In In Proceedings of the Fifth International World Wide Web Conference
WWW5, May 6-10, Paris, France, volume Poster Proceedings, pages 71-79. EPGL.

Browsing Digital Libraries with the Aid of Self-Organizing Maps

Krista Lagus, Samuel Kaski, Timo Honkela, and Teuvo Kohonen
Helsinki University of Technology
Neural Networks Research Centre
Rakentajanaukio 2 C, FIN-02150 Espoo, Finland
e-mail: Krista.Lagus@hut.fi

Abstract **---** Powerful methods for exploring and searching collections of free-form textual documents are needed to control the flood of digital information emerging from various sources. In this article we present a method, WEBSOM, for automatic organization of document collections based on full-text analysis using the Self-Organizing Map. The document collection is ordered on the map in such a way that similar documents lie near each other. The WWW-based user interface provides the basic functionalities necessary for intuitive exploration of the document space visualized with a map: moving on the map, zooming, and examining individual documents. We apply the method to the Usenet newsgroup *comp.ai.neural-nets*.

1 Introduction

One of the most time-consuming tasks of the Web users in particular is to find relevant information from the vast material available. Efficient search tools such as search engines have quickly emerged to aid in this endeavor. The basic problem with traditional search methods such as searching by keywords or by indexed contents 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. In these situations, devising search expressions is often like trying to find some object in total darkness, and the results are poor. If there were something like a map of the document collection at hand, a map where documents were ordered meaningfully according to their content, then even a vague knowledge of the connections of the desired information to something already familiar would be useful. Maps might help the exploration first by giving an idea of what the information space looks like, and then by guiding one to the information of interest. Furthermore, a visualized map of the information landscape might reveal surprising connections between different areas of knowledge, and even be an invaluable aid in exploring totally new areas.

The Self-Organizing Map (SOM) (Kohonen, 1982 and 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 word 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. Easy exploration of the document landscape may then be provided via WWW.

2 Browsing document maps with the WWW

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. A detailed description of the method will be given in section 3. 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.

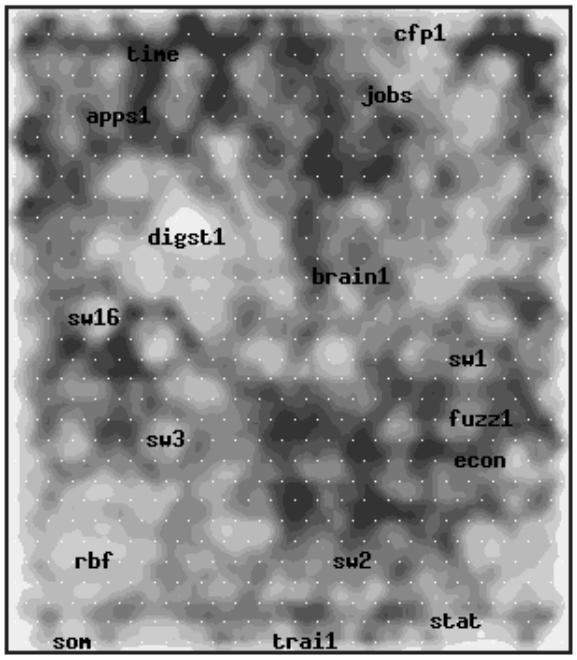
2.1 Document material: The Usenet newsgroup comp.ai.neural-nets

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 to use as input for the WEBSOM. We have organized a collection of 4600 full-text articles that appeared in the Usenet newsgroup "comp.ai.neural-nets" during the latter half of 1995, containing approximately a total of 1 200 000 words. The articles are colloquial, mostly rather carelessly written short documents that contain little topical information to organize them properly. Furthermore, spelling errors are not uncommon even in the most central words. After a map has been formed new articles can be added on the map without recomputing the whole map. In the end of January 1995 the map contains some 5000 documents.

2.2 Viewing the document collection

The view of the whole map offers a general overview on the whole document collection (Fig. 1). 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. 3 in 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. 3).

