PROCESSING AND CATEGORIZATION OF CZECH WRITTEN DOCUMENTS USING NEURAL NETWORKS

Pavel Mautner, Roman Mouček[*]

Abstract: The Kohonen Self-organizing Feature Map (SOM) has been developed for clustering 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 text documents. Within the project WEBSOM, some methods based on SOM have been developed. These methods are suitable either for text documents information retrieval or for organization of large document collections. All methods have been tested on collections of English and Finnish written documents. This article deals with the application of WEBSOM methods to Czech written documents collections. The basic principles of WEBSOM methods, transformation of text information into the real components feature vector and results of documents classification are described. 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.

Key words: *Document categorization, WEBSOM, SOM, ART-2, neural networks, document semantics*

Received: March 26, 2010
Revised and accepted: February 8, 2012

1. Introduction

Today huge collections of documents are affordable in electronic libraries or on the Internet. Finding relevant information in these collections of documents is often a difficult and time consuming task. Efficient search tools such as search engines have quickly emerged to aid in this endeavor. Traditional search methods are based on asking a suitable query (e.g. query based on keywords from a searched domain) and following matching document contents with keywords included in the query.

[*]Pavel Mautner – corresponding author, Roman Mouček,
Department of Computer Science and Engineering, University of West Bohemia in Plzeň, E-mail: `mautner@kiv.zcu.cz`, `moucek@kiv.zcu.cz`

Since a free word order is possible in natural inquiries it can happen that searching engine produces a long list of irrelevant citations.

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

2. State of the Art

In the past many document categorization methods have been developed. Apte [1] used optimized rule-based induction methods which automatically discovered classification patterns that can be used for general document categorization. Kwok [18] and Joachims [8] applied Support Vector Machines (SVM) technique which allowed users easy incorporation of new documents into an existing trained system. Yang and Chute [17] used a training set of manually categorized documents to learn word-category associations, and used these associations to predict categories of arbitrary documents. A Linear Least Squares Fit technique is employed to estimate the likehood of these associations. Lai and Lam [12] developed a similarity based textual document categorization method called the generalized instance set (GIS) which integrates advantages of linear classifiers and k-nearest neighbor algorithm by generalization of selected instances. Manninen and Pirkola [14] used a self-organizing map to classify textual documents (emails) to categories. They created a method for automatic classification of abstract, open-ended, and thematically overlapping emails in which the boundaries of different classes may be relatively vague. Merkl and Rauber [16] compares two models of self-organizing neural networks (Adaptive Resonance Theory and self-organizing map) which are used to content-based classification of textual documents. Lagus et al. [9], [11] developed a method called WEBSOM which utilizes a self-organizing map algorithm for organizing collections of text documents into a visual document map. The map was also tested in Semantic Web environment [5]. Dittenbach et al. [13] employed the Growing Hierarchical Self-Organizing Map for hierarchical classification of documents from CIA World Factbook.

Most articles cited above deal with categorization of English written documents. Honkela et al. [6] used the WEBSOM method for creating maps of multilingual document collections, but only English and Finnish documents are used in their test. Therefore, we decided to apply a similar principle to categorization of Czech written documents to determine if it is possible to use self-organizing neural networks for categorization of documents with a different grammar structure than the English grammar.

This paper deals with the application of WEBSOM method for Czech written document categorization and its modification in which the ART-2 neural network is used as a document categorizer. The paper is organized as follows. Sections 2 and 4 provide basic information about the architecture and features of neural networks used for document processing and categorization. Section 3 describes document representation using a feature vector, word category creation and document categorization. The results of experiments and possible future extension of this work are summarized in Section 5.

3. System Architecture for Czech Written Documents Processing and Categorization

3.1 Basic WEBSOM architecture

The WEBSOM method [7] is based on a two layer neural network architecture (see Fig. 1). The first layer of WEBSOM processes an input feature vector and creates so called Word Category Map (WCM). The WCM is a self-organizing map (SOM) which has organized words according to similarities in their role in different context. Each unit of the SOM corresponds to a word category that contains a set of similar words. The word similarity is based on their averaged context in which the word occurs in collection of input documents. Each input document is then encoded by WCM as the histogram of the corresponding word categories. This histogram is processed by the second layer, the Document Map (DM), which maps the histogram of word categories into the corresponding document class. Similar documents then activate topologically similar output units of the Document map. The document map is also formed by SOM algorithm.

