels on the word category map may be looked up by hash coding.

The document map was found to reflect relations between the newsgroup articles; similar articles tended to occur near each other on the map as was seen in the previous chapter. Not all nodes were 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. The visualized clustering tendency or density of the documents in different areas of the "digital library",

[3]In the experiment, the Gaussian convolution kernel had the full width at half maximum of two map spacings on the word category map consisting of 15 by 21 nodes.

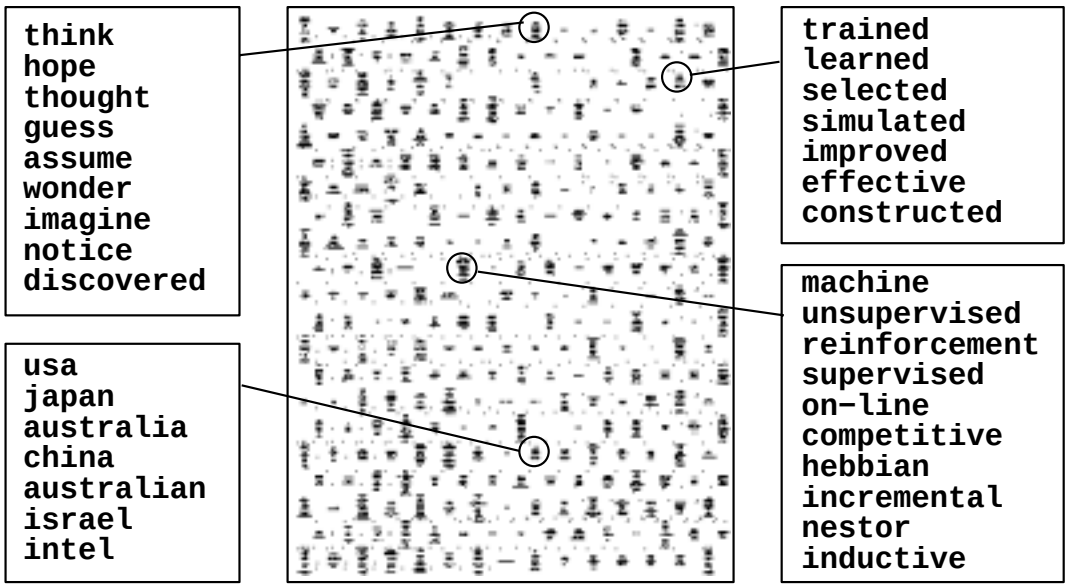


Figure 2: Examples of some clear "categories" of words on the word category map of the size of 15 by 21 nodes. The word labels of the map nodes have been shown with a tiny font on the map grid, and four nodes have been enlarged in the insets.

presented with the gray scale on the document map, can be used to aid in finding related articles.

The map was computed using the CNAPS neurocomputer, and fine-tuned with the SOM_PAK software (Kohonen et al., 1996).

2.5 Previous work on organizing documents with the SOM

Several studies have been published on self-organizing maps that map words into grammatical and semantic categories (Honkela et al., 1995; Miikkulainen, 1993; Ritter and Kohonen, 1989; Ritter and Kohonen, 1990; Scholtes, 1993). The SOM has also been utilized previously to form a small map based on titles of scientific documents by Lin et al. (Lin et al., 1991). Scholtes has developed, based on the SOM, a neural filter and a neural interest map for information retrieval (Scholtes, 1991a; Scholtes, 1991b; Scholtes, 1992; Scholtes, 1993). Merkl (Merkl, 1993; Merkl et al., 1994) has used the SOM to cluster textual descriptions of software library components. Quite recently we have also been informed of a SOM-based approach for organizing WWW-pages studied at the University of Arizona AI lab.

3 WEBSOM Browsing Interface

By virtue of the Self-Organizing Map algorithm, the documents are positioned on a two-dimensional grid, viz. the map, so that related documents appear close to each other. We have developed a WWW-based browsing environment, which utilizes the order of the map to aid in exploring the document space. The basic idea is that the user may zoom at any map area by clicking the map image to view the underlying document space in more detail.

