Map of WSOM'97 Abstracts — Alternative Index

Krista Lagus

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

Abstract

The collection of abstracts of WSOM'97 articles was organized using a simplified form of the WEBSOM method. The result offers an approximate visual index to the workshop. The resulting map shows no clear general organizing criterion; instead the grounds for clustering certain documents seem to vary accross areas, sometimes even within a single node across pairs of articles. It is suggested that this results from the inherent large dimensionality of the document space.

1 Introduction

Abstracts of articles accepted for presentation in the Workshop on Self-Organizing Maps 1997 were organized using a simplified form of the WEBSOM method. The resulting map offers a visual overview of the workshop contents. The interactive version of the map can be explored at the WWW address `http://nucleus.hut.fi/wsom97/abstractmap/`.

The collection contained 58 short abstracts, on the average 106 words per abstract excluding articles (a, the, an), and including the title of the abstract. Many of the abstracts discussed both theoretical and application aspects, and often several different topics were mentioned in a single abstract.

The abstract collection differed from the previous application targets of the WEBSOM method in two significant ways. First, the size of this collection was very small, only 58 documents, whereas the other experiments were conducted on collections of thousands or even hundreds of thousands of documents [5, 2, 4, 3]. Second, in the earlier experiments the material consisted of Usenet newsgroup articles, which differ considerably from scientific abstracts both in the style of writing and often also in how many different "topics" worth visualizing as clustering on the map a single article typically addresses.

An abstract of a scientific article may, for example, introduce the main components of the methodology used, discuss the relations of the work to earlier studies, mention performed simulations and comparisons, as well as 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, the actual local dimensionality of the document space may be considerably greater in the case of scientific abstracts than with most Usenet newsgroup articles.

2 Method

For details of the WEBSOM method, see the description of the full WEBSOM system, e.g., elsewhere in this proceedings[1]. The small number of documents to be organized had two important consequences for the method: First, the word categorization was not needed for speeding the computation. Second, there was very little contextual information available of any other than the most common words, and thus any statistics of word contexts collected from the material would have been rather random-like. This also meant that the word category map could not be taught the same way as in the classical WEBSOM method.

As a consequence, either the word category map could not be used at all, or it should be organized based on contextual information gathered from some other document collection. The latter approach was tried, using 20000 articles from the Usenet newsgroup comp.ai.neural-nets to provide samples of word contexts, but the resulting word category map was not precise enough for successfully organizing the SOM-centered abstract collection. Our hypothesis is that this was due to the different styles of writing in the abstracts vs. the news articles. A change of text collection might have provided support for this hypothesis, but due to the lack of a suitable collection, we decided not to use the word category map. Instead of encoding the abstracts as histograms on word categories, they were encoded as histograms of words.

The total vocabulary, after numbers and other non-word codes were excluded, was 1723 unique words. Of these were excluded words that were too common in general, as well as SOM-related words such as "self-organizing" and "SOM", because all the abstracts are SOM-related. The size of the final coding vocabulary was 455 words. After encoding the documents as histograms on the vocabulary, the document vectors were organized on a self-organizing map of 6×10 units.

3 The Results

The resulting map is shown in the figures, labeled by manually selected words (Fig. 1) and by first authors of the articles (Fig. 2).

After viewing the nodes, many rather interesting small clusters seem to emerge, for example the text processing, finance and speech recognition areas. Also a more extensive area of related theoretical issues such as topology, topography, measures of organization and growing or dynamic SOMs 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.

Comparing the map with the division of the papers into sessions may also be of interest. Although the division of articles into workshop sessions may not be the canonical truth about clustering the documents, it is still the only human-made clustering available, and as such can be used as some kind of "ground truth". The spread of each session on the map is depicted in Fig. 3.

Two sessions, "Text" and "Financial analysis", are rather well clustered on the map (a poster on financial analysis is also located in the finance area of the map). In both sessions a single article becomes mapped rather separated from the rest, apparently due to dissimilar vocabulary. The isolated finance paper that falls under the visualization cluster, discusses the use of SOM in the analysis and visualization of statistical data, in this case of financial nature.

The "Theory" session is also quite clearly in the upper left corner of the map, 3 out of 5 papers even falling into the same node. Again one article is rather isolated, though. A majority of the articles in the session "Monitoring and data mining" appear on the bottom half of the map, near the visualization label. The "Temporal sequence processing" session seems to be divided into two clusters on the map, one in the upper left corner and the other near a node labeled "Local linear" in the middle left area of the map. Also some of the posters that were related to temporal processing were mapped into the upper left corner. The session called "Probabilistic interpretations", with only two abstracts received electronically from authors, appears near the upper left corner, again in the main area of theory-oriented organization. The last two sessions, "Image processing" and "Pattern and signal processing", do not form any notable clusters on the map, and neither, as expected, does the miscellaneous category "other".

