Text Document Retrieval by Feed-forward Neural Networks

Lenka Skovajsova*
Institute of Applied Informatics
Institute of Informatics
Slovak Academy of Sciences in Bratislava
Dubravska cesta 9, 845 07 Bratislava, Slovakia
lenka.skovajsova@savba.sk

Abstract

The paper deals with text document retrieval from the given document collection by using neural networks, namely cascade neural network, linear and nonlinear Hebbian neural networks and linear autoassociative neural network. With using neural networks it is possible to reduce the dimension of the document search space with preserving the highest retrieval accuracy.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Clustering; I.2.6 [**Learning**]: Connectionism and Neural Nets

Keywords

Text document retrieval, Document space dimension reduction, Latent semantic indexing, Cascade neural network, Hebbian neural networks, Autoassociative neural network, Document clustering

1. Introduction

Text documents became a part of our everyday life. Lots of papers in our offices were replaced by lots of documents in the computers and on the internet. As the amount of documents became more and more higher, the chance to find the right information is more and more lower. Because of this, the attempts to retrieve the right information in the sufficient amount of time are made.

Information retrieval is a wide research area mainly on the internet. One part of information retrieval is the retrieval of the text documents what is the focus of work.

When the number of documents is large, the amount of the data in the text document space can be reduced by various methods. One possibility to reduce the document space is to divide the documents into groups - clusters, with similar documents in each cluster. The work deals with text document space dimension reduction and their clustering by neural networks.

Work is divided into five sections. In the first section there is introduction into text document retrieval. In the second section of the work there is presented work done in the text document retrieval area, and the basic terms in this area are explained. Next, the basic information retrieval models which were used for text documents are there described. In the third section, the aims of the work are presented which are oriented in the proposal of text document retrieval models using analytical solution and by neural networks. In the fourth section, the experiments made are described and the results are shown. In the fifth section, the whole work is concluded.

2. State of the ART

Text document retrieval system consists from the two main parts: the user part and the text document retrieval part. At the beginning, the user poses the query to the system, the system creates from the query the inner representation and compares it with the representation of the documents. Most relevant documents connected with the query are sorted by the relevance and returned to the user as an answer. Documents with higher relevance are shown to the user as a first, documents with lower relevance are shown at the end of the document list.

Information retrieval models differ. They differ in the user query representation, they differ in the representation of documents and, also, they differ in the way of assigning relevant documents to the query. The query can be represented as a pattern, as a keyword, and in the different structure forms. Documents can be represented by inverted index, by matrix of keywords and documents, database, knowledge base etc.

It is often advantageous to group similar documents into several groups before they are asked for. Documents in the document collection can be grouped by two manners: by classification and by clustering. Classification is made by assigning documents to the predefined categories, clustering is somewhat different. The cluster, to which document belongs is not known before, the document is assigned there on the base of its properties. The documents

*Recommended by thesis supervisor: Prof. Igor Mokriš Defended at Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovakia on December 16, 2010.

Skovajsova, L. Text Document Retrieval by Feed-forward Neural Networks. Information Sciences and Technologies Bulletin of the ACM Slovakia, Vol. 2, No. 2 (2010) 70-78

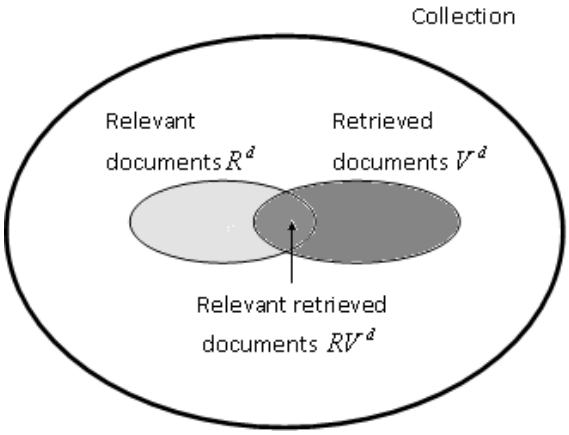

Figure 1: Precision and recall

with similar properties are placed in one cluster [13, 15]. Clustering can be divided on the hierarchical clustering and non-hierarchical clustering [13]. Hierarchical clustering methods can be divided on agglomerative and divisional clustering.

Text document retrieval system assigns higher relevance between the documents and query to documents that are more similar to the query. To ensure that the system works correctly, the retrieval parameters are computed. The most known parameters are the precision **P** and recall **R** (Figure 1), and the F-measure **F** computed from them.

Precision P is the division of the relevant documents found (relevant answer set, Ra), to the all documents found by the system (answer set, A).

$$\mathbf{P} = \frac{|Ra|}{|A|} \tag{1}$$

where $|.|$ gives the number of elements in the set.

Recall R is division of all relevant documents found by the system (Ra) to the number of all relevant documents (Rl)

$$\mathbf{R} = \frac{|Ra|}{|Rl|} \tag{2}$$

F-measure F is computed from precision and recall:

$$\mathbf{F} = \frac{1.5\mathbf{RP}}{0.5\mathbf{P} + \mathbf{R}} \tag{3}$$

Precision, recall and F–measure can be computed only for one query at a time. There is also possibility to compute the average precision, recall and F–measure as an average of the obtained values from each tested query.

There are three main approaches to the information retrieval nowadays. They are boolean approaches [1, 2], vector space model approaches [3, 8, 12], and the n-gram approaches [10, 14].