The Websom browsing interface is implemented as a set of HTML[4] documents that can be viewed using a graphical WWW browser.

3.1 Document material: Usenet newsgroup comp.ai.neural-nets

The WEBSOM method is readily applicable to any kinds of collections of textual documents. To ensure that the method works in realistic situations, we selected material that is difficult enough from the textual analysis point of view – articles from a Usenet newsgroup. They are colloquial, mostly rather carelessly written short documents that contain little topical information to organize them properly. We have organized a collection of 4 600 full-text documents containing approximately a total of 1 200 000 words with the WEBSOM method. The collection consists of all the articles that have appeared during the latter half of 1995 in the Usenet newsgroup "comp.ai.neural-nets". After a map has been formed new articles can be added on the map without recomputing it. In the end of January 1995 the map contains some 5 000 documents.

3.2 Viewing the document collection

The view of the whole map offers a general overview on the whole document collection (Fig. 3). The display may be focused to a zoomed map view, to a specific node, and finally a single document. The four view levels are shown in Fig. 4 in the increasing order of detail. The first two levels display the graphical map, first the general view and then a closer look on the selected area. As one goes deeper into the details by moving to the next level, first the contents of an individual node are revealed, and finally a document is seen (the view in the lower right corner of Fig. 4).

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

Clickable arrow images are provided for moving around on the zoomed map level and for moving between between neighboring map nodes on the node level (see Fig. 4).

[4] HyperText Markup Language

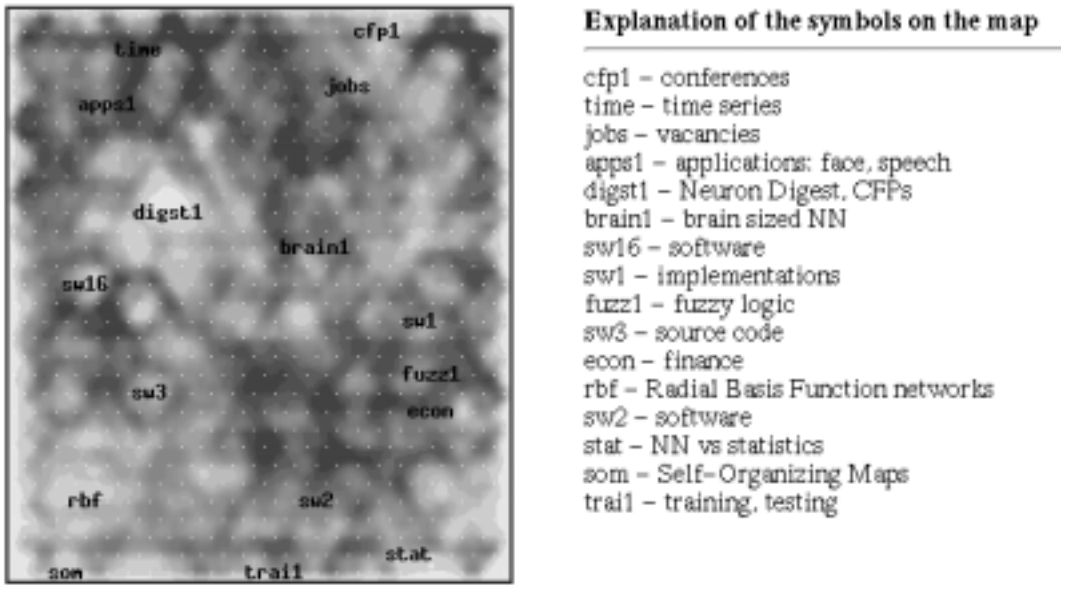


Figure 3: A map of 4600 documents. The small dots denote the nodes on the map of the size of 24 by 32 nodes, and the density, or clustering tendency in different parts of the map has been indicated by shades of gray. White denotes a high degree of clustering whereas black denotes long distances, "ravines", between the clusters.