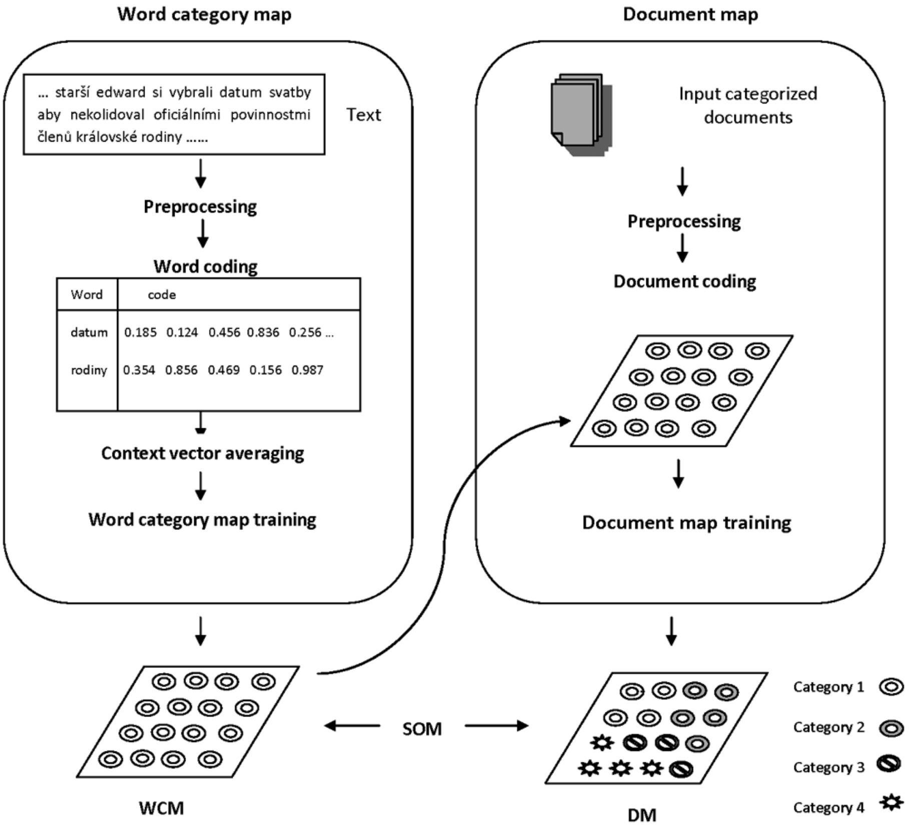

Fig. 1 *Basic architecture of WEBSOM model.*

The SOM is an artificial neural network developed by Theuvo Kohonen. It has been described in several research papers and books [10], [4], [3]. The purpose of the self-organizing feature map is basically to map a continuous high-dimensional space into a discrete space of lower dimension (usually 1 or 2). The basic structure of the SOM is illustrated in Fig. 2. The map contains one layer of neurons, ordered to two-dimensional grid, and two layers of connections. In the first layer of connections, each element is fully connected (through weights) to all feature vector components. Computations are feedforward in the first layer of connections: the network computes the distance between the input vector F_{v_i} and each of the neuron weight vectors $w_{i,j}$ by the following formula:

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

where t is the time point in which the output is observed, $F_{v_i}(t)$ are components of input vector and $w_{ij}(t)$ are components of corresponding neuron weight vector, N is the number of feature vector components, and M is the number of WCM units (in the Document map) or length of the context vector (in the Word Category Map).

The second layer of connections acts as a recurrent excitatory/inhibitory network, whose aim is to realize a winner-take-all strategy, i.e. the only neuron with highest activation level $d_j(t)$ is selected and signed as the best matching unit (BMU). The weight vector $w_{ij}(t)$ of this neuron then corresponds to the vector which is the most similar to the input feature vector F_v.

The document categorization by the WEBSOM method proceeds in the following manner. At first, an input document is parsed and particular words are preprocessed and translated into a feature vector (see Section 3). The feature vector of the input word is clustered by WCM and BMU value of the input vector is saved into WCM output vector F_{wov}. The size of this vector is the same as the number of neurons in WCM map. In case the BMU's of different input words are of the same value, the corresponding components of the WCM output vector are averaged. After the processing of all the words of the input document, the WCM

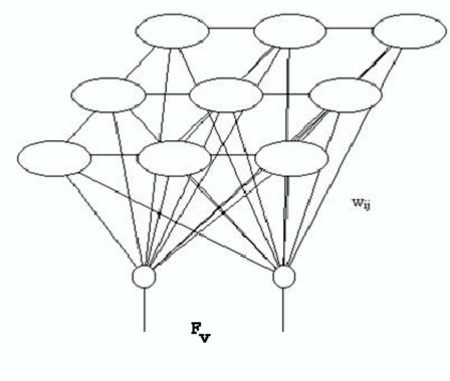

Fig. 2 *Kohonen's Self-organizing Feature Map.*

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 [9] that the similar documents activate similar output units of the DM.