The boolean model is the first text document retrieval model used in the world. The queries in the boolean model are represented as a keywords separated by the operators AND, OR, NOT, and BUT. The documents in the boolean model are stored in an inverted index. The system finds documents that are in the set represented by the query expression.

In the vector space model, each document is represented as a vector of keywords, and when the document representations are placed as a columns in the matrix, such a matrix is called vector space model (VSM) matrix. Query is also represented as a vector of keywords. The relevance of the document to the query is computed as a cosine similarity between document and query representation.

Representation by n-grams is based on the assumption, that high probability is assigned to the words, that co-occur and low probability is assigned to the words that never co-occur.

Problem of the text document retrieval can be solved also by neural networks. Nowadays, there are different neural networks used for text document retrieval [6, 7, 9, 16]. Generally, we can divide neural networks for text document retrieval into three main categories: feed-forward neural networks (spreading activation neural network, COSIMIR) [9], Kohonen neural networks (WEBSOM, GHSOM) [7, 16], and recurrent neural networks (ART and ARTMAP neural networks, Hopfield neural network) [6].

The most often used model for text document retrieval is Vector Space Model. Because the VSM matrix is for large document collections very large and sparse, so there is a need to reduce the text document space and simultaneously preserve the accuracy of the document similarity on the maximum value. One of the dimension reduction techniques used in this work is Principal Component Analysis, PCA [4, 5, 11] that reveals principal components of the vector space model matrix. Principal components show the greatest variance or the largest energy places of the analyzed data. There are several approaches to compute the PCA, for example by Singular Value Decomposition (SVD). In this work the problem is solved by SVD and by neural networks.

Feature space dimension reduction methods by neural networks can be divided to linear and nonlinear methods and there can be used for example Hebbian neural networks or autoassociative neural network. All mentioned neural networks are described in the section 4.

3. The Thesis Objectives

In the work, we use vector space model for text document representation due to its suitable representation, it is not too large for large document collections as n-gram models and, it is not so imprecise as is the boolean model, and, with vector space model, the dimension reduction can easily be done. The vector space model can be represented by the spreading activation neural network. Using this paradigm, the cascade neural network model for text document retrieval can be proposed.

In this work, also, the text document space dimension reduction methods are proposed by two main approaches, and they are:

1. SVD on the base of LSI model,
2. Linear Hebbian neural network with Oja learning rule, nonlinear Hebbian neural network with Oja learning rule, and linear autoassociative neural network,

to find principal components of VSM matrix.

On the base of this, the objectives of the work can be divided into three areas:

1. The proposal of text document retrieval models by Vector Space Model and Feed-forward Neural Networks.
2. The proposal of text document retrieval models and text document space dimension reduction by Principal Component Analysis and LSI model and their retrieval by analytical solution.
3. The proposal of text document retrieval models and text document space dimension reduction by Principal Component Analysis and LSI model and their retrieval by feed-forward neural network solutions.

4. Proposal of the text document retrieval models

The text document retrieval model used in the work has these properties. Input query is represented as a set of keywords, the document collection is represented as a VSM matrix, and the document relevance is computed on the base of cosine similarity.

Text document retrieval part of the system (recall the division of the text document retrieval system in the introduction) more precisely has three main subsystems as it is on the Figure 2. First subsystem is the user query subsystem where user poses a query, second is the indexing subsystem where the documents and query are compared, and third is the administrator or documents subsystem where the documents are stored.

4.1 Proposal of text document retrieval model by cascade neural networks

This system has complex structure which can be simplified by substitution of the relations between the subsystems by neural networks, where first neural network solves the relation between user query and keywords of documents and second neural network solves the relation between keywords and relevant documents. On the base of this paradigm the cascade neural network can be proposed, where the first problem can be solved by the three-layer feed-forward neural network, and second problem can be solved by spreading activation neural network. First neural network can be used between user query subsystem and the indexing subsystem and second neural network can be used between indexing subsystem and document subsystem.

4.1.1 Proposal of Keyword Recognition Model on the base of the Three-Layer Feed-Forward Neural Network

The first neural network used is between the query subsystem and the indexing subsystem and it is the three–layer

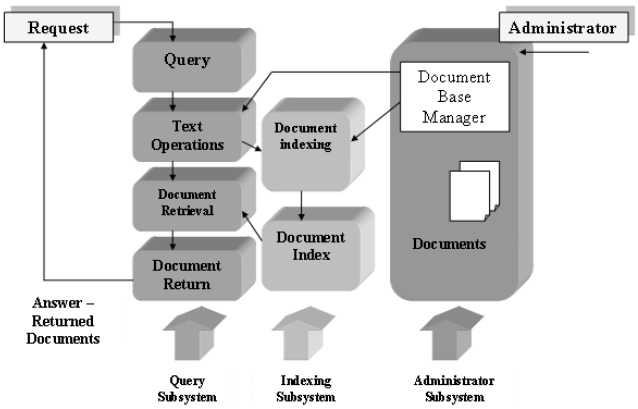

Figure 2: Text Document Retrieval System

feed–forward neural network where input layer represents the query in natural language, middle layer represents its inner representation and output layer represents the set of keywords. The user poses a query in the natural language on the input and the neuron in the output represents the recognized keyword. The backpropagation algorithm is used for training, and the output is computed by following formulas

$$net_{zj} = \sum_{i=1}^{n} w_{ij} k_i + \theta_{zj}, j = 1 \dots M \tag{4}$$

$$z_j = f(net_{zj}) = \frac{1}{1 + e^{-net_{zj}}} \tag{5}$$

$$net_{kj} = \sum_{i=1}^{m} vv_{ij} z_i + \theta_{kj} \tag{6}$$

$$x_j = f(net_{kj}) = net_{kj} \tag{7}$$

Three-layer feed-forward neural network was trained on the training set of 164 different types of twenty keywords and the used keywords were recognized with precision approximately 0.9959. But in the case of recognition of words that are not keywords, the reached precision was much lower.