Some tentative analysis of the general organization of the abstract map can also be made. Exploration of the map leads to the impression that some areas were organized more based on similarity in theoretical issues (e.g., the area of topology, topography and organization measures), whereas in other areas the common factors were more often related to the application area, such

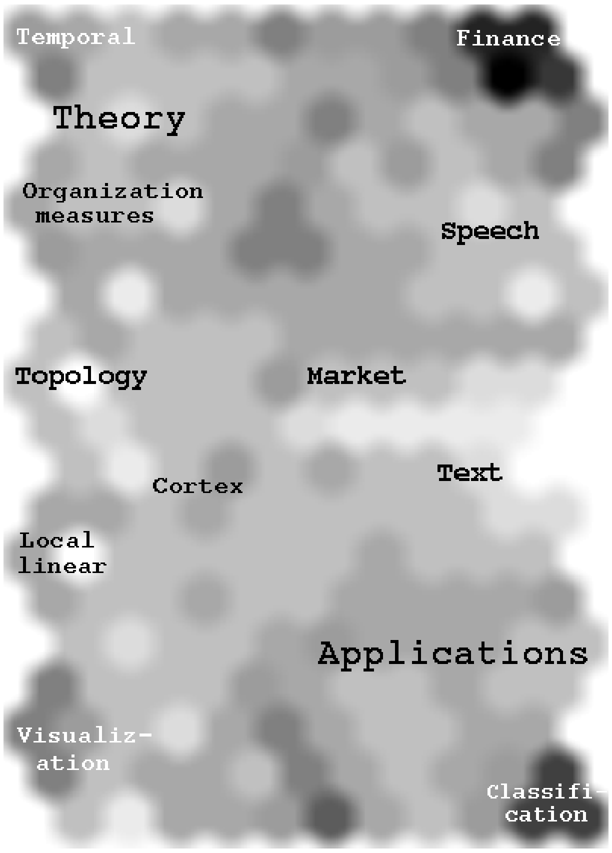

Figure 1: The WSOM'97 abstract map labeled manually. Some areas seem to be organized mainly based on similarity in theoretical considerations (area of topology and organization measures), whereas other areas have a common application area, such as text processing on the right side. Still others have commonalities in both, such as speech recognition. The labels “Theory” and “Applications” indicate the perceived grounds of organization: on the upper left side, articles more often seem to be found together because of the theoretical commonalities, whereas lower and right-side areas (despite a single node in the lower right corner) appear to be organized mainly according to application considerations.

as in the text processing area. Still other areas seemed to have both kind of commonalities. An example is speech recognition, where 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 abstract in the same node. This general approximate division is depicted in Fig. 1 by the words “Theory” and “Applications”.

Also the session plots offer some support to the view that theoretically oriented (or organized) discussions appear more on the upper left area, whereas application-oriented articles (or those theoretical articles that were drawn to the clusters due to related application area) can be found more from the right side and lower area of the map. It is also evident from the session plots that this division is by no means clear-cut.

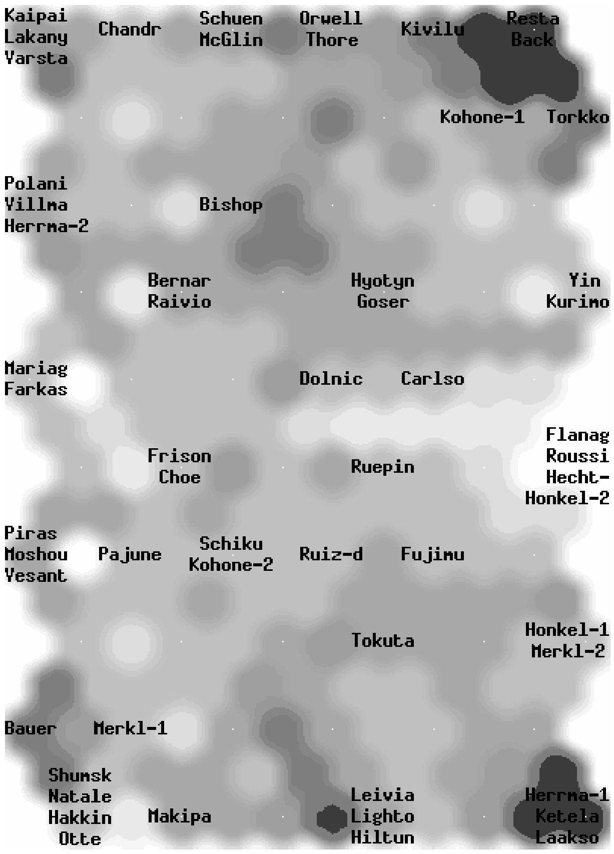

Figure 2: The WSOM'97 abstract map labeled with a prefix of the first author's last name, and with the page number of the article in this proceedings. Pleease note that not all abstracts seem to fall in their "optimal" cluster. This may be partially due to differences in word choices and also because of the high degree of relations between the articles — not all relations could be visualized at the same time.

