COMPARISON OF NEURAL NETWORKS USED FOR PROCESSING AND CATEGORIZATION OF CZECH WRITTEN DOCUMENTS

Pavel Mautner and Roman Mouček
Department of Computer Science and Engineering, University of West Bohemia in Pilsen
Univerzitní 8, 30614 Pilsen, Czech Republic

Keywords: Document categorization, WEBSOM, ART-2.

Abstract: The Kohonen Self-organizing Feature Map (SOM) has been developed for the clustering of input vectors and for projection of continuous high-dimensional signal to discrete low-dimensional space. The application area, where the map can be also used, is the processing of collections of text documents. The basic principles of the WEBSOM method, a transformation of text information into a real components feature vector and results of documents classification are described in the article. The Carpenter-Grossberg ART-2 neural network, usually used for adaptive vector clustering, was also tested as a document categorization tool. The results achieved by using this network are also presented here.

1 INTRODUCTION

Today a huge collection of documents is accessible in electronic libraries or on the Internet. Finding relevant information in this collection of documents is often difficult and time consuming task. Efficient search tools such as search engines have quickly emerged to aid in this endeavor.

To make searching faster, the categorization of documents according their content has become a widely used method. Based on the keywords included in the query it is possible to estimate the query class (or domain) and then to make the search space narrower. It reduces either searching time or the length of the list of references.

In the past, many document categorization methods have been developed (Manning et al., 2007). One of the interesting approaches to a document representation and consequential classification was presented by Finish scientists in (Kaski et al., 1998). Their approach is called WEBSOM and it is based on Kohonen self-organizing feature map (Kohonen, 2001). The method was developed for automatic processing and categorization of English (or Finish) written internet documents and consecutive information retrieval in these documents.

This paper deals with the application of the WEBSOM method for Czech written document categorization and its modification, in which an ART-2 neural network is used as a document categorizer. The paper is organized as follows. Section 2 provides basic information about architecture and features of neural networks used for document processing and categorization, Section 3 describes principles of document representation by a feature vector, word category construction and documents categorization. The results of experiments and possible future extension of this work are summarized in Section 4.

2 SYSTEM ARCHITECTURE

2.1 Basic WEBSOM Architecture

The WEBSOM method is based on a two layer neural network architecture (see Figure 1). The first layer of the WEBSOM, the Word Category Map (WCM), processes an input feature vector. The second layer, the Document Map (DM), categorizes input documents according to information from WCM output. Both layers of the WEBSOM are based on Kohonen self-organizing feature map (SOM).

The SOM is an artificial neural network developed by Theuvo Kohonen. It has been described in several research papers and books (Kohonen, 2001), (Fiesler and Beale, 1997), (Fausett, 1994). Its purpose is to map a continuous high-dimensional space into a discrete space of lower dimension (usually one or two dimensional space). The map contains one layer of neu-

Mautner P. and Mouček R..
COMPARISON OF NEURAL NETWORKS USED FOR PROCESSING AND CATEGORIZATION OF CZECH WRITTEN DOCUMENTS.
DOI: 10.5220/0003116205100513
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2010), pages 510-513
ISBN: 978-989-8425-28-7

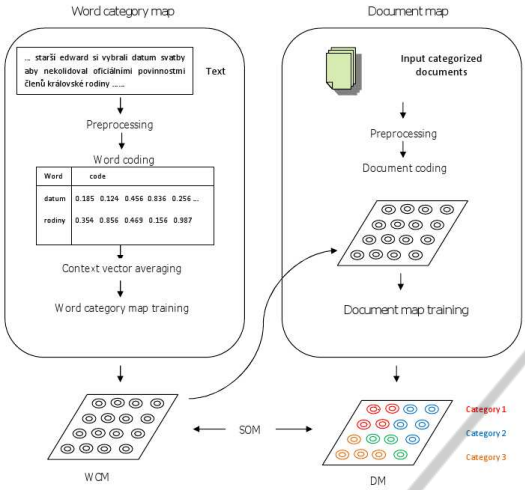

Figure 1: Basic architecture of the WEBSOM.

rons, ordered to two-dimensional grid, and two layers of connections. In the first layer of connections, each neuron is fully connected (through weights) to all feature vector components. Computations are feedforward in the first layer of connections: the network computes a distance between the feature vector F_v and each of the neuron weight vectors w_j by the following formula:

$$d_j(t) = \sum_{i=0}^{N-1} (F_{v_i}(t) - w_{ij}(t))^2, \qquad j = 1,2,\ldots,M, \quad (1)$$

where t is the time point, in which the output is observed, $F_{v_i}(t)$ are components of feature vector and $w_{i,j}(t)$ are components of neuron weight vector, N is the number of feature vector components, and M is the number of neurons (and also WCM units).

The second layer of connections acts as a recurrent excitatory/inhibitory network, whose aim is to realize winer-takes-all strategy, i.e. only the neuron with the highest activation level $d_j(t)$ is selected and signed as the best matching unit (BMU).