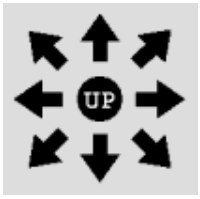
In a typical session, the user might start from the overall map view, and proceed to examine further a specific area, perhaps later gradually wandering to close-by areas containing related information. Clickable arrow images are provided for moving around on the zoomed map level (Fig. 2 a) and for moving between between neighboring map units on the node level (Fig. 2 b). After finding a particularly interesting node, one may use it as a ‘‘trap’’ or ‘‘document bin’’ which can be checked regularly to see if new interesting articles have arrived.



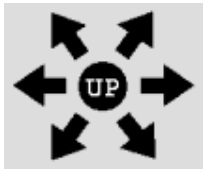
Explanation of the symbols on the map

cfp1 – conferences
time – time series
jobs – vacancies
apps1 – applications: face, speech
digst1 – Neuron Digest, CFPs
brain1 – brain sized NN
sw16 – software
sw1 – implementations
fuzz1 – fuzzy logic
sw3 – source code
econ – finance
rbf – Radial Basis Function networks
sw2 – software
stat – NN vs statistics
som – Self-Organizing Maps
trai1 – training, testing

Figure 1. *The overall view on the document collection. The user may click any area on the map to get a zoomed view. Each white dot marks a map node. Color denotes the density, or the clustering tendency of the documents. Light-colored areas are clusters and dark areas between the clusters more sparsely occupied zones.*



a)



b)

Figure 2. *a) Arrows are provided for moving into the neighboring areas on the zoomed map. A click in the middle takes one to the whole map view, and clicking any arrow causes a movement of the map view to the selected direction by half of the width of the current view. b) In the actual map nodes, an image with six arrows can be used for moving between neighboring nodes on the hexagonal map grid.*