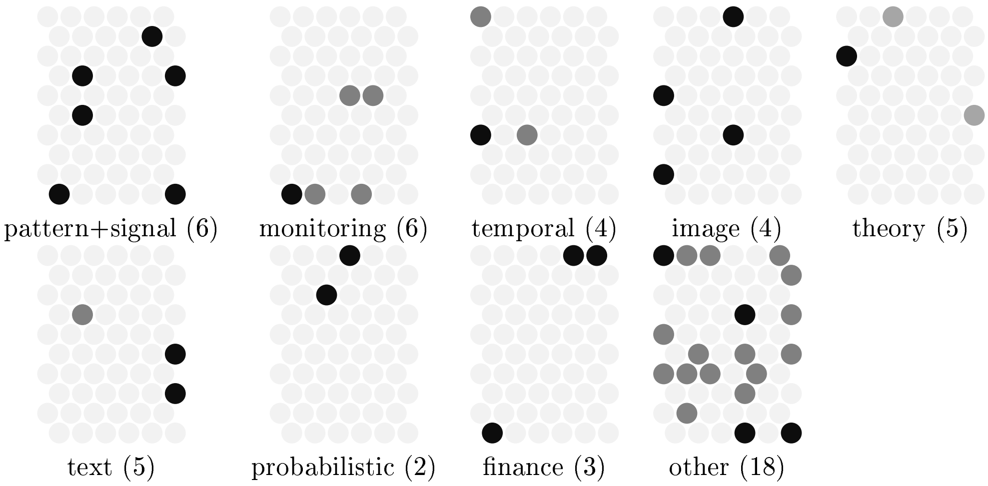

Figure 3: The spread of articles from each session was lotted on the abstract map. The category "other" includes plenaries, invited papers and posters, none of which belonged to sessions. A black dot indicates two or more papers in a unit, a grey dot implies one paper.

4 Conclusions and Discussion

After one has explored the map thoroughly it becomes evident that there are indeed some rather clean units or clusters, in which all or most of the abstracts clearly have a "common denominator". Still, in many nodes no single common aspect can be found. It seems that in several nodes different pairs of articles have different issues in common.

The organization on different areas of the map seems to locally visualize different dimensions of the input space. Thus, the search for a few clear overall dimensions is in vain, especially when the actual input dimension is large even locally. This phenomenon was to be expected, since a scientific article seldom discusses only a single or a few issues. Nevertheless, the map visualizes some of the structures of the abstract collection in a useful manner, and uncovers relationships that were not apparent from the workshop program.

There are many possible and useful views to any collection of documents, some easier to uncover automatically than others. Often at any given time we may be mainly interested in one particular view, such as methodological organization, or organization based on application area. Designining methods for selecting and weighting the document features according to the desired view or emphasis would thus be an important research issue for improving the visualization of document collections.

Acknowledgements

I would like to thank Samuel Kaski, Teuvo Kohonen, Timo Honkela and Janne Sinkkonen for their helpful comments regarding this article.

References

[1] T. Honkela, S. Kaski, K. Lagus, and T. Kohonen. Websom — self-organizing maps of document collections. In *Proc. of Workshop on Self-Organizing Maps 1997 (WSOM'97)*, Espoo, Finland, June 1997.

[2] S. Kaski, T. Honkela, K. Lagus, and T. Kohonen. Creating an order in digital libraries with self-organizing maps. In *Proc of World Congress on Neural Networks (WCNN-96)*, 1996.

[3] T. Kohonen. Exploration of very large databases by self-organizing maps. In *Proceedings of International Conference on Neural Networks (ICNN'97)*, Houston, Texas, USA, 1997. to appear.

[4] T. Kohonen, S. Kaski, K. Lagus, and T. Honkela. Very large two-level som for the browsing of newsgroups. In *Proc. of International Conference on Artificial Neural Networks*, 1996.

[5] K. Lagus, T. Honkela, S. Kaski, and T. Kohonen. Self-organizing maps of document collections: A new approach to interactive exploration. In *Proc. of the 2nd Int. Conf. on Knowledge Discovery and Data Mining (KDD-96)*, pages 238–243, Menlo Park, California, 1996. AAAI Press.