3.3 Exploration example

As a detailed example of the results, we will examine a specific area in the ordered document map. In the lower right corner of the map (Fig. 3) the WEBSOM has positioned articles related to financial issues and fuzzy logic. The same area is portrayed in more detail in Fig. 5.

A closer look at the contents of a few nodes shows that a continuum can be found in this area between discussions of economic applications on the one hand and the fuzzy set theory and neural networks methodology on the other. The most strictly economy-related node, the titles of which are shown below, contains articles related to financial applications of neural networks, specifically bankruptcy prediction.

- Neural nets & finance
- Re: Neural nets & finance
- Neural nets & bankruptcy prediction
- Re: NN applications?

In a nearby node (see below) there are discussions about both fuzzy logic and economical issues, such as stock forecasting.

- Fuzzy Logic, Neural Networks, and Genetic Algorithms
- Neural Nets and Stock Forcasting
- Re: Neural Nets and Stock Forcasting
- Fuzzy Logic, Neural Nets, GA courses
- Re: "Growing" a Neural Net
- Re: neural-fuzzy

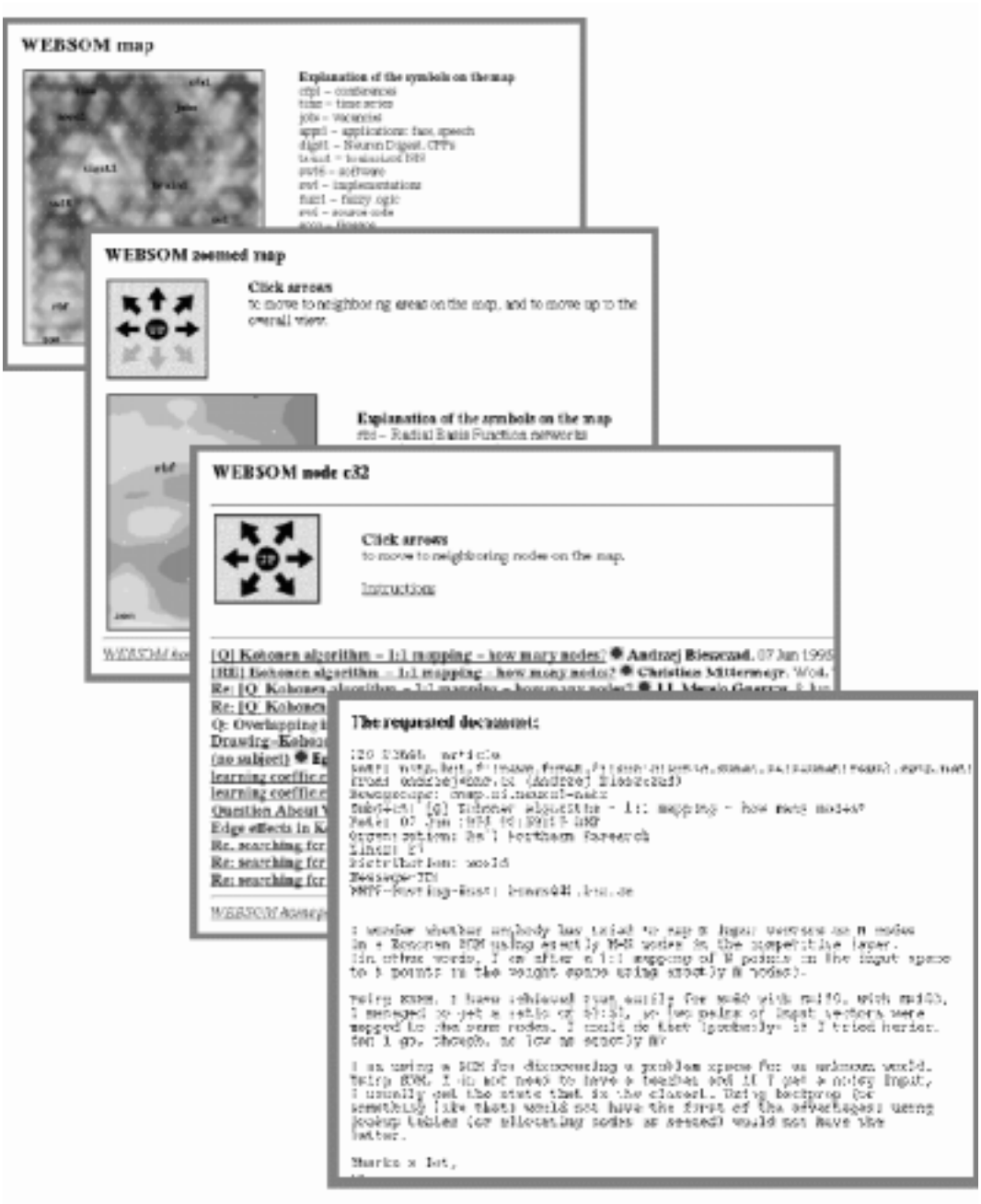
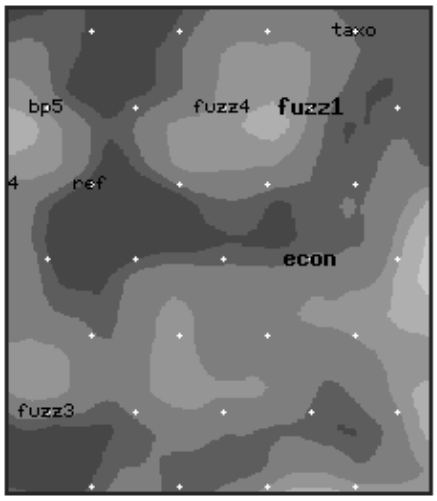