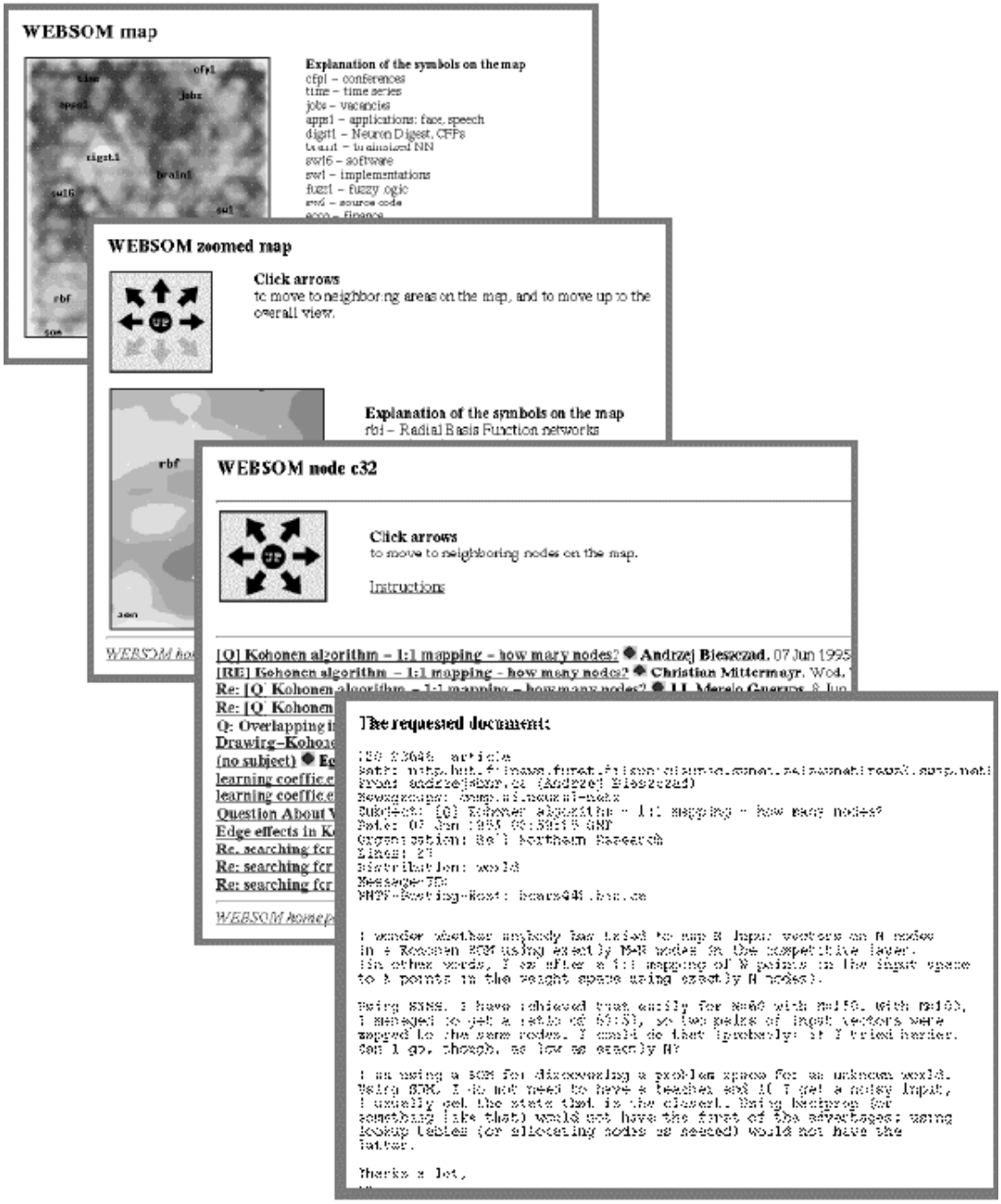
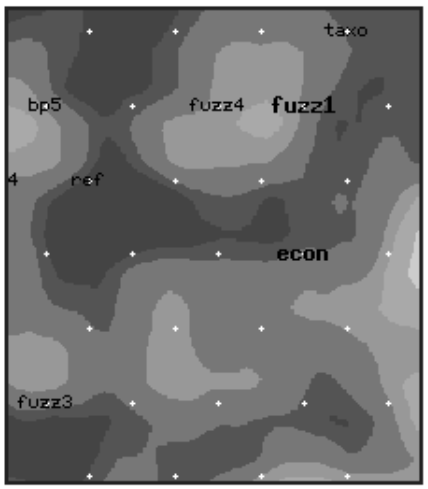


Figure 3. *The four different view levels: the whole map, the zoomed map, the map node, and the document view, presented in the order of increasing detail. Moving between the levels or to neighboring areas on the same level is done by mouse clicks on the images or on the document links. Once an interesting 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.*

2.3 An 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. 4.

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, portrayed in Fig. 5, contains articles related to financial applications of neural networks, specifically bankruptcy prediction. In a nearby node (Fig. 6) there are discussions both about fuzzy logic and economical issues such as stock forecasting. In its neighbor (Fig. 7), the discussion has moved altogether away from economic issues to fuzzy logic and neural methods. A click on any of the titles in the node brings the corresponding document into view.



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 4. *A zoomed view on the map. The labels reveal the main topics in the area, namely finance and fuzzy logic.*

Neural nets & finance ◆ **Bonaz**, 30 Sep 1995, Lines: 11.
Re: Neural nets & finance ◆ **Patrick Eiler**, Tue, 03 Oct 1995, Lines: 29.
Re: Neural nets & finance ◆ **Patrick Eiler**, Tue, 03 Oct 1995, Lines: 29.
Neural nets & bankruptcy prediction ◆ **Bonaz**, 5 Oct 1995, Lines: 6.
Re: Neural nets & finance ◆ **J. Feenstra**, 5 Oct 1995, Lines: 36.
Re: Neural nets & bankruptcy prediction ◆ **Peter Robinson**, Fri, 6 Oct 1995, Lines: 17.
Re: NN applications? ◆ **Gary Russel /ADVISOR L. FAUSETT**, Wed, 25 Oct 1995, Lines: 15.