3.2 Document categorization using ART neural network

In subsection 3.1, the document categorization system based on the Kohonen map was described. In that system the document map creates clusters of similar documents and has 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 become the BMUs. The labeling process can be complicated because there are no clear borders between document clusters and thereby also between document categories.

This problem can be solved using another neural network with similar properties as the Kohonen map but with simple outputs which correspond to the document categories accurately. Whereas the document separation based on topic similarity is often required, the ART network was selected as a good candidate for document categorization. The ART (Adaptive Resonance Theory) network developed by Carpenter and Grossberg [2] is also based on clustering, but its output is not a map but direct information about an output class (document category). There are several ARTs (ART-1, ART-2, ARTMAP) differing by architecture and input feature vector type (binary or real valued). For our work, the ART-2 network, processing real-valued feature vector was used. The simplified architecture of this network is illustrated in Fig. 3.

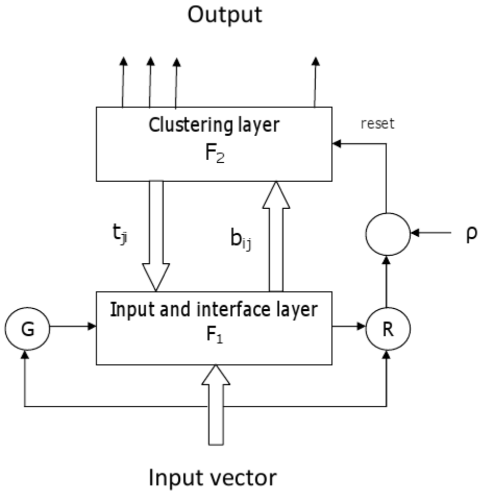

Fig. 3 *ART-2 architecture.*

The network consists of two layers of elements labeled F_1 (input and interface units) and F_2 (cluster units), each fully interconnected with the others, and supplemental unit G and R (called *gain control unit* and *reset unit*) which are used to control the processing of the input data vector and the creation of the clusters.

The input and interface layer F_1 has the same number of processing units as it is the size of the feature vector. The clustering layer F_2 consists of as many units as it is the maximum number of document categories. Interconnection of F_1 and F_2 layers is realized through the set of weight vectors labeled b_{ij} and t_{ji} saving the template of each cluster. Weight vectors can be modified according to the properties of input feature vector. For detailed description of ART network see [3] or [2]. In short, the ART-2 operation can be summarized in the following steps:

1. An input feature vector is preprocessed in the input and interface layer F_1 and it is compared with templates saved in weight vector b_{ij}. The comparison process is realized by neurons of F_2 layer as inner product of input feature vector and weight vector b_{ij}. For simplification, we will assume that k clusters (k is lower then the maximum number of clusters) were created meanwhile.

2. The neuron with the highest output is labeled as a winner and it is verified if the similarity between an input vector and the corresponding template satisfies the preadjusted criterion (vigilance threshold ρ).

3. If yes, the input vector is submitted to the cluster represented by the winner unit of F_2 and the corresponding weights b_{ij} and t_{ji} are modified.

4. If not, the winner unit of F_2 is blocked, and the process is repeated from step 2 until a neuron of F_2 satisfying preadjusted criterion is found.

5. If all k neurons of F_2 is blocked the new k+1-th cluster is created and the corresponding input vector F_v becomes its template (the weights of newly activated neuron are adapted).

The modified architecture of document categorization system using the ART-2 network is illustrated in Fig. 4.

4. Feature Vector for 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 a self-organizing map, which processes a real-valued input vector, it is essential to transform an input text to a suitable feature vector.