Figure 4: The four different view levels: the whole map, the zoomed map, the map node, and the document view, presented in the order of increasing detail. Moving between the levels or to neighboring areas on the same level is done by mouse clicks on the images or on the document links. Once an interesting area on the map has been found, exploring the related documents in the neighboring areas is simple. This can be contrasted with the traditional information retrieval techniques where the users cannot know whether there is a considerable number of relevant documents just "outside" their search results.



Explanation of the symbols on the map
taxo – taxonomy
bp5 – bp errors
fuzz4 – fuzzy logic, stock forecasting
fuzz1 – fuzzy logic
ref – reference queries
econ – finance
fuzz3 – fuzzy encoding

Figure 5: A zoomed view on the map. The labels reveal the main topics in the area: finance and fuzzy logic.

In a neighbor of the previously presented nodes, the discussion has moved altogether away from economic issues to fuzzy logic and neural methods. The corresponding titles of the articles have been listed below.

- neural nets application in paint manufacturing
- Re: GO programming question useing nueral nets.
- Fuzzy Min-Max Neural Networks Code Needed
- Critics on Fuzzy and Neural Net. Control
- Re: Critics on Fuzzy and Neural Net. Control
- Fuzzy Neural Net References Needed
- Re: Fuzzy Neural Net References Needed
- Distributed Neural Processing

As one can notice, spelling errors are not uncommon even in the most central words ("nueral", "forcasting"). We consider it important to develop methods that take into account also this typical phenomenon of everyday use of natural language. It may be necessary to emphasize here that the analysis of the documents is based on their text as a whole, not on the titles only.

With the WWW-browser one may click any of the titles in the node to bring the corresponding document into view.

4 Conclusions and Future Directions

In this work we have presented a novel methodology for ordering collections of documents, together with a browsing interface for exploring the resulting ordered map of the document space. The method, called the WEBSOM, performs a completely automatic and unsupervised full-text analysis of the document set using Self-Organizing Maps. The result of the analysis, an ordered map of the document space, displays directly the similarity relations of the subject

matters of the documents; they are reflected as distance relations on the document map. Moreover, the density of documents in different parts of the document space can be illustrated with shades of color on the document map display.