The document categorization by the WEBSOM method proceeds by the following manner. At first, an input document is parsed and each word is preprocessed and translated into a feature vector (see the following section). The feature vector is clustered by the WCM and a BMU value of the feature vector is saved into the WCM output vector. After processing all the words from the input document, the WCM output vector is presented to the input of the document map (DM). The document map processes the WCM output vector and activates one of the output units (BMU of the document map) which corresponds to the category of the input document. It can be shown (Kaski et al., 1998), that similar documents activate similar DM output units.

2.2 Document Categorization using ART Neural Network

In subsection 2.1 the document categorization system based on Kohonen map was described. In that system the document map creates clusters of similar documents, which have to be calibrated after the training process. Within the calibration process, the output units of the document map are labeled according to the input documents categories, for which they have become the BMUs. The labeling process can be complicated because there are not clear borders between document clusters.

To simplify this problem another neural network, with simple outputs, which correspond to the document categories accurately, was used. Since the document separation based on topic similarity is often required, the ART (Adaptive Resonance Theory) network was selected as a good candidate for document categorization. The ART network developed by Carpenter and Grossberg (Carpenter and Grossberg, 1988) is also based on clustering, but its output provides direct information about output class (document category). There are several ARTs (ART-1, ART-2, ARTMAP) differing by their architecture and input feature vector type. For our work, the ART-2 network, processing real-valued feature vector was used. For detailed description of ART network see (Fausett, 1994) or (Carpenter and Grossberg, 1988).

The modified architecture of a document categorization system using ART-2 network is illustrated in Figure 2.

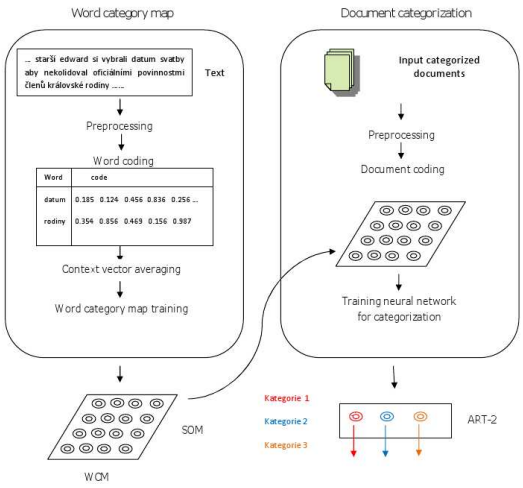

Figure 2: Modified architecture of system using ART-2 network for document categorization.

3 DOCUMENT REPRESENTATION

In section 2 the system architecture for document categorization was presented. With respect to the fact that input layer of the document processing system uses the self-organizing map, which processes a real-valued input vector, it is essential to transform an input text to its numerical representation.

In (Kaski et al., 1998) the representation of documents by the averaged context vectors was presented. The averaged context vectors are generated upon the context of the words in the document collection by the following process:

1. Each word s_i in the vocabulary, which was created for given document corpus, is evaluated by a unique random real vector w_i of dimension n.
2. The input document corpus is searched, and all occurrences of word s_i are found.
3. The context of the word s_i is found, i.e. m words, which precede/follow the word s_i are taken from each document containing this word, and the vectors pw_i (average of all vectors of m-tuple of w_i preceding the word s_i) and nw_i (average of all vectors of m-tuple of w_i following the word s_i) are evaluated.
4. The average context vector cw_i of the word s_i is created from values pw_i, w_i, nw_i by the following way:

$$cw_i = \begin{pmatrix} pw_i \\ \varepsilon w_i \\ cw_i \end{pmatrix}, \qquad (2)$$

where ε is a weight of vector representing the word s_i.

It is evident that the words occurring in the similar context have a similar context vector and they belong to the same category. Based on this assumption, it is possible to train the word category map.

4 RESULTS AND FUTURE WORK

All neural network based systems for document categorization described in this paper were implemented in Java. They can be downloaded and used for non-commercial purpose.

The systems were tested on corpus of 6000 documents containing Czech Press Agency news. The whole corpus has included approximately 146 000 words, stop and insignificant words were removed from the corpus. The documents were categorized by hand into four categories, then the results were compared with the results of automatic categorization. Distribution of documents into categories was the following:

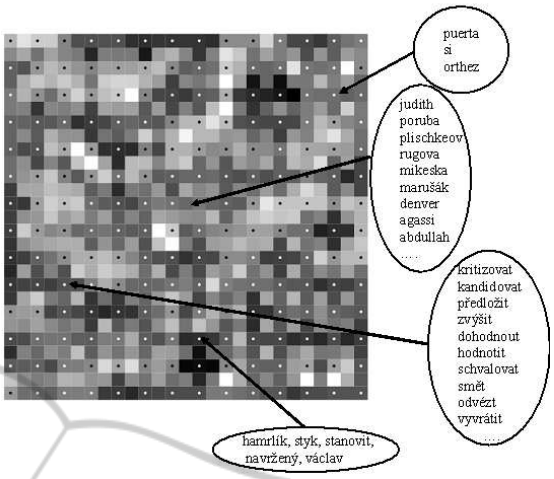

Figure 3: Trained Word Category Map.

document category	% of all documents
sport	44
policy	51
foreign actuality	3
society	2

With regard to the low numbers of documents representing some categories (e.g. there were approximately 80 documents about society and 200 documents dealing with foreign actuality in the corpus), a set of 160 documents (40 documents from each category) was selected for training the word category map and neural-based categorizers. A vocabulary of words generated from the training set of documents was created and all words with frequency of occurrence smaller than predefined threshold were removed from the vocabulary. Then the vocabulary was used for training the WCM.