4.1.2 Proposal of Text Document Retrieval Model on the base of the Spreading Activation Neural Network

The second neural network used between the indexing subsystem and the administrator subsystem is the Spreading Activation Neural Network.

Spreading activation neural network serves for the text document retrieval, where document collection is represented by the VSM matrix. VSM matrix has following structure:

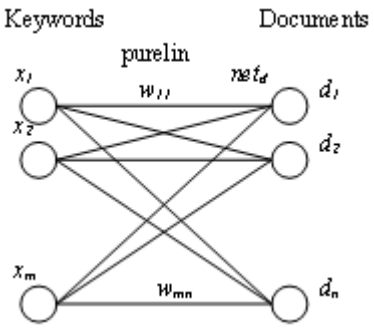

Figure 3: Spreading Activation Neural Network

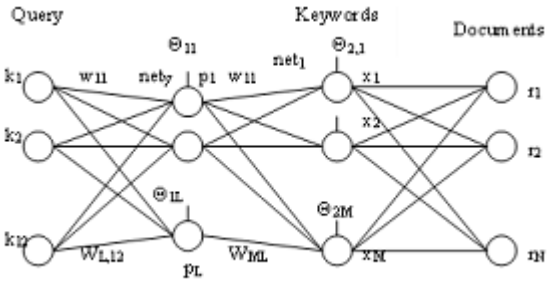

Figure 4: Cascade Neural Network Model

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ x_{m1} & x_{m2} & \dots & x_{mn} \end{pmatrix} \tag{8}$$

where x_{ij} is the frequency of the keyword i in the document j. Spreading activation neural network (Figure 3) has on the input the vector of keywords that represent the user query and on the output the vector of relevances of all the documents in the collection to the user query. The weights of the spreading activation neural network are the same as the elements of VSM matrix.

$$\mathbf{W} = \mathbf{X} \tag{9}$$

and the relevance of the documents to the given query is computed as

$$\mathbf{r} = \mathbf{qX} \tag{10}$$

so the spreading activation neural network has no training phase.

Spreading activation neural network is in experiments on the collection of 90 documents and 20 keywords shown as suitable for text document retrieval. Spreading Activation Neural Network has simple structure, but for large document collections it is very large, so it is suitable to reduce the number of elements in the VSM and Spreading Activation Neural Network

Three-layer feed-forward neural network and spreading activation neural network can be connected. Input to the three-layer feed-forward neural network is the user query in natural language and the output is the correct keyword. The keyword serves as an input to the spreading activation neural network where the output represents the vector of relevances of all the documents in the collection. So the output of the first neural network serves as an input to the second neural network. By combining these two neural networks the cascade neural network model is created (Figure 4).

4.2 Proposal of the Text Document Retrieval Model by the Dimension Reduction of the Text Document Space on the Base of Latent Semantic Indexing Model

From the reason of high dimension of VSM matrix when the document collection is large, it is suitable to reduce the dimension of the document space, that represents recognition feature space. The manner by which the text documents are reduced in the work is principal component analysis PCA, by which the set of keywords reduces to the much smaller feature set. The obtained model is the Latent Semantic Model. Most often method of the text document space dimension reduction is the singular value decomposition SVD of the VSM matrix, which represents the analytical solution.

VSM matrix is by the SVD reduced to the three submatrices:

$$\mathbf{X} = \mathbf{USV^T} \tag{11}$$

where $\mathbf{U}$ is the matrix of left singular vectors of the matrix $\mathbf{XX^T}$, $\mathbf{S}$ is diagonal matrix of positive singular values and $\mathbf{V}$ is the matrix of right singular vectors of the matrix $\mathbf{X^TX}$. By reducing the certain number of singular values, the size of matrices $\mathbf{U}$ and $\mathbf{V}$ is reduced to the $\mathbf{U_r}$ and $\mathbf{V_r}$, where r is the dimension after reduction and it is the same number as is the number of reduced singular values. Reduced VSM matrix then can be rewritten as

$$\mathbf{X_r} = \mathbf{U_r S_r V_r^T} \tag{12}$$

On the base of $\mathbf{U_r}$, $\mathbf{S_r}$ and $\mathbf{V_r}$ matrices, the representation of documents in the reduced feature space can be computed as

$$\mathbf{d_r} = \mathbf{dU_r} inv\,(\mathbf{S_r}) \tag{13}$$

and it can be compared with query representation

$$\mathbf{q_r} = \mathbf{qU_r} inv\,(\mathbf{S_r}) \tag{14}$$

where $\mathbf{d}$ is the representation of the document by the vector of keywords, $\mathbf{d_r}$ is the representation of the document