The vector space model [15] is a suitable method for document representation. In this method the stored documents are represented as binary vectors where the vector components correspond to the words of vocabulary. The component value is equal to 1 if the respective word is found in the document; otherwise the value is 0. Instead of binary values, real values can be used. Then each component corresponds to some function of the frequency of particular word occurrence in the document. The main problem of the vector space method is a large vocabulary in

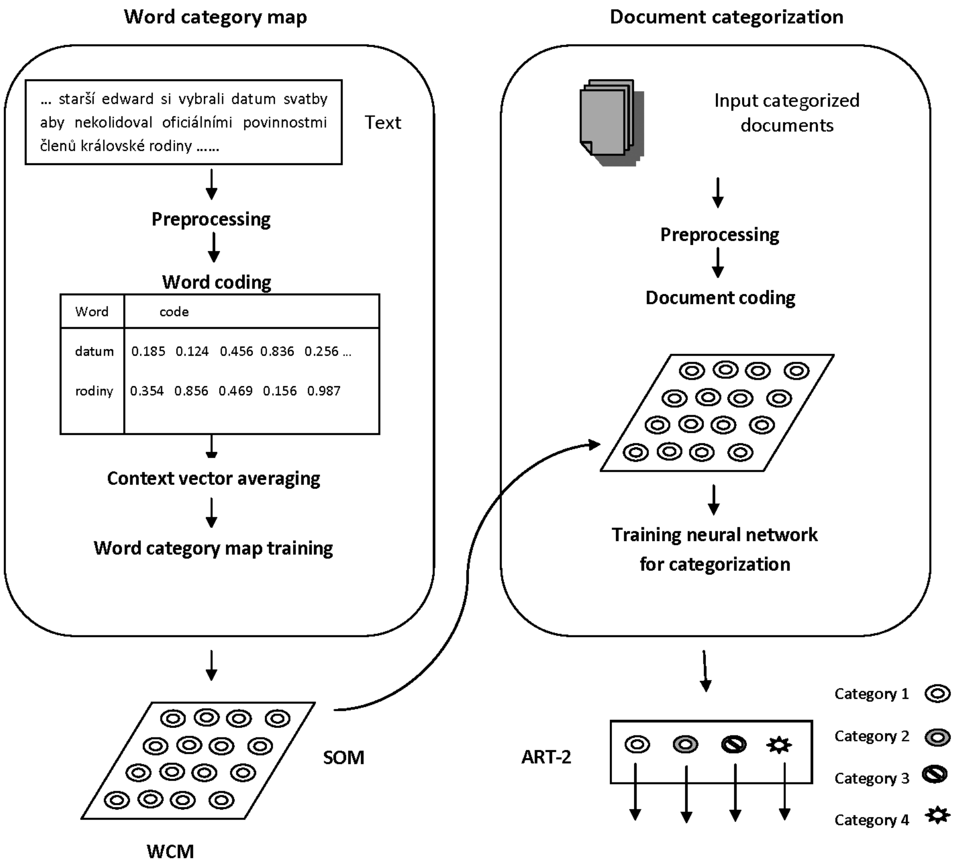

Fig. 4 *Modified architecture of system for document categorization.*

any sizable collection of free-text documents, which results in a vast dimensionality of the documents vector.

Another method of document representation is a technique called Latent Semantic Indexing (LSI) [15]. In LSI, the document-by-word matrix is analyzed using singular value decomposition (SVD) and the least significant elements of the resulting latent representation are discarded. After this, each document is represented as a linear combination of low-dimensional (typically between 100 and 200 dimensional) latent representations of the document vectors.

In [9] 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 a given document corpus, is evaluated by the 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 preceding/following 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:

$$cw_i = \begin{pmatrix} pw_i \\ \epsilon w_i \\ cw_i \end{pmatrix}, \tag{2}$$

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

In Fig. 5 the process of formation of the context vector for the word *vojáků* ($m = 1$) is illustrated.

