ICANN'96, Bochum, Germany, July 16-19, 1996.
Lecture Notes in Computer Science, vol. 1112,
pages 269-274. Springer, Berlin.

Very Large Two-Level SOM for the Browsing of Newsgroups

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

Helsinki University of Technology
Neural Networks Research Centre
Rakentajanaukio 2 C, FIN-02150 Espoo, Finland

Abstract. On January 19, 1996 we published in the Internet a demo of how to use Self-Organizing Maps (SOMs) for the organization of large collections of full-text files. Later we added other newsgroups to the demo. It can be found at the address http://websom.hut.fi/websom/. In the present paper we describe the main features of this system, called the WEBSOM, as well as some newer developments of it.

1 Introduction

When organizing large collections of free-form full-text document files that contain no keywords, e.g. the newsgroups in the Internet, it is difficult to base their analysis on traditional search expressions. The main information one can resort to in the classification of such documents is statistical.

SOMs of document collections have previously been constructed on the basis of their word histograms (published works are [5], [6], [7], [10], [11], [12]). Thereby, however, the size of the selected vocabulary cannot be large.

In other studies (cf., e.g., [1], [2], [8], [9], [11], and several others) it has also been found that short segments of text, such as triples of successive words, and in particular their statistical frequencies can effectively distinguish words according to their semantic roles. Self-organized maps of meaningful clusters of words are then formed.

In this work, in order to encode a document, we first formed a histogram of the *clusters* of its words on a SOM of the above type. Such histograms of different documents were then organized by a second SOM, which created another clustered display, namely, the document map. The various nodes in this second SOM can be seen to contain closely related documents, such as discussions on the same topics, answers to the same questions, calls for papers, publications of software, related problems (such as financial applications, ANNs and the brain), etc.

The first map contained 315 neurons with 270 inputs each. The second map had 49 152 neurons with 315 inputs each. The number of documents used for training and being mapped in this experiment was 131 500.

When provided with suitable means for communication, our system (dubbed the *WEBSOM*), can also be used as a kind of "agent" for the automatic searching of documents.

2 Detailed Description

2.1 Preprocessing of Text

First we eliminated some non-textual information (e.g., ASCII drawings and automatically included signatures) from the newsgroup articles. Numerical expressions and special codes were replaced by special symbols using heuristic rules.

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

In order to emphasize the subject of an article and to reduce erratic variations due to different discussion styles, a number of common words was discarded from the vocabulary. There were 31 000 000 words processed. The size of the vocabulary, after discarding the rare words, was 22 000, from which 3 500 common words were still removed manually.

2.2 Formation of the Word Category Map

The first SOM, with 270 inputs and 315 map units, was formed and labeled using the whole text material as training data. Each word of the vocabulary was represented by a random code. Each "code," relating to word position i, was a random vector $x_i \in \Re^{90}$, every component of which was drawn from a uniform scalar distribution. The encoded words were concatenated into a single string of word symbols.

For each different (remaining) word in the corpora we then computed its *averaged statistical feature vector*

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

where i can now be any position in the string where the same code x_i of this word is found, and ε is a small numerical constant, e.g. equal to 0.2. These feature vectors were applied as inputs to the first SOM.

The nodes of the SOM were labeled by inputting the feature vectors once again and finding the winner node for each. A node was thus labeled by all the words the corresponding feature vector of which selected this node for a winner.

2.3 Formation of the Histograms

In the encoding of documents, the text of each document separately was preprocessed as described in Sec. 2.1. When the encoded string of its words was scanned, the occurrence of each word was counted and recorded at that node of the first SOM which was labeled according to this word.

If the documents belong to different groups, such as the newsgroups in the Internet, the counts can be further *weighted* by the information-theoretic *entropies* (Shannon entropies) of the words, defined in the following way. Denote by $n_g(w)$ the frequency of occurrence of word w in group g ($g = 1, \ldots, 20$), and

by $P_g(w)$ the probability that the word w belongs to group g. The entropy H of this word is defined as:

$$H(w) = -\sum_g P_g(w) \log P_g(w) \approx -\sum_g \frac{n_g(w)}{\sum_{g'} n_{g'}(w)} \log \frac{n_g(w)}{\sum_{g'} n_{g'}(w)} , \quad (2)$$

and the weight $W(w)$ of word w is defined as

$$W(w) = H_{\max} - H(w) , \quad H_{\max} = \log 20 . \quad (3)$$

2.4 Formation of the Document Map

Before using the *histograms* obtained in Sec. 2.3 as inputs to the second SOM, the *document map*, they were further *blurred* using a convolution with a symmetric Gaussian kernel, the full width at half maximum of which was two lattice units. The blurring increases invariance in classification. This map, with 315 inputs and 49 152 map units, was then computed as explained in Sec. 4.