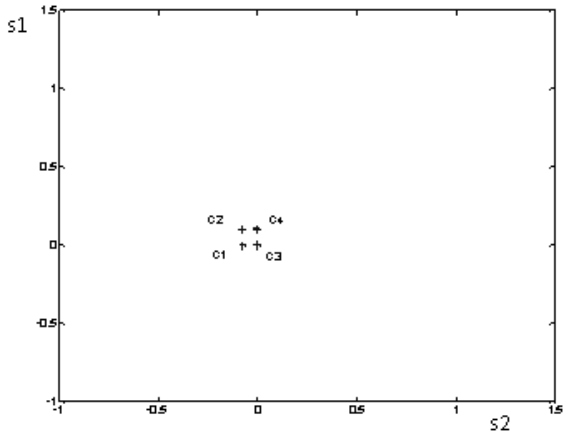

Figure 5: Clustering of documents by SVD

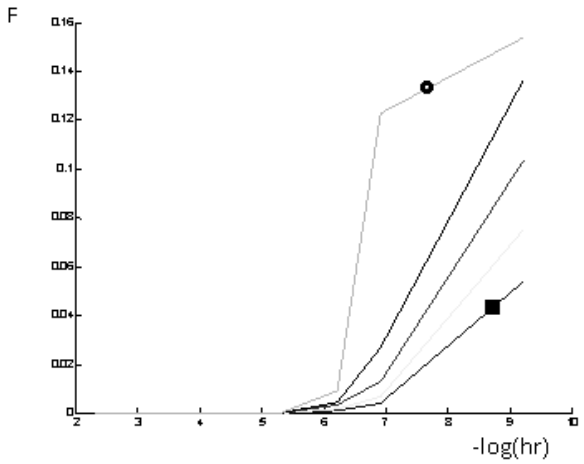

Figure 6: F–measure of document retrieval by Singular Value Decomposition

in the reduced feature space, $\mathbf{q}$ is the representation of query as a vector of keywords and $\mathbf{q_r}$ is the query representation in the reduced feature space.

The similarity of the document with the query is computed on the base of the cosine measure of their reduced representations in the feature space.

$$sim\,(\mathbf{d_r}, \mathbf{q_r}) = \frac{\mathbf{d_r}.\mathbf{q_r}}{|\mathbf{d_r}||\mathbf{q_r}|} \tag{15}$$

where $\mathbf{d_r}.\mathbf{q_r}$ is the inner product of the document and query in the reduced feature space and $|.|$ gives the length of the particular vector.

When the document representations are reduced to two–dimensional feature space, they can be depicted as a points in the two-dimensional coordinate system. As we will see, the clusters of similar documents in this coordinate system were formed with mutually similar documents in each cluster. In other words, by dimension reduction of the document space, similar documents are positioned near each other.

For the testing of the dimension reduction approaches, the three document collections are used. First document collection has 60 documents and 60 keywords, the matrix has dimension of 60x60. When the methods used work correctly, the three clusters are formed from the collection and in each cluster there are twenty documents which are always the same. The second testing collection is made from the part of Reuters[1] document collection and has 900 documents and 135 keywords. When the experiments were correct, the documents were divided into four clusters. The third document collection is also part of the Reuters document collection with 600 documents and 2000 keywords. If the experiments were correct, this document collection in the reduced space can be divided into two clusters.

As an example, four clusters are made from the collection of 900 documents and 135 keywords by SVD and they are shown on the Figure 5.

For each query there were chosen relevant documents (Rl) from the VSM matrix and then the answer set (A) of documents from the reduced document collection from the LSI was created. By comparison of these two sets of documents and averaging it over all tested queries, the precision, recall and F-measure were computed, see the equations (1), (2), (3).

The highest precision, recall and also F-measure was obtained by the model reduced to the 10 dimensions (Figure 6, circle over line) and the lowest values were obtained by the reduction to two dimensions (Figure 6, square over line), what was expected because the higher dimension after reduction, the more relevant documents is found in the answer set.

SVD has one disadvantage, and it is that the larger the document collection is, the more complex is the computation of SVD. Due to this disadvantage it is sometimes better to replace SVD by training the neural network.

4.3 Proposal of the Text Document Retrieval Models on the base of Feed-forward Neural Networks

Principal component analysis PCA is solved by SVD, that enables the dimension reduction of the document space.

The singular value decomposition is for large collections too complex, so it can be replaced by the training algorithms of neural networks.

For retrieval of the text documents there were used different types of neural networks in this work. The first type of neural network is the Hebian neural network with Oja learning rule, the second type is the autoassociative neural network.

[1] http://www.daviddlewis.com/resources/testcollections/reuters21578/

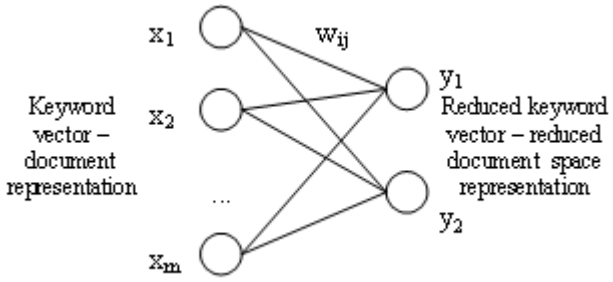

Figure 7: Linear Hebbian Neural Network

4.3.1 Proposal of the Text Document Retrieval Model on the base of The Linear Hebbian Neural Network with Oja Learning Rule