Figure 5. *A node containing discussions on financial applications. This node is labeled "econ" in the zoomed map of Fig. 4.*

Fuzzy Logic, Neural Networks, and Genetic Algorithms ◆ **Hamid Berenji**, 20 Jul 1995, Lines: 54.
Neural Nets and Stock Forcasting ◆ **david rosner**, Sun, 06 Aug 1995, Lines: 17.
Neural Nets for Stock Forcasting ◆ **david rosner**, Sun, 06 Aug 1995, Lines: 17.
Re: Neural Nets and Stock Forcasting ◆ **Avatar/Silver**, 6 Aug 1995, Lines: 16.
Re: Neural Nets for Stock Forcasting ◆ **Harit P Trivedi**, 11 Aug 1995, Lines: 13.
Re: Neural Nets and Stock Forcasting ◆ **Peter Nolan**, 13 Aug 1995, Lines: 20.
Re: Neural Nets for Stock Forcasting ◆ **Andrew Lohbihler**, 24 Aug 95, Lines: 22.
Fuzzy Logic, Neural Nets, GA courses ◆ **IIS Corp**, Fri, 22 Sep 1995, Lines: 52.
Re: "Growing" a Neural Net ◆ **rudolph@inet.uni-c.dk**, 9 Dec 1995, Lines: 15.
Re: neural-fuzzy ◆ **Ari Eisenberg**, 14 Dec 1995, Lines: 18.

Figure 6. *A node containing articles related to stock forecasting using neural networks and fuzzy logic.*

neural nets application in paint manufacturing ◆ **ashi shah**, 30 May 1995, Lines: 5.
Re: GO programming question useing nueral nets. ◆ **Jorrit Tyberghein**, 8 Jun 1995, Lines: 36.
Fuzzy Min-Max Neural Networks Code Needed ◆ **Tim Purschke**, 26 Jun 1995, Lines: 18.
Critics on Fuzzy and Neural Net. Control ◆ **Ali Hariri**, 7 Jul 1995, Lines: 9.
Re: Critics on Fuzzy and Neural Net. Control ◆ **Ulf Nordlund**, 10 Jul 1995, Lines: 13.
Fuzzy Neural Net References Needed ◆ **Dean Alderucci**, 25 Oct 1995, Lines: 11.
Re: Fuzzy Neural Net References Needed ◆ **Maurice Clerc**, 27 Oct 1995, Lines: 27.
Re: Fuzzy Neural Net References Needed ◆ **Derek Long**, 27 Oct 1995, Lines: 24.
Distributed Neural Processing ◆ **Jon Mark Twomey**, 28 Oct 1995, Lines: 12.
Distributed Neural Processing ◆ **Jon Mark Twomey**, 28 Oct 1995, Lines: 12.
Re: neural-fuzzy ◆ **TiedNBound**, 11 Dec 1995, Lines: 10.

Figure 7. *A node containing articles on the use of neural and fuzzy methods.*

3 The WEBSOM method

The problem addressed by the WEBSOM method is to automatically *order*, or *organize*, collections of arbitrary free-form textual documents 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 histogram of its words, whereby for computational reasons the order of the words is neglected. Often 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 synonymic words, 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). At the same time the semantic similarity of the words can be taken into account in encoding the documents.

3.1 The SOM

To provide a general understanding of what the Self-Organizing Map is and why it is an especially suitable method for ordering large collections of text documents, the following expedition of thought may be helpful:

Consider an information processing system, such as the brain, which must learn to carry out very different tasks, each of them well. Let us assume that the system may assign different tasks to different sub-units that are able to learn from what they do. Each new task is given to the unit that can best complete the task. Since the units learn, and since they receive tasks that they can do well, they become even more competent in those tasks. This is a model of specialization by competitive learning. Furthermore, if the units are interconnected in such a way that also the *neighbors* of the unit carrying out a task are allowed to learn some of the task, the system slowly becomes ordered so that units near each other have similar abilities, and the abilities change slowly and smoothly over the whole system. This is the general principle of the Self-Organizing Map (SOM). The system is called a *map* and the task is to imitate, i.e., *represent* the input as well as possible. The representations become ordered according to their similarity relationships in an unsupervised learning process. This property makes the SOM useful for organizing large collections of data in general, including document collections.