A large proportion of the articles in this Usenet newsgroup are either short queries and answers, or long articles dealing with a multitude of topics (e.g., digests). The short articles may not provide enough context for their proper placement on the map, while it might be most appropriate to place the longer articles to multiple locations. These deficiencies in the data may explain the partial overlap of the different discussion topics on the map. We intend to investigate whether scientific document collections were even better clustered.

References

Honkela, T., Kaski, S., Lagus, K., and Kohonen, T. (1996). Exploration of full-text databases with self-organizing maps. Submitted to ICNN-96, Washington D.C.

Honkela, T., Pulkki, V., and Kohonen, T. (1995). Contextual relations of words in Grimm tales analyzed by self-organizing map. In Fogelman-Soulié, F. and Gallinari, P., editors, *Proc. ICANN-95, Int. Conf. on Artificial Neural Networks*, volume II, pages 3–7, Paris. EC2 et Cie.

Kaski, S., Honkela, T., Lagus, K., and Kohonen, T. (1996). Creating an order in digital libraries with self-organizing maps. Submitted to WCNN-96, San Diego, California.

Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biological Cybernetics*, 43:59–69.

Kohonen, T. (1995). *Self-Organizing Maps*. Springer, Berlin, Heidelberg.

Kohonen, T., Hynninen, J., Kangas, J., and Laaksonen, J. (1996). SOM_PAK: The self-organizing map program package. Technical Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science, Espoo, Finland.

Lagus, K., Kaski, S., Honkela, T., and Kohonen, T. (1996). Browsing digital libraries with the aid of self-organizing maps. Submitted to the Fifth International World Wide Web Conference, Paris, France.

Lin, X., Soergel, D., and Marchionini, G. (1991). A self-organizing semantic map for information retrieval. In *Proc. 14th. Ann. Int. ACM/SIGIR Conf. on R & D In Information Retrieval*, pages 262–269.

Merkl, D. (1993). Structuring software for reuse - the case of self-organizing maps. In *Proc. IJCNN-93-Nagoya, Int. Joint Conf. on Neural Networks*, volume III, pages 2468–2471, Piscataway, NJ. IEEE Service Center.

Merkl, D., Tjoa, A. M., and Kappel, G. (1994). A self-organizing map that learns the semantic similarity of reusable software components. In *Proc. ACNN'94, 5th Australian Conf. on Neural Networks*, pages 13–16.

Miikkulainen, R. (1993). *Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon, and Memory*. MIT Press, Cambridge, MA.

Ritter, H. and Kohonen, T. (1989). Self-organizing semantic maps. *Biological Cybernetics*, 61:241–254.

Ritter, H. and Kohonen, T. (1990). Learning 'semantotopic maps' from context. In Caudill, M., editor, *Proc. IJCNN-90-WASH-DC, Int. Joint Conf. on Neural Networks*, volume I, pages 23–26, Hillsdale, NJ. Lawrence Erlbaum.

Scholtes, J. C. (1991a). Kohonen feature maps in full-text data bases: A case study of the 1987 Pravda. In *Proc. Informatiewetenschap 1991, Nijmegen*, pages 203–220, Nijmegen, Netherlands. STINFON.

Scholtes, J. C. (1991b). Unsupervised learning and the information retrieval problem. In *Proc. IJCNN'91, Int. Joint Conf. on Neural Networks*, pages 95–100, Piscataway, NJ. IEEE Service Center.

Scholtes, J. C. (1992). Neural nets for free-text information filtering. In *Proc. 3rd Australian Conf. on Neural Nets, Canberra, Australia, February 3-5.*

Scholtes, J. C. (1993). *Neural Networks in Natural Language Processing and Information Retrieval.* PhD thesis, Universiteit van Amsterdam, Amsterdam, Netherlands.