The first manner of using the linear PCA is the linear Hebbian neural network with Oja learning rule (Figure 7). It is a two-layer neural network, where the input layer has so many neurons, how many keywords is in the VSM matrix, and the output layer has so many neurons as is the number of dimensions in the reduced feature space. Weights are at the begining of training set to small random values. The learning algorithm is unsupervised. The document from the document collection comes on the input, then the output is computed and on the base of input, weights and output the weight change in the next time step is computed. Here are the formulas:

$$y_i(t) = \sum_p w_{ip}(t) x_{ip}(t) \tag{16}$$

$$w_{ij}(t+1) = w_{ij}(t) + \gamma(t) y_i(t) \left(\mathbf{x_j}(t) - y_i(t)\mathbf{w_i}(t)\right) \tag{17}$$

where $\mathbf{x}$ represents one document from the VSM matrix, $\mathbf{y}$ represents the output of the neural network, i.e. document representation in the reduced feature space, w_{ij} represents the connection weight between j-th input neuron and i-th output neuron, $\mathbf{W}$ represents the weight matrix, t is discrete time step, and $\gamma(t)$ is the learning parameter. After training the neural network, the document representation appears on the input and the reduced document representation is on the output.

Linear Hebbian neural network with Oja learning rule has one advantage against SVD. The whole proces of SVD computation is replaced by training the neural network. Training the Hebbian neural network gives sometimes better output then the SVD. After training on all three collections separately, the neural network created the expected clusters, so it works correctly (Figure 8).

Linear Hebbian neural network gives the best results of precision, recall, and F-measure (Figure 9) by the reduction to ten dimensions (circle over line) and the worst results by the dimension reduction to two dimensions (square over line).

4.3.2 Proposal of the Text Document Retrieval Model on the base of Nonlinear Hebbian Neural Network with Oja Learning Rule

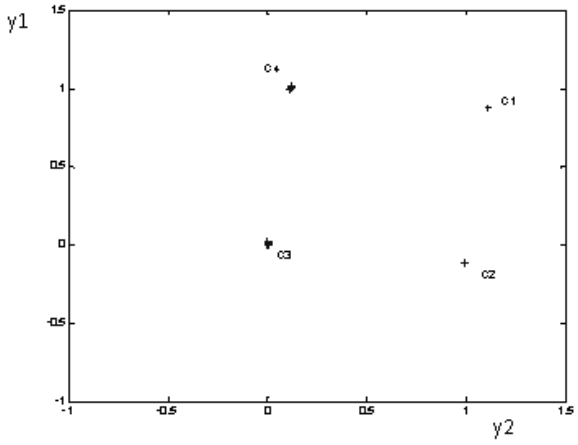

Figure 8: Clustering of documents by the Linear Hebbian Neural Network with Oja Learning Rule

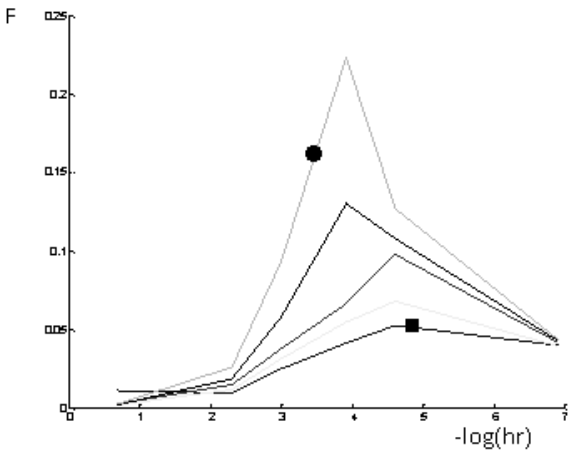

Figure 9: F–measure of document retrieval by Linear Hebbian Neural Network

Principal component analysis has also its nonlinear version, represented by the nonlinear Hebbian neural network with Oja learning rule. In order to compare the performance and the accuracy with the linear neural network, it is now analyzed.

Nonlinear Hebbian neural network has the same topology as linear Hebbian neural network (Figure 7), only in the output layer, there is nonlinear activation function, $y = g(y)$, where $g(y)$ is nonlinear output function, in this paper we used for comparison $tanh()$ function. The learning rule is represented by the formula:

$$w_{ij}(t+1) = w_{ij}(t) + \gamma(t) \left(g(y_i(t))x_j(t) - g(y_i(t))\right) \mathbf{w}_i(t) \tag{18}$$

Nonlinear Hebbian neural network gives due to the nonlinear activation function different results as a linear Hebbian neural network or SVD (Figure 10), but there were also expected clusters created. The performance accuracy

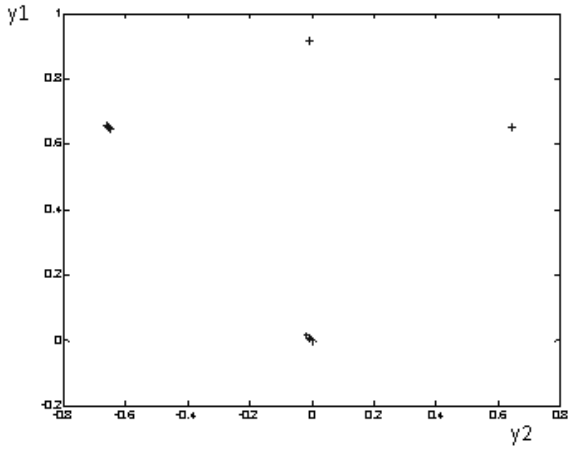

Figure 10: Clustering of documents by the Nonlinear Hebbian Neural Network with Oja Learning Rule

is not so good as in the linear Hebbian neural network, but it can be used for document clustering too.