2.5 Practical Computation of Large Maps

With large maps, both winner search and updating (especially of large neighborhoods in the beginning) are time-consuming tasks. With a parallel SIMD computer, such as the 512-processor neurocomputer CNAPS at our disposal, this can be made fairly rapidly, in a few dozens of minutes.

The local-memory capacity of the CNAPS, however, has so far restricted our computations to 315-input, 768-neuron SOMs. Recently [4] we have been able to multiply the sizes of the SOMs by two solutions: 1. Good initial values for a much larger map can be *estimated* on the basis of the asymptotic values of a smaller map, like the one computed with the CNAPS, by a local interpolation procedure. There is room for a much larger map in a general-purpose computer, and the number of steps needed for its fine tuning is quite tolerable. 2. In order to accelerate computations, the winner search can be speeded up by storing with each training sample an address pointer to the old winner location. During the next updating cycle, the approximate location of the winner can be found directly with the pointer, and only a *local search* around it needs to be performed. The pointer is then updated. In order to guarantee that the asymptotic state is not affected by this approximation, updating with a full winner search was performed intermittently, after every 30 training cycles.

3 Browsing Interface

The document space is presented at three basic levels of the system hierarchy: the map, the nodes, and the individual documents (Fig. 1). Any subarea of the map can be selected and zoomed by "clicking." One may explore the collection by following the links from hierarchy level to another. It is also possible to move to neighboring areas of the map, or to neighbors at the node level directly. This hierarchical system has been implemented as a set of WWW pages. They can be explored using any standard graphical browsing tool. A complete demo is accessible in the Internet at the address http://websom.hut.fi/websom/.

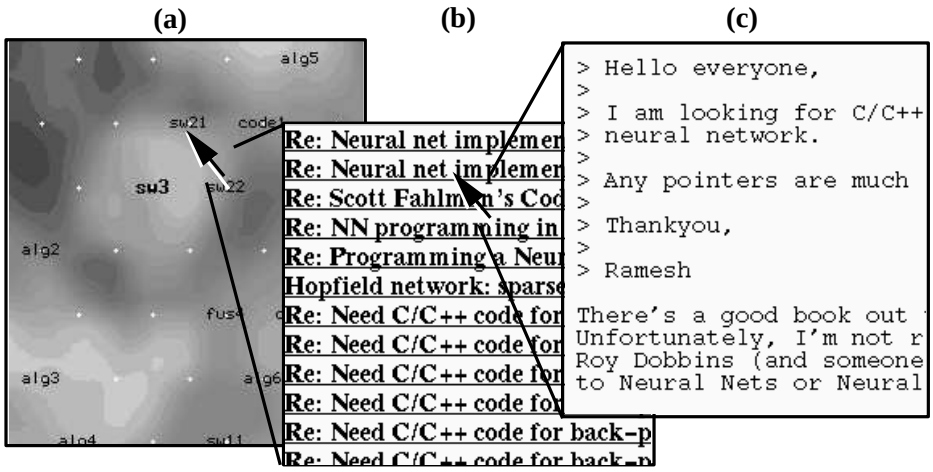

Fig. 1. Sample scenes of the WEBSOM interface. (a) Part of a zoomed document map display. The clustering tendency is visualised using the gray scale. (b) The map node contents. (c) An example of a newsgroup article picked up from the node.

4 Experiments

The word category map was computed with the CNAPS and fine-tuned on a general-purpose computer. The document map contained initially 768 units and it was computed with the CNAPS as discussed in Sec. 2.5 using 131 500 documents for training. It was then enlarged into 49 152 units by interpolation. Finally the whole document material was used to fine-tune the map on a general-purpose computer. The resulting document map is presented in Fig. 2, with separate images displaying the distribution of each group on the same map. The separation of the groups is presented by the confusion matrix in Table 1.

The actual distribution figures were too large to be included here. If sufficiently interested, ask for a paper copy of the original article from the main author.

Fig. 2. Distributions of documents in different newsgroups on the WEBSOM of size $192 \times 256 = 49\,152$ units. Each small display only contains articles from a single newsgroup. The shade of the dots indicates the number of articles the unit contains: the darker the dot, the larger the number.