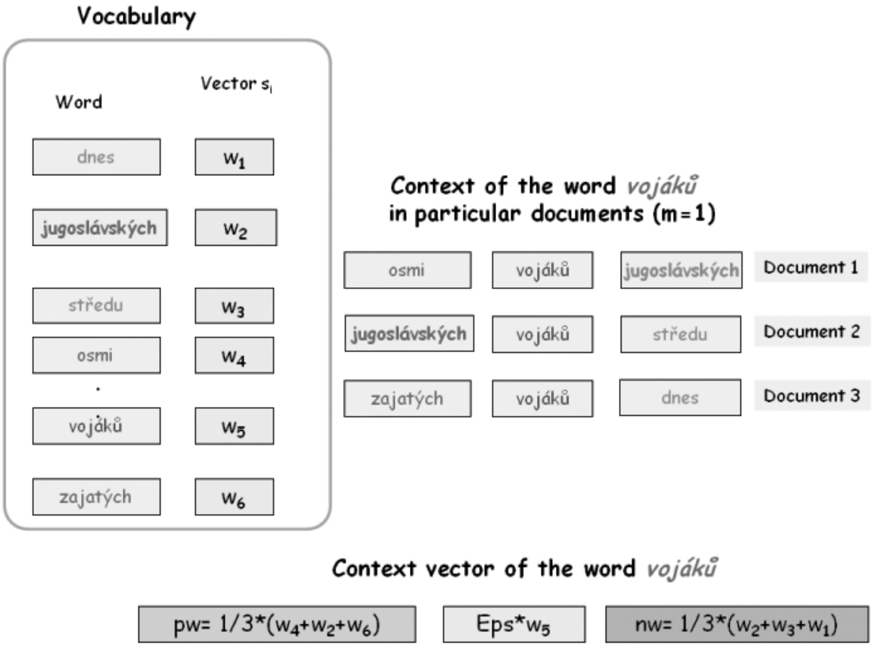

Fig. 5 *Formation of context vector.*

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

5. Training Neural Networks

5.1 Training Word Category Map (WCM)

All documents from a training set are processed to train the WCM. For each word of the document, the context vector is evaluated and it is fed to the input of the WCM. According to Eq. 1 the output of the WCM is evaluated and the winner unit is determined. The weight of the winner unit and its neighbors are updated by the following equation:

$$w_{ij}(t+1) = w_{ij}(t) + h_{ci}[F_{v_j} - w_{ij}] \qquad j = 1, 2, \ldots, M; \qquad i = 1, 2, \ldots, N, \tag{3}$$

where M is the number of WCM units, N is the number of context vector components, $w_{ij}(t+1)$ is a new weight vector component value, $w_{ij}(t)$ is an old vector component value and $h_{ci}(t)$ is a neighborhood kernel (see [10]).

The WCM is calibrated after the training process by inputing the F_{v_i} again and labeling the best-matching nodes according to symbols corresponding to the w_i parts of F_{v_i} (see Fig. 5). Each node may become labeled by several symbols, often synonymous or belonging to the same closed category, thus forming "word categories" in the nodes. Sample categories are illustrated in Fig. 6.

5.2 Training Document Map (DM)

The DM, also based on the self-organizing map, is trained by the same way as the WCM (the weights of map neurons are set up according to Eq. 3, but the WCM output vector is used for input to the DM, see Section 2).

The DM is also calibrated after the training process, now by inputting the F_{wov} acquired for each document of the training set. Each node of the DM may be now labeled by the topic of the document, for which the node is activated.

5.3 Training ART-2 Categorizer

Foreasmuch as the ART-2 is also trained without a teacher, the training of the network is similar to the training of the document map, i.e. the smoothed WCM output vector is presented to the input and interface layer of the ART-2 and after the alignment of the vector to the output cluster (see Section 3.2), the corresponding weights of the winning unit $\mathbf{J}$ are updated by the following equations:

$$\begin{aligned} t_{J,i}(n+1) &= \alpha d u_i + [1 + \alpha d(d-1)] t_{J,i}(n) & (4)\\ b_{i,J}(n+1) &= \alpha d u_i + [1 + \alpha d(d-1)] b_{i,J}(n) & (5) \end{aligned}$$

where d and α are ART-2 input parameters, u_i is the output of the interface sublayer and $\mathbf{b}(t+1), \mathbf{t}(n+1]$ and $\mathbf{b}(n), \mathbf{t}(n)$ are new and old weights vectors respectively (see [3] for details).