Nonlinear Hebbian neural network was analyzed on each of the three collections, and by using different nonlinear functions. As a best nonlinear functions there were shown the tanh and logsig, where the expected clusters were created and the worst behavior has shown the function radbas, where no clusters were created. The functions logsig and radbas are defined by the formulas

$logsig(n) = \frac{1}{1+e^{-n^2}}$

$radbas(n) = e^{-n^2}$

Nonlinear Hebbian neural network gives best F-measure (Figure 11) when it is reduced to ten dimensions and the worst F-measure, when it is reduced to two dimensions.

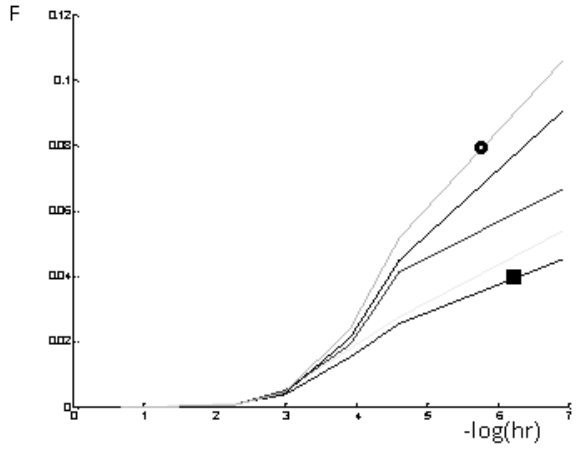

Figure 11: F–measure of document retrieval by Nonlinear Hebbian Neural Network

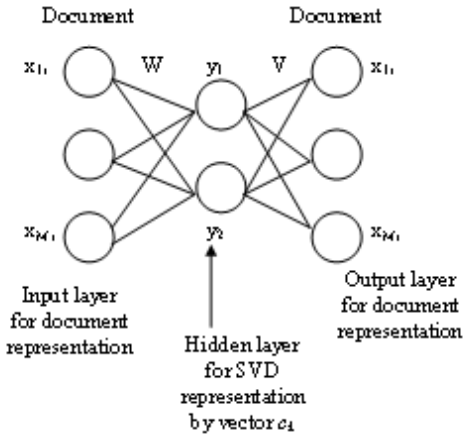

Figure 12: Autoassociative Neural Network

4.3.3 Proposal of the Text Document retrieval Model on the base of Linear Autoassociative Neural Network

Linear autoassociative neural network has three layers (Figure 12). The input and output layers have the same number of neurons, representing the vector of keywords. The number of neurons in the hidden layer equals the dimension of the reduced feature space. All layers in the linear autoassociative neural network have linear activation function. The same document $\mathbf{x_k}$ represents both, input and expected output. The training algorithm is unsupervised backpropagation, that means, on the base of the expected output that equals the input, only the input to the neural network for training is needed. After the training the document is brought on the input and the values of hidden neurons are computed, which represent the document in the reduced feature space, given by the formula

$$\mathbf{c} = \mathbf{x}_k\mathbf{W} \tag{19}$$

where $\mathbf{c}$ represents the coordinates of document representation in the reduced feature space, $\mathbf{x_k}$ is input and expected output document and $\mathbf{W}$ is the weight matrix between the input and the hidden layer.

Linear autoassociative neural network gives similar results as linear Hebbian neural network with Oja learning rule, the correct clusters were always formed (Figure 13). F-measure for comparison is depicted on the figure 14. In comparison with linear Hebbian neural network, the autoassociative neural network is faster, but not so accurate.

4.4 Assessment

There were applied four methods (SVD, linear Hebbian neural network, nonlinear Hebbian neural network, and autoassociative neural network) on three document collections (60x60, 900x135, 600x2000) that is twelve experiments together. In each experiment, the precision, recall and F-measure was computed. From the obtained results it is obvious that for given training sets, the linear autoassociative neural network is faster than Hebbian neural networks, but is less precise, then the linear Hebbian neural network. By several trainings of the neural net-

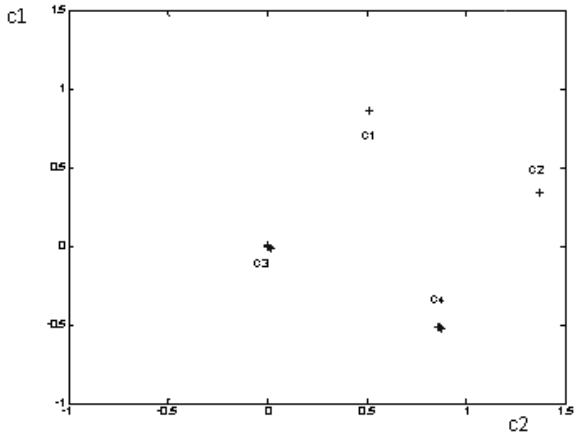

Figure 13: Clustering of documents by Linear Autoassociative Neural Network

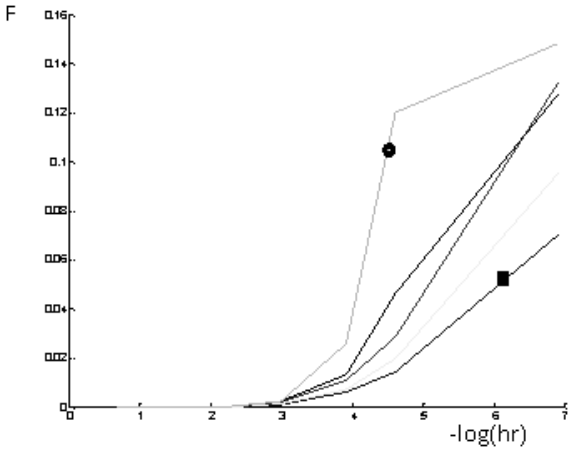

Figure 14: F–measure of document retrieval by Linear Autoassociative Neural Network

work, the formed clusters were not always positioned on the same place, because of random setting the weights on the beginning of the training, but the right clusters were always formed.