	1)	2)	3)	4)	5)	6)	7)	8)	9)	10)	11)	12)	13)	14)	15)	16)	17)	18)	19)	20)
1)	4157	48	143	135	72	288	151	23	52	148	167	165	114	29	225	178	194	126	99	189
2)	141	547	30	89	26	276	120	9	8	41	34	93	37	6	91	63	94	48	33	82
3)	263	23	1137	53	14	115	42	9	8	28	17	51	24	7	81	44	43	43	27	62
4)	345	100	62	770	33	115	103	10	25	102	125	71	97	42	119	60	84	79	40	103
5)	128	20	41	26	790	961	248	40	14	29	50	99	50	7	126	97	129	63	33	83
6)	234	84	79	56	277	5935	1095	186	40	110	137	323	162	21	274	209	291	158	94	224
7)	123	63	56	58	122	1389	3988	420	44	123	169	282	147	23	290	170	319	134	70	243
8)	49	23	37	18	53	597	846	660	20	43	64	95	57	13	110	68	134	66	31	95
9)	85	20	7	23	9	58	65	20	1751	570	547	196	238	99	184	182	137	33	31	200
10)	164	38	16	52	16	136	116	31	368	5688	931	350	464	173	291	321	289	72	66	407
11)	114	20	20	42	29	143	136	33	301	910	5632	390	606	188	353	316	304	81	52	327
12)	203	77	41	44	69	501	351	59	164	399	456	4967	394	91	484	383	590	108	106	499
13)	120	24	26	74	28	160	145	30	127	434	630	354	5535	293	593	705	254	63	38	352
14)	48	9	5	17	11	34	28	10	81	242	306	137	474	1608	234	205	95	23	23	143
15)	170	33	41	75	57	318	241	37	99	267	361	387	572	150	5736	597	286	68	75	416
16)	145	43	15	38	39	174	147	35	111	290	352	385	668	139	759	5833	250	55	57	448
17)	172	49	29	55	51	397	260	45	89	239	248	398	167	41	300	252	6578	142	126	355
18)	155	25	48	62	30	285	172	26	17	66	78	152	72	16	122	82	181	1414	72	115
19)	211	24	44	44	30	159	69	15	26	68	73	159	62	7	119	121	198	60	1145	129
20)	152	30	35	49	31	180	173	40	118	320	383	371	292	76	429	436	427	91	70	6263

Table 1. This confusion matrix indicates how the articles from a newsgroup are distributed to the map units dominated by the various groups (numbered as in Fig. 2). Each map unit was labeled according to the most frequent group in that unit. Each row describes into which groups the articles were distributed. Some similar groups like the philosophy groups (6, 7, and 8) and the movie-related groups (9, 10, and 11) contain similar discussions and are easily mixed.

References

1. S. Finch and N. Chater, "Unsupervised methods for finding linguistic categories," in *Proc. of ICANN-92*, vol. 2, pp. 1365-1368, North-Holland, 1992.
2. T. Honkela, V. Pulkki, and T. Kohonen, "Contextual relations of words in Grimm tales analyzed by self-organizing map," in *Proc. of ICANN-95*, vol. 2, pp. 3-7, EC2 et Cie, 1995.
3. T. Kohonen, *Self-Organizing Maps*. Berlin, Heidelberg: Springer, 1995.
4. T. Kohonen, *Speedy SOM*, Report A33, Laboratory of Computer and Information Science, Helsinki University of Technology, 1996.
5. X. Lin, D. Soergel, and G. Marchionini, "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, 1991.
6. D. Merkl, A. M. Tjoa, and G. Kappel, "A Self-Organizing Map that Learns the Semantic Similarity of Reusable Software Components," in *Proc. ACNN'94, 5th Australian Conference on Neural Networks*, Brisbane, Australia, pp. 13–16, 1994.
7. D. Merkl, "Content-Based Document Classification with Highly Compressed Input Data," in *Proc. of ICANN-95*, vol. 2, pp. 239-244, EC2 et Cie, 1995.

8. R. Miikkulainen, *Subsymbolic Natural Language Processing: An Integrated Model of Scripts, Lexicon, and Memory*. Cambridge, MA: MIT Press, 1993.
9. H. Ritter and T. Kohonen, "Self-organizing semantic maps," *Biol. Cyb.*, vol. 61, no. 4, pp. 241–254, 1989.
10. J. C. Scholtes, "Unsupervised learning and the information retrieval problem," in *Proc. IJCNN'91*, pp. 18–21, IEEE Service Center, 1991.
11. J. C. Scholtes, *Neural Networks in Natural Language Processing and Information Retrieval*. PhD thesis, University of Amsterdam, Amsterdam, Netherlands, 1993.
12. J. Zavrel, *Neural Information Retrieval*, MA Thesis, University of Amsterdam, Amsterdam, Netherlands, 1995.

This article was processed using the LaTeX macro package with LLNCS style