NEURAL NETWORK MODEL OF SYSTEM FOR INFORMATION RETRIEVAL FROM TEXT DOCUMENTS IN SLOVAK LANGUAGE

Igor MOKRIŠ, Lenka SKOVAJSOVÁ
Institute of Informatics, Slovak Academy of Sciences, Bratislava, Slovak Republic,
E-mail: mokris@valm.sk, skovajsova@valm.sk

SUMMARY

The aim of the paper is to describe the information retrieval model which retrieves the information from the text documents in Slovak language and which, for this purpose, uses the neural networks. This model comes from linguistic and conceptual approach for the analysis of text documents in Slovak language. The neural network model, based on multilayer perceptron and spreading activation network type, accepts the structure of conceptually and linguistically oriented model, where problems of document database creation and document indexing for keyword determination are solved. Proposed structure of the neural network model solves the problem of the document retrieval on the base of user's question. However, learning algorithm and neural network invariance, come from utilization of the neural networks, enable the decrease of the computational complexity of the language analysis algorithm.

Keywords: *Information Retrieval, Queries, Keywords, Text Documents, Neural Networks, Slovak Language*

1. INTRODUCTION

The aim of this paper is to describe the development of the information retrieval system, which retrieves the information from the text documents in Slovak language by neural networks and comes from the information retrieval system using statistical, conceptual, and linguistic model [6].

With growing number of information in the space of the information retrieval system, there is also growing need of the effective information retrieval in this space. There were, therefore, developed many different information retrieval systems [2,11]. This paper shows the possibility to simplify the computational complexity of the information retrieval process for documents in Slovak language by neural networks.

Slovak language is, as oppose to many other languages, difficult to process on a computer. Nouns, adjectives, pronouns, numerals are inflected differently; moreover, verbs are timed differently, too. Therefore, it is difficult to recognize keywords in Slovak text. Keywords are words, which take part in a document indexing. Next problem arises with synonyms, which are the words with different shape, but similar meaning. Another problem develops with homonyms, which are words with similar shape but with different meaning. Furthermore, another problem occurs with phrases, which contain more than one word, and so on. Situation can also be more complicated with varied text structures which are not based on words, but, for example, on dates, various numberings, etc.

Various approaches are used for the Slovak text analysis. Most common are statistical approach, linguistic approach and knowledge-based approach. The statistical approach analyses words in text documents by comparing them with keywords. The keyword set can be made manually, or can be created automatically from the documents [16].

The linguistic approach extracts linguistic units from the text [14]. Linguistic unit can be phoneme, morpheme, lexeme, and so on. Linguistic text analysis consists of the partial analyses, as phonological, morphological, syntactical, semantic and pragmatic analysis. Result is whole representation of the text, where every relevant data from the point of the content are marked. These are clearly identified language units and relationships between them.

Knowledge-based approach uses concepts or parts of text, which are associated with context for document indexing [18]. Knowledge model can be defined as a formal description of documents, concepts and relationships between them, with emphasis on their semantics. These concepts create structures with related documents in given document domain. This model is named domain model. Knowledge-based approach uses mainly semantic nets, existential graphs, conceptual graphs and ontologies. More detailed description of ontologies can be found in [17,19,21]. Aggregated description of ontological languages can be found on the world wide web [5,7,12,13,15,20], etc.

Because the information retrieval process based on the statistical, conceptual, and linguistic approach is very difficult, complex and time consuming, it is advantageous to use the neural networks for the development of the information retrieval system [1,3,9]. Well-trained neural network is able to determine stems in the words better and, therefore, there are no problems with complicated Slovak text language analysis. Besides, the neural networks in their life phase from point of their invariance on inflecting of Slovak words, perform faster linguistic analysis of Slovak documents than linguistically oriented information retrieval system. However, the

This work was supported by Science and Technology Assistance Agency under the contract No. APVT-51-024604.

system proposed with neural networks can have some disadvantages, for instance, training of neural networks, when new documents are added to the document base and contain new keyword.

2. STATISTICAL, LINGUISTIC AND KNOWLEDGE-BASED INFORMATION RETRIEVAL SYSTEM

Model of the information retrieval system with neural networks comes from the model based on statistical, linguistic and knowledge-based approach, which expresses document content and document relevance. User specifies the query for that system and system returns a document subset relevant to his query.