From the obtained precisions, recalls and F-measures, depicted in the work, it can be seen, that the best results were obtained by the collection 60x60, what can be expected, because it was the smallest and least sparse collection. Less precise outcomes belong to the 600x2000 collection and the worst result gave the 900x135 collection, what can be expected again, because it was most sparse. Highest average F-measure had the LSI model and the lowest average F-measure had the nonlinear Hebbian neural network. But all four models gave similar results.

The highest difference between the LSI model and the neural networks used in this work is that the matrices obtained by the SVD are fixed for dimension reduction r. Against it, when we want to change the dimension after the reduction by the neural network, we have to train the network again. On the opposite side, neural networks have one advantage against the LSI model, they work with only one document representation at a time, whilst in the LSI we need whole document collection at once to compute the SVD. From the compared algorithms all can be used for the dimension reduction and also for the document clustering, as they have the similar results. Used neural networks use unsupervised learning algorithms. By each of the algorithms, the number of clusters is not known in advance.

5. Conclusions

In the state of the art, the basic terms are described and the related work is presented. From the state of the art it follows that the most suitable query representation for our purposes is in the form of keywords and the document collection representation by the VSM matrix. The easiest representation of VSM matrix is by the spreading activation neural network.

The text document retrieval part of system has three main subsystems, and they are the query subsystem, indexing subsystem and the administrator subsystem. The transitions between these subsystems can be replaced by the neural networks, where input of the first neural network represents the user query subsystem and the output of the first neural network represents the indexing subsystem. The output of the first neural network represents also the input to the second neural network – spreading activation neural network. The output of the second neural network represents the relevance of documents – the administrator subsystem. Such a model is called cascade neural network model.

The VSM matrix represented by a spreading activation neural network is for large collections too large, so it is suitable to reduce the dimension of the document space. It can be solved by SVD, that creates the reduced query representation and the reduced representation of the documents, and so it is possible to reduce the complexity of document representation. Model that uses the SVD for dimension reduction is called the Latent Semantic Indexing (LSI) model.

Because the SVD is very complex for large VSM matrices, there appears possibility to replace the SVD by training the neural network. Neural networks used in the work search for the principal components of the VSM matrix. There are used three types of neural networks, linear and nonlinear Hebbian neural network and autoassociative neural network. All of these neural networks can be used not only for the dimension reduction but also for the document clustering.

The main contribution of this work is the analysis and synthesis of described algorithms for dimension reduction and clustering the text documents and for text document retrieval by means of neural networks in comparison with LSI and SVD.

Acknowledgements. This contribution has been supported by the grant VEGA 2/7098/27 and VEGA 2/0211/09

References

[1] Cordon, O., Moya, F., Zarco, C.: *A New Evolutionary Algorithm Combining Simulated*

Annealing and Genetic Programming for Relevance Feedback in Fuzzy Information Retrieval Systems. Soft Computing- A Fusion of Foundations, Methodologies and Applications, Vol. 6, Number 5, August 2002, pp. 308-319.

[2] Herrera-Viedma, E.: *Modelling the Retrieval Process for an Information Retrieval System using an Ordinal Fuzzy Linguistic Approach.* Journal of the American Society for Information Science and Technology, Vol. 52, , Issue 6, 2001, pp. 460-475.

[3] Hotho A., Nurnberger, A., Paas, G.: *A Brief Survey of Text Mining.* LDV-forum, Volume 20(1), 2005, pp. 19-62.

[4] Jutten, C., and Karhunen, J.: *Advances in Blind Source Separation (BSS) and Independent Component Analysis (ICA) for Nonlinear Mixtures.* Int. J. Neural Systems , Vol. 14, No. 5, 2004, pp. 267-292.

[5] Karhunen, J., and Ukkonen, T.: *Generalizing Independent Component Analysis for Two Related Data Sets.* In Proc. of the IEEE 2006 Int. Conf. on Neural Networks / 2006 IEEE World Congress on Computational Intelligence (IJCNN2006/WCCI2006), Vancouver, Canada, July 2006, pp. 1822-1829.

[6] Kondadadi R., Kozma R.: *A Modified Fuzzy ART for Soft Document Clustering.* International Joint Conference on Neural Networks, World Congress on Computational Intelligence, Honolulu, Hawaii, 2002, pp. 2545-2549

[7] Lagus K., Kaski S.,Kohonen T.: *Mining Massive Document Collection by the WEBSOM method.* Elsevier, ISSN 0020-0025, 2004, pp. 135-156.

[8] Lan, M., Tan, C., Low, H., and Sung, S.: *A comprehensive comparative study on term weighting schemes for text categorization with support vector machines.* In Special interest Tracks and Posters of the 14th international Conference on World Wide Web (Chiba, Japan, May 10 – 14, 2005). WWW '05. ACM Press, New York, NY, 2005, pp. 1032-1033.