The size of the WCM (the first layer of the classification system) was chosen in order to place approximately 25 words into each category (i.e. the map contains approximately 40 neurons for 1000 words). The word category map was trained by numeric vectors representing the words in the dictionary.

The result of the training of the WCM and an example of word categories are illustrated in Figure 3. It is apparent that some output units respond to the words only from a specific syntactic or semantic category (nouns, first name and surname etc.), while other units respond to the words from various syntactic or semantic categories.

The Document Map consists of nine neurons ar-

Table 1: Results of document categorization using Document Map.

DM unit number	Number of documents (in %) for category:			
	Sport	Policy	Foreign Actuality	Society
1	2.5	16.2	8.5	0
2	45	9.4	16.5	20
3	20	2.4	0	0
4	22.5	25.5	25	0
5	0	0	0	0
6	10	25.5	41.5	60
7	0	0	0	0
8	0	0	0	0
9	0	21	8.5	20

Table 2: Results of document categorization using ART-2 categorizer.

ART-2 output unit number	Number of documents (in %) for category:			
	Sport	Policy	Foreign Actuality	Society
1	8.4	11.4	53.7	17.7
2	0.4	2.3	1.4	2.4
3	0.1	0	0	0
4	14.7	5.6	0.5	8.9
5	5.8	58.3	10.4	44.4
6	0.2	0.1	0.5	0
7	56.3	14.7	16.3	13.3
8	5.7	2.9	4.1	4.4
9	8.4	4.7	13.1	8.9

ranged to 3x3 grid. The map receives and processes the vectors from the WCM output convolved by Gaussian mask and produces the output which corresponds to the category of the input document. After the training, the DM output units were labeled manually.

The association of documents from particular categories to the clusters, which are represented by the DM map output units are presented in Table 1. It is evident that the unit 2 is mostly activated for the sport category, units 4 and 6 are activated especially for category policy, etc.

The ART-2 network was developed to give a comparable output with the SOM based categorizer. The ART-2 categorizer has nine output units (i.e. the network can create at most nine clusters). The set of documents used for training of the SOM based categorizer was also used here. The number of actually created clusters was strongly dependent on the parameter ρ (vigilance threshold). In our case parameter $\rho = 0.98$ was used because most documents were submitted to only one cluster if ρ had a smaller value. The results of categorization using ART-2 categorizer are presented in Table 2. The meaning of values in the table is similar as for the SOM based categorizer. Documents with sport, policy and foreign actuality topics are well separated (see the values for units 7, 5 and 1 respectively), documents dealing with society news were mostly submitted to the same cluster as documents about policy (output unit 5).

The comparison of SOM and ART-2 based categorizers is quite difficult and it is still investigated. Since the changes in the SOM network parameters affect the resulting clusters less than it is in the case of ART-2 network, the results seem to be more natural. The advantage of SOM categorizer is a low number of parameters. The ART-2 is very sensitive to parameters setting. There are seven parameters of the network (including ρ mentioned above), which have to be set up before training the network. If the parameters are chosen properly, the network can give better categorization results then SOM categorizer.

In our future work we plan to focus on the following tasks, which could improve the results of document categorization:

- introduction of another feature set for word description,
- application of other supervise-trained neural networks (e.g. multilayer perceptron, LVQ, etc.) as a second layer
- usage of more sophisticated approaches for comparison of categorization results

ACKNOWLEDGEMENTS

This work was supported by grant no. 2C06009 CotSewing.

REFERENCES

Carpenter, G. A. and Grossberg, S. (1988). The art of adaptive pattern recognition by a self-organizing neural network. *Computer*, 21(3):77–88.

Fausett, L. V. (1994). *Fundamentals of Neural Networks*. Prentice Hall, Englewood Cliffs, NJ.

Fiesler, E. and Beale, R., editors (1997). *Handbook of Neural Computation*. Oxford University Press.

Kaski, S., Honkela, T., Lagus, K., and Kohonen, T. (1998). Websom-self-oganizing maps of document collections. *Neurocomputer*, pages 101–117.

Kohonen, T. (2001). *Self-Organizing Map*. Springer-Verlag, Berlin Heidelberg.

Manning, C. D., Raghavan, P., and Schütze, H. (2007). *An Introduction to Information Retrieval - Preliminary Draft*. Cambridge University Press.