On fig.1 there is developed the information retrieval system for text documents in Slovak language based on statistical, linguistic and knowledge-based approach [6]. The structure of the system is modular and has separated query subsystem, indexation subsystem and document subsystem. Query subsystem is used for preprocessing of an entering query and acquisition and returning of the relevant documents to the user. It consists of query processing module, retrieval module, relevance feedback solution module and relevant arrangement and result list creation module.

In the query-processing module the query is formulated and preprocessed. In the retrieval module, the inner representation of query is compared with index of documents from the indexation part. If there are some marked documents sent as a new query, then they can be processed in the feedback solution module. In the end the retrieved documents from knowledge domain model part are arranged and sent to the user by arrangement and result list creation module.

In the indexation subsystem there is language analysis part and indexation part. The language analysis part consists of processing, morphological, lexical, syntactic and semantic module. These modules are used for the language analysis of the text documents. Indexation part consists of attribute index module, full text index module and conceptual index module. This part is used for indexing of documents in the document base module. Conceptual index is index associated with the concept vector and belongs to each document as its index.

Full text index is index, which uses the matrix of keywords and documents. Each document in this matrix is represented by the keyword vector. Each keyword belonging to given document is expressed by its relative frequency in the document [10]. This representation is called vector space model.

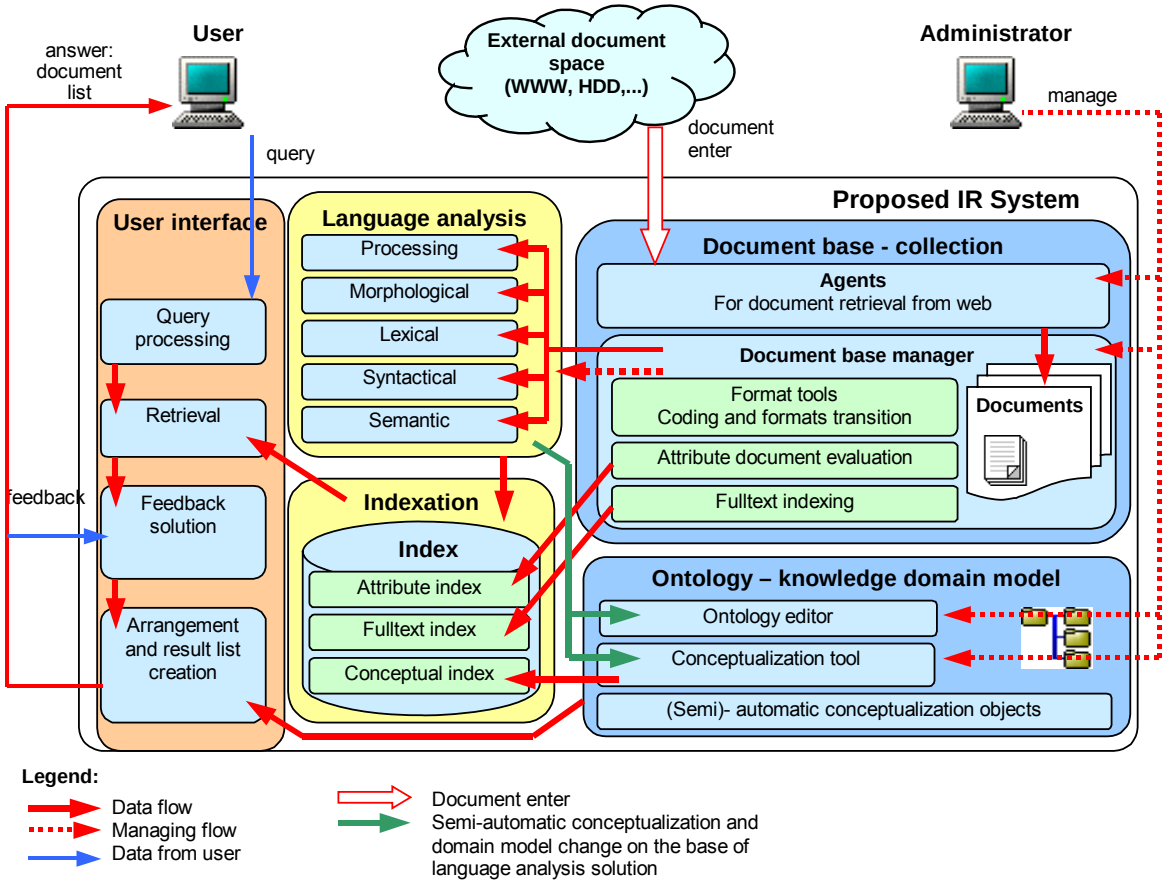

Fig. 1 Modular structure of information retrieval system

The document subsystem consists of two parts. First is the document base part and second one is the knowledge-based domain model part. The document base part consists of the agent module and document manager module. The document manager module consists of format tools module, attribute document evaluation and full text indexing module. Agent module serves for acquiring the documents from the web. Document manager module is used for managing the documents. Format tool module transits its coding and format. Attribute evaluation module evaluates the attributes in the document text and full text indexing module is used for assigning the index to documents in the document base.

The document base part consists of ontology editor module, conceptualization module and (semi-) automatic conceptualization object module. Ontology editor module is used for creation and modification of ontology. Conceptualization tool module with language analysis part is used for the conceptualization of documents.

3. NEURAL NETWORK INFORMATION RETRIEVAL SYSTEM

Because of the modular structure complexity the information retrieval system can be divided into three different subsystems. There are administrator subsystem, indexation subsystem and user subsystem (fig. 2).

The administrator subsystem guaranties the administration of the documents. Administrator determines the document base from the document set. Document base manager then provides the system representation of the documents. He also determines a suitable model for document storage and creates the system representation of the documents. Indexation subsystem solves two tasks. Firstly, the creation of an index and secondly, the creation of a query representation that is comparable with the document index.

User subsystem processes user query and searches for relevant documents. Firstly, user puts a query. User subsystem processes this query and assigns a keyword to it. Then the indexation subsystem indexes the query, which is then compared with the document index. The administrator subsystem retrieves relevant documents and sends them to user according to this comparison. The user can use feedback, in which the user marks the most relevant documents from the set of retrieved documents and consequently sends it as a new query. The system creates a new query from these documents and searches again the document base.

These three subsystems of the information retrieval system can be represented as a three layer model (fig. 3).

The first sublayer of this system is a query sublayer, the second one is the keyword sublayer and the third one is the document sublayer. The user enters a query, which is associated with a keyword, according to which the relevant documents are retrieved from the document sublayer.

Transition of the information from query sublayer into keyword sublayer and transition of information from keyword sublayer to document sublayer can be replaced by neural networks, as is expressed on fig. 4 [4].

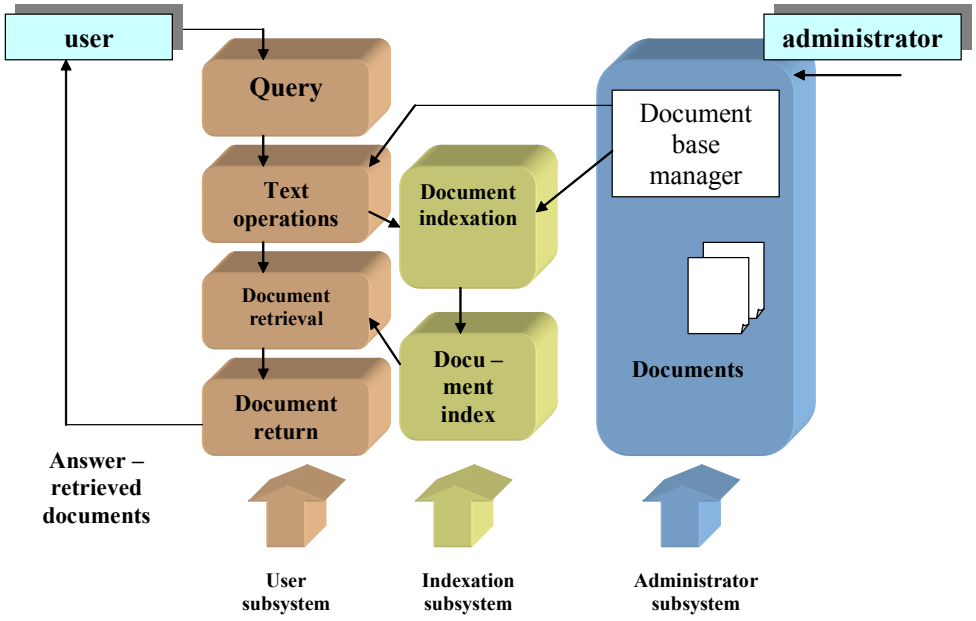

Fig. 2 Simplified information retrieval system structure

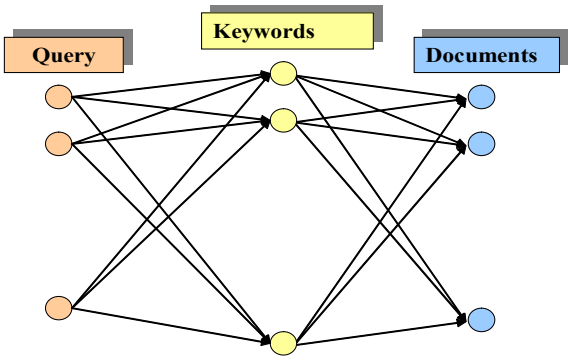

Fig. 3 Three layer information retrieval system

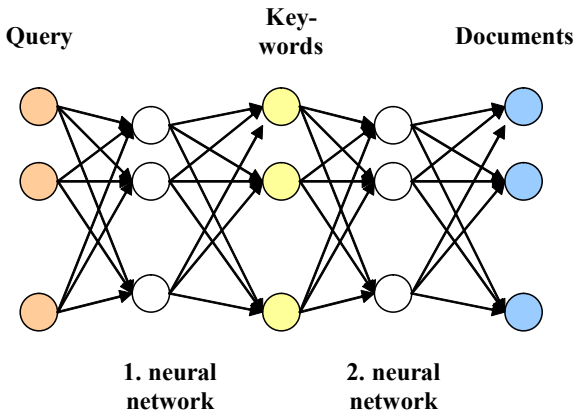

Fig. 4 Neural network information retrieval model

4. DEVELOPED NEURAL INFORMATION RETRIEVAL SYSTEM DESCRIPTION

First neural network (fig. 5, multilayer perceptron - back propagation type) consists of three layers, i.e., input layer, hidden layer and output layer [8]. Input layer is created of N input neurons x_1, …, x_N, where each neuron represents one character of a query, i.e. input layer represents one word. Hidden layer is created by M neurons y_1, …, y_M, which express the inner query representation. Output layer is created by L neurons k_1, …, k_L, where each neuron represents one keyword.

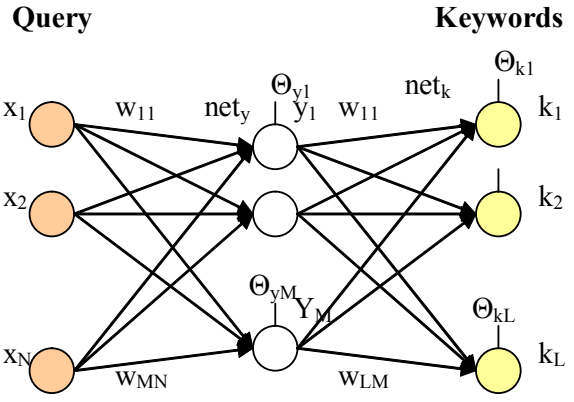

Fig. 5 Neural network for determination of keywords

Hidden layer is used for the query representation based on the formula

$$net_{yj} = \sum_{i=1}^{N} w_{ij} x_i(t) + \theta_{yj} \text{ , j=1...M} \quad (1)$$

The neuron y_j of the hidden layer is defined by sigmoid transition function in a form

$$y_j = f(net_{yj}) = \frac{1}{1+e^{-net_{yj}}}, \quad (2)$$

Outer value net_{kj} of hidden layer is defined as follows:

$$net_{kj} = \sum_{i=1}^{M} w_i y_i(t) + \theta_{kj} \text{ ,} \quad (3)$$

The neuron k_l of the output layer is defined by linear transition function in a form

$$f(net_{kj}) = net_{kj} \quad (4)$$

For learning this neural network a backpropagation algorithm was used.

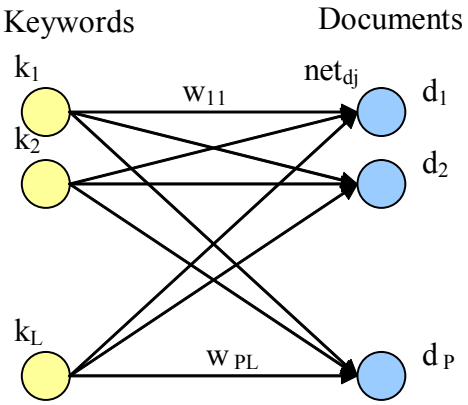

Fig. 6 Neural network for determination of relevant documents

Second neural network (spreading activation network) consists of an input layer created by keywords k_1, ..., k_L and an output layer created by documents d_1, ..., d_P, where each input neuron represents one keyword and each output neuron represents one document or document related subset of document base.

The neuron of the input layer of spreading activation network is the same as the neuron of the output layer of first neural network defined by (4).

The neuron of the output layer of spreading activation network is defined by linear transition function in the form

$$f(net_{dj}) = net_{dj} \quad (5)$$

Weights w_{p1} of neural network can be determined on the base of relative frequency matrix

of the keywords $\mathbf{F}_{Lp}$ in matrix of keywords and documents. Relative frequency matrix of keywords and documents is also called the vector space model [10]. Vector space model is created by rows of keywords k_1, ..., k_L and columns of documents d_1, ..., d_p by relation

$$F(L\times P)=\begin{pmatrix}k_1\\k_2\\\dots\\k_L\end{pmatrix}=\begin{array}{c}\begin{matrix}d_1 & d_2 & & d_P\end{matrix}\\\begin{pmatrix}f_{11} & f_{12} & \dots & f_{1P}\\f_{21} & f_{22} & \dots & f_{2P}\\\dots & \dots & \dots & \dots\\f_{L1} & f_{L2} & \dots & f_{LP}\end{pmatrix}\end{array} \quad (6)$$

where:

- L is the number of keywords,
- P is the number of documents,
- k_i is i-th keyword,
- d_j is j-th document,
- f_{ij} is relative frequency of the i-th keyword in the j-th document called also keyword weight.

Matrix obtained as a vector space model is normed in such manner in order to obtain the relevance in interval <0,1> .

The spreading activation network is not trained, its weights are determined on the base of elements f_{ij} of matrix of normed vector space model, which are assigned into weights of neural network

$$\mathbf{W}_{ij}=\mathbf{F}_{ij}, \quad i=1..P, j=1..L \quad (7)$$

Developed neural network model of information retrieval system is depicted in fig. 7.

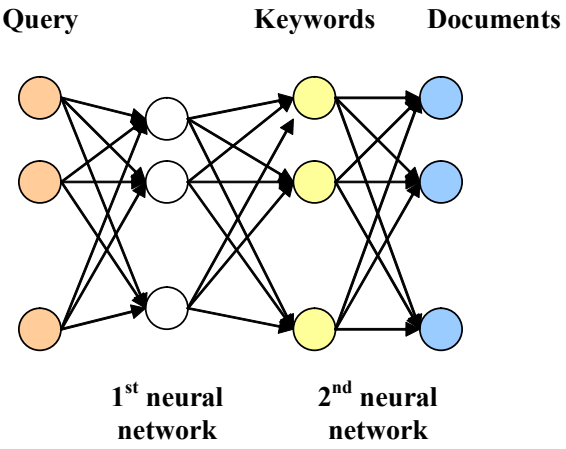

Fig. 7 Developed neural network model of information retrieval system

5. CONCLUSION

On the base of the above mentioned model, MATLAB program, which was used for testing the correctness of developed model on the neural network base, was created.

The developed model consists of two neural networks (fig. 7). They are the 1st neural network (NN_1) between query layer and keyword layer and the 2nd neural network (NN_2) between the keyword layer and the document layer. In the query layer there are 12 neurons - each for one character, which represents the user query. In the keyword layer, there are 20 neurons, where each neuron indicates one keyword. In the document layer, there are 90 neurons, each for one document. The length of each document is approximately about 50 words.

The first neural network NN_1 is trained with a query training set QTrS, involving 164 queries and a keyword training set KwTrS, which involves 20 keywords. Query set is created on the base of appropriate grammatical forms of Slovak language for appropriate chosen words and the keyword set is created on base of the root shapes of chosen words. Learning process is oriented on the association between the entered queries and the keywords, where keywords from KwTrS are created from associated queries from QTrS.

The second neural network NN_2 uses keyword training set KwTrS and document set DS. This network is not trained because its weights are assigned by the matrix of normed vector space model (7).

Within the model testing two experiments were made. First experiment was made with the first query test set $QTsS_1$, consisting of 185 queries, which contained different grammatical word forms. These forms were then used for the root bases of the keyword training set KwTrS. For questions from $QTsS_1$ belonging keywords from KwTrS were found and for them the documents from the document set DS with achieved precision of 0,9959 were found.

Second experiment was made with second test set $QTsS_2$, where 100 queries were involved but no keyword belonged to it and this fact influences that no documents can be returned to the user. From the results obtained, it follows, that the system reacted to chosen query training set with precision of 0.97.

From the whole assessment of the experiment, it follows, that the approach used has a perspective and provides next possibilities for its widespread.

But the main advantage of this approach is an important decrease in the computational complexity of the information retrieval process from the text documents in Slovak language in relation to the model proposed according to the linguistically oriented approach [6].

REFERENCES

[1] Berg, J., Schuernie, M.: Information Retrieval Systems using Associative Conceptual Space. European Symposium on Artificial Neural Networks. 1999, ISBN 2-600049-9-X, pp. 351-356.

[2] Clarke, I., Sandberg, O., Wiley, B., Hong, T. W.: Freenet: A Distributed Anonymous Information Storage and Retrieval System. Springer, Berlin, 2000.

[3] Crestani, F.: Implementation and Evaluation of a Relevance Feedback Device Based on Neural

Networks. In: J. Mira and J. Cabestany, (eds), From Natural to Artificial Neural Computation. International Workshop on Artificial Neural Networks, Volume 930 of Lecture Notes in Computer Science, 1999, pp. 597–604.
[4] Cunningham, S.J., Holmes, G., Littin, J., Beale, R., Witten, I.H.: Applying Connectionist Models to Information Retrieval. In: S. Amari, and N. Kasobov (eds.), Brain-Like Computing and Intelligent Information Systems, Springer-Verlag, 1997, pp. 435-457.
[5] Ontological engineers handbook. http://www.cyc.com/doc/handbook/oe/oe-handbook-toc-opencyc.html
[6] Furdík, K.: Information Retrieval in Natural Language by Hypertext Structures. [PhD thesis], FEI TU Košice, 2003, (in Slovak).
[7] Gruber, T. R.: A Translation Approach to Portable Ontology Specifications. Knowledge Acquisition, 1993.
[8] Gurney, K.: An Introduction to Neural Networks. ISBN 1-85728-503-4, UCL Press, 1997.
[9] Chen, H.: Machine learning for Information Retrieval: Neural Networks, Symbolic Learning, and Genetic Algorithms. Journal of the American Society for Information Science, 46 (3), 1995, pp. 194 - 216.
[10] Vector Space Model (VSM). http://isp.imm.dtu.dk/thor/projects/multimedia/textmining/node5.html
[11] Kohonen, T., Hynninen, J., Kangas, J, Laaksonen, J.: SOM-PACK - The Self Organizing Map Program Package. Helsinki, University of Technology, Finland, 1996.
[12] Motta, E.: Reusable Components for Knowledge Models: Principles and Case Studies in Parametric Design. IOS Press, Amsterdam, 1999.
[13] OKBC home page. http://www.ksl.standford.edu/software/OKBC/
[14] Páleš, E.: Sapfo - Slovak Para – Phraser. 1. issue. ISBN 80-224-0109-9, VEDA, Bratislava, 1994, (in Slovak).
[15] Protégé user guide, 2000. http://protege.stanford.edu/publications/UserGuide.pdf
[16] Raghavan, V. V., Wong, S. K. M.: A Critical Analysis of Vector Space Model for Information Retrieval. Journal of the American Society for Information Science, Vol.37 (5), 1986, pp. 279-87.
[17] Sensus.http://www.isi.edu/natural_language /projects /ontologies.html
[18] Sowa, F. J.: Knowledge Representation: Logical, Philosophical, and Computational Foundations. Brooks Cole Publishing Co., Pacific Grove, CA, 2000.
[19] SUMO. http://ontology.teknowledge.com
[20] OWL - Web Ontology Language Overview. http://www.w3.org/TR/owl-features/
[21] WordNet. http://www.cogsi.priceton.edu/~wn/

BIOGRAPHY

Igor Mokriš (prof) was born in 1948. He received MSc degree in technical cybernetics from the Faculty of Electrical Engineering, Technical University Košice in 1972, PhD degree in technical cybernetics from the Faculty of Electrical Engineering, Slovak Technical University Bratislava in 1