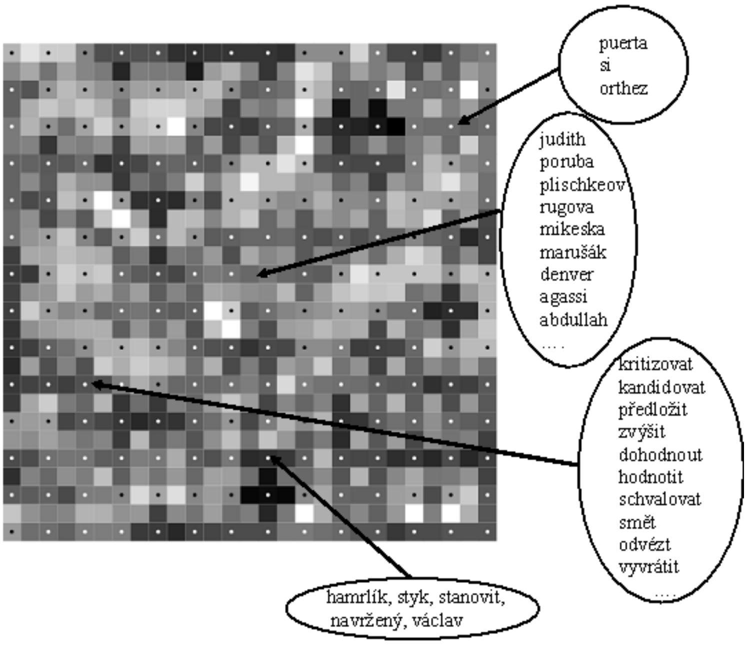

Fig. 6 *Trained Word Category Map.*

6. Results and Future Work

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

The systems were tested on the corpus of 6 000 documents containing Czech Press Agency news. The whole corpus has included approximately 146 000 words, the stop and insignificant words were removed from the corpus. These documents were categorized by hand into 4 categories so that 3 experts independently classified input documents into the categories and the resulting document category was selected by the voting rule. Then these results were compared with the results of automatic categorization. The distribution of the documents into categories was as follows:

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

With regard to the low number of documents representing some categories (e.g. there were approximately 80 documents about society and 200 documents dealing with foreign actuality available in the corpus), a set of only 160 documents (40 documents from each category) was selected for training the word category map and neural-based categorizers. The 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 1 000 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 Fig. 6. 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.

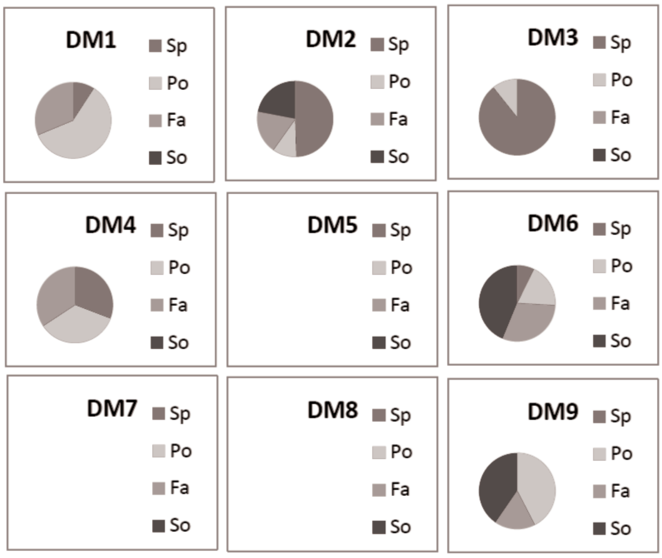

Fig. 7 *Distribution of document categories (Sport -* **Sp***, Policy -* **Po***, Foreign Actuality -* **Fa***, Society -* **So***) into trained Document map units (see Tab. I).*

The Document Map consists of 9 neurons arranged into 3x3 grid. The map receives and processes the vectors from the output of WCM convolved by Gaussian mask and produces the output which corresponds to the category of the given input document. After the training, the output units of DM were labeled manually by document categories.

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

Tab. I *Results of document categorization using Document Map.*

The association of documents from particular categories to the clusters, which are represented by the DM output units, are presented in Tab. I. It is evident that unit 2 is mostly activated for the sport category, units 4 and 6 are activated especially for category policy, etc. (see Fig. 7).

The ART-2 network was developed to have a comparable output with the SOM based categorizer. The categorizer has 9 output units (i.e. the network can create at most 9 clusters). The set of documents used for training SOM based categorizer was also used here. The number of actually created clusters was strongly dependent on the parameter ρ (vigilance threshold). The threshold was chosen experimentally to achieve the best results of categorization. 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 Tab. II. The meaning of values in the table is similar as for the SOM based categorizer. In this case, 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 training time for both categorizers strongly depends on the size of the networks and the number of training epochs. For network sizes mentioned above and 500 training epochs for each layer the computation time was between 2 and 4 hours (for CPU AMD Duron 650 MHz and operation system MS Win XP). Relations between training time and number of training epochs were also studied in the related work, where a computational complexity for different sizes of neural networks and number of training epochs are published (see [19]).