3.2 Preprocessing text

Before applying the SOM on the document collection we automatically removed some non-textual information (e.g., ASCII drawings and automatically included signatures) from the newsgroup articles. Numerical expressions and several kinds of common code words were categorized with heuristic rules into a few classes of special symbols.

To reduce the computational load the words that occurred only a few times (in this experiment 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 variations caused by the different discussion styles, which were not of interest in this experiment, a group of common words that were not supposed to discriminate any discussion topics were discarded from the vocabulary. In the actual experiment, 800 common words of the total of 2500 were removed.

3.3 The word category map

On the word category map words are clustered into an ordered set of word categories. On the resulting map conceptually related words fall into the same or nearby word categories, which in turn are found in the neighboring map nodes. The ordering is formed by the SOM algorithm based on the average short contexts of the words (Ritter and Kohonen, 1989).

In our experiment, the word contexts were of length three, thus consisting of one preceding and one following word in addition to the word to be encoded. The ***i***th word in the sequence of words is represented by an ***n***-dimensional (here 90) real vector ***x(i)*** with random number components. The

averaged context vector of a word may be expressed as follows:

$$X(i) = \begin{bmatrix} \mathrm{E}\{x(i-1)|x(i)\} \\ 0.2x(i) \\ \mathrm{E}\{x(i+1)|x(i)\} \end{bmatrix} , \qquad (1)$$

where **E** denotes the conditional average over the whole text corpus. Now the real-valued vectors ***X(i)***, in our experiment of dimension 270, constitute the input vectors given to the word category map. During the training of the map, the averaged context vectors ***X(i)*** were used as input.

After the map has self-organized, each map node corresponds to a set of input vectors that are close to each other in the input space. The map node may thus be expected to approximate a set of similar inputs, here averaged word contexts. The process by which words are associated with map nodes is called calibration. In calibration, each node on the map is labeled with all the inputs ***X(i)*** for which the node is the best match, i.e., the closest representative. The word in the middle of the context, that is, the word corresponding to the ***x(i)*** part of the context vector, was used as the actual label of the node. In this method a unit may become labeled by several symbols, often synonymic or describing alternative or opposing positions or characteristics. Examples of map nodes have been illustrated in Fig. 9. Usually interrelated words that have similar contexts appear close to each other on the map.

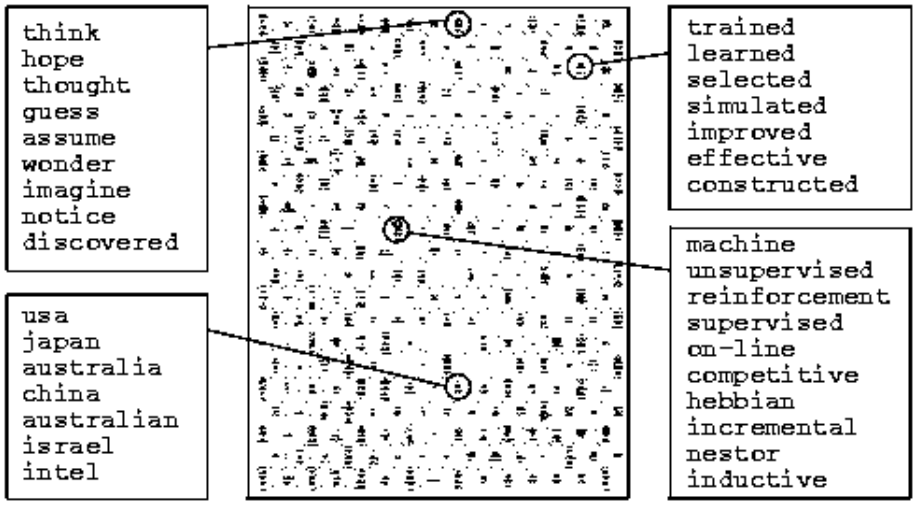


Figure 9. *Sample categories illustrated on a word category map. Similar words tend to occur in the same or nearby map nodes, forming ''word categories''. The map was computed using a massively parallel neurocomputer CNAPS, and fine-tuned with the SOM_PAK software (Kohonen, 1996).*