[9] Mandl, T.: *Tolerant and Adaptive Information Retrieval with Neural Networks.* Global Dialog, Science and Technology, Thinking the Future. EXPO 2000, Hannover, 2000.

[10] Morin, F., Bengio, Y.: *Hierarchical Probabilistic Neural Network Language Model.* In Robert G. Cowell and Zoubin Ghahramani, editors, AISTATSâĂŹ05, 2005, pp. 246-252.

[11] Raju, K., Ristaniemi, T., Karhunen, J., and Oja, E.: *Jammer Suppression in DS-CDMA Arrays Using Independent Component Analysis.* IEEE Trans. on Wireless Communications, Vol. 5, No. 1, January 2006, pp. 77-82.

[12] Scheir, P., Lindstaedt, S. N.: *A Network Model Approach to Document Retrieval taking into account Domain Knowledge.* LWA 2006, Lernen-Wissendeckung-Adaptivitat, Hildesheim, Universitat Hildesheim, 2006, pp. 154-158.

[13] Treeratpituk, P., Callan, J.: *An Experimental Study on Automatically Labeling Hierarchical Clusters using Statistical Features.* Annual ACM Conference on Research and Development in Information Retrieval, 2006, pp. 707-708.

[14] Wang, S. et. al.: *Combining Statistical Language Models via the Latent Maximum Entropy Principle.* Machine Learning, Volume 60, Numbers 1-3, 2005, pp. 229-250.

[15] Wei, X., Croft, W. B.: *LDA-based Document Models for ad-hoc Retrieval. Annual Conference on Research and Development in Information Retrieval*, Seattle, Washington, USA, 2006, pp. 178-185.

[16] Zhong, S.: *Efficient Streaming Text Clustering.* Neural Networks, Issues 5-6, July-August 2005, pp. 790-798.

Selected Papers by the Author

I. Mokriš, L. Skovajsová. Neural Network Model of System for Information Retrieval from Text Documents in Slovak Language. *Acta Electrotechnica et Informatica*, 5(3): 36–41, 2005.

I. Mokriš, L. Skovajsová. Development of Neural Network Information Retrieval System from Text Documents. In *3rd Slovakian – Hungarian Joint Symphosium on Applied Machine Intelligence*, pages 123–131, Herl'any, Slovakia, 2005.

I. Mokriš, L. Skovajsová. Information Retrieval by means of Vector Space Model of Document Representation and Cascade Neural Networks. In *1st Workshop on Intelligent and Knowledge Oriented Technologies, WIKT 2006*, pages 102–105, Bratislava, Slovakia , 2007.

L. Skovajsová, I. Mokriš. Text Document Space Dimension Reduction by Latent Semantic Model. In *1st Workshop on Intelligent and Knowledge Oriented Technologies, WIKT 2006*, pages 106–109, Bratislava, Slovakia, 2007.

I. Mokriš, L. Skovajsová. Feed-Forward Neural Networks for Information Retrieval from Text Documents in Natural Language. *Acta Electrotechnica et Informatica*, 8(2): 3–10, 2008.

I. Mokriš, L. Skovajsová. Proposal of Latent Semantic Model for Document Set Representation by Neural Network. In *2nd Workshop on Intelligent and Knowledge Oriented Technologies, WIKT 2007*, Faculty of EE and Informatics, Technical University of Košice, pages 102–105, 2008.

I. Mokriš, L. Skovajsová. Proposal of Cascade Neural Network Model for Text Document Space Dimension Reduction by Latent Semantic Indexing. In *SAMI 2008 – 6th IEEE International Symposium on Applied Machine Intelligence and Informatics*, pages 79–84. Herl'any, Slovakia, January 21–22, 2008.

I. Mokriš, L. Skovajsová. Document Space Dimension Reduction by Latent Semantic Analysis and Hebbian Neural Network. In *6th IEEE International Symposium on Intelligent Systems and Informatics, IEEE, SISY 2008*, pages 25-27, Subotica, Serbia, September 25–27, 2008.

I. Mokriš, L. Skovajsová. Comparison of Two Document Clustering Techniques which Use Neural Networks. In *ICCC2008 – 6th IEEE International Conference on Computational Cybernetics, IEEE*, pages 75–78, Stará Lesná, Slovakia, November 27–29, 2008.

L. Skovajsová, I. Mokriš. Document Space Dimension Reduction by Nonlinear Hebbian Neural Network. In *SAMI 2009 – 7th IEEE International Symposium on AppliedA Machine Intelligence and Informatics*, pages 89–91, Herl'any, Slovakia, January 30–31, 2009.

L. Skovajsová, I. Mokriš. Nonnegative Factor Analysis for Text Document Clustering. In*Proceedings of the 9th WSEAS International Conference on Simulation, Modelling and Optimization*, pages 345–349, Budapest Tech, Hungary, Sept. 3 – 5, 2009.

L. Skovajsová, I. Mokriš Bigradient Learning Algorithm for Dimension Reduction of Text Document Space. In *ICCC 2009 – 7th IEEE International Conference on Computational Cybernetics*, IEEE Catalog Number CFP09575-CDR, pages 125 – 128, Palma de Mallorca, November 26–29, 2009

L. Skovajsová, I. Mokriš Text Document Retrieval by Neural Networks. In *The Tenth International Conference on Informatics Proceedings, Informatics 2009*, pages 312–318, Herl'any, 23.–25. 11., 2009.