Comparison of SOM and ART-2 based categorizers is quite difficult considering a different topological structure of output layers of both networks. These networks were also used in [16], where the authors presented only networks outputs without any comparison. Since the clustered data are organized in a two dimensional array in SOM network (see Fig. 7) and neighboring units then contain similar data which belong to the same category (neighboring units are updated in the same time during

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

Tab. II *Results of document categorization using ART-2 categorizer.*

the map training, see Eq. 3). 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 setting the parameters of the network. There are 7 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 than the SOM categorizer.

At the time of finishing this paper, other similar experiments were conducted (another set of categories, various combinations of neural networks for WCD and DM, optimal number of training methods, etc.). However, current results show that the application of neural networks for Czech written document categorization is not as successful as we expected. As a result, the possible follow-up activities, like building larger data collections or optimizing the code for them, are not planned.

Acknowledgement

This work was supported by grant MŠMT No. 2C06009 Cot-Sewing. We also thank a group of students who helped to implement and test the software tools which were used for processing the corpora described in this paper.

References

[1] Apte C., Damerau F., Weiss S. M.: Automated learning of decision rules for text categorization. ACM Transactions on Information Systems, **12**, 1994, pp. 233–251.

[2] Carpenter G. A., Grossberg S.: The art of adaptive pattern recognition by a self-organizing neural network. Computer, **21**, 1988, pp. 77–88.

[3] Fausett L. V.: Fundamentals of Neural Networks. Prentice Hall, Englewood Cliffs, NJ, 1994.

[4] Fiesler E., Beale R., eds.: Handbook of Neural Computation. Oxford University Press, 1997.

[5] Honkela T., Pöllä M.: Concept mining with self-organizing maps for the semantic web. In: Principe J., Miikkulainen R., eds.: Advances in Self-Organizing Maps. Volume 5629 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2009, pp. 98–106.

[6] Honkela T., Laaksonen J., Törrö H., Tenhunen J.: Media map: A multilingual document map with a design interface. In Laaksonen, J., Honkela, T., eds.: Advances in Self-Organizing Maps. Volume 6731 of Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2011, pp. 247–256.

[7] Honkela T., Kaski S., Lagus K., Kohonen T.: Websom – self-organizing maps of document collections. In: Neurocomputing, 1997, pp. 101–117.

[8] Joachims T.: Text categorization with suport vector machines: Learning with many relevant features. In: Nedellec C., Rouveirol C., eds.: ECML. Volume 1398 of Lecture Notes in Computer Science., Springer, 1998, pp. 137–142.

[9] Kaski S., Honkela T., Lagus K., Kohonen T.: Websom-self-oganizing maps of document collections. Neurocomputer, 1998, pp. 101–117.

[10] Kohonen T.: Self-Organizing Map. Springer-Verlag, Berlin Heidelberg, 2001.

[11] Lagus K., Honkela T., Kaski S., Kohonen T.: Websom for textual data mining. Artif. Intell. Rev., **13**, 1999, pp. 345–364.

[12] Lai K. Y., Lam W. In: Document Categorization Using Multiple Similarity-Based Models, 2001, pp. 1–16.

[13] Rauber A., Merkl D., Dittenbach M.: The growing hierarchical self-organizing map: Exploratory analysis of high-dimensional data. IEEE Transactions on Neural Networks, **13**, 2002, pp. 1331–1341.

[14] Manninen T., Pirkola J.: (Classification of textual data with self-organising map: Neural computing as email filtering method).

[15] Manning C. D., Raghavan P., Schütze H.: An Introduction to Information Retrieval – Preliminary Draft. Cambridge University Press, 2007.

[16] Merkl D., Rauber A.: Document classification with unsupervised artificial neural networks, 2000.

[17] Yang Y., Chute C. G.: An example-based mapping method for text categorization and retrieval. ACM Trans. Inf. Syst., **12**, 1994, pp. 252–277.

[18] Yau Kwok J. T.: Automated text categorization using support vector machine. In: Proceedings of the International Conference on Neural Information Processing (ICONIP), 1998, pp. 347–351.

[19] Valenta M.: Aplikace neuronových sítí v oblasti zpracování česky psaných dokumentů (Application of Neural Networks for Czech-written Document Processing). Bachelor thesis, University of West Bohemia, 2009 (in Czech).