3.4 The document map

With the aid of the word category map the documents can be encoded as *word category histograms*. Very closely related words (that are in the same category on the map) then contribute identically to the code formed of the document.

It would also be advantageous if words in *similar* categories would contribute similarly to the codes of the documents. This is indeed possible to achieve since the word category map is ordered. The relations of the categories are reflected in their distances on the map. Therefore, the contributions of

nearby categories can be made similar by *smoothing* the histogram on the word category map, whereby the encoding becomes more invariant to the choices of words made by the writers of the documents. A moderately narrow (e.g., diminishing to the half of the maximum value within one map spacing) Gaussian function was found to be a suitable smoothing kernel in our recent experiments.

A representative sample of the encoded documents is presented as input to the SOM, which organizes them by unsupervised learning. After the learning process is complete the density of the documents, i.e. the clustering tendency in different regions of the document space, can be visualized on the document map to illustrate the relations of different map locations (Fig. 1). The map can be explored using the browsing interface described in section 2.

3.5 Previous work on organizing documents with the SOM

Several studies have been published on SOMs that map words into grammatical and semantic categories (e.g. Honkela et al.,1995; Miikkulainen, 1993; Ritter and Kohonen, 1989, 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. (1991). Scholtes has applied the SOM extensively in natural language processing (e.g., Scholtes, 1992a, 1992b) and developed a neural filter along with a neural interest map for information retrieval (Scholtes, 1991a, 1991b, 1992a, 1993). Merkl (1993, 1994) has used the SOM to cluster textual descriptions of software library components.

Quite recently we have also been informed of an approach studied at the University of Arizona AI Lab for organizing WWW-pages.

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, reflects 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 the documents in different parts of the document space can be illustrated with shades of color on the document map display.

The present version of the WEBSOM interface contains the basic functionality needed for exploring the document collection: moving on the document map, zooming into the map and viewing the contents of the nodes. Many different directions for enhancing the interface are possible. For example, the search for the location of the user’s own document on the map would form an ideal starting point for exploration. Most of the useful enhancements would greatly benefit from using client-side script languages such as Java, since the operations are often quite intensive computationally and therefore cannot be provided for widespread use in the server side. A three-dimensional interface based, e.g., on VRML might offer even more flexibility in representing the document space.

The WEBSOM (1996) tool is already available for exploring collections of Usenet newsgroups.

References

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

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

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

T. Kohonen, J. Hynninen, J. Kangas, and J. Laaksonen (1996) SOM_PAK: The Self-Organizing Map Program Package. Report A31, Helsinki University of Technology, Laboratory of Computer and Information Science.

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

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

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

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

H. Ritter and T. Kohonen (1989) Self-organizing semantic maps. *Biol. Cybern.*, 61(4):241--254.

H. Ritter and T. Kohonen (1990) Learning "semantotopic maps" from context. In M. Caudill, ed., *Proc. IJCNN-90-WASH-DC, Int. Joint Conf. on Neural Networks*, vol. I, pp. 23--26. Lawrence Erlbaum, Hillsdale, NJ.

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

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

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

J. C. Scholtes (1992b) Resolving linguistic ambiguities with a neural data-oriented parsing (DOP) system. In I. Aleksander and J. Taylor, eds., *Artificial Neural Networks, 2*, vol. II, pp. 1347--1350. North-Holland, Amsterdam, Netherlands.

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

WEBSOM home page (1996) Available at http://websom.hut.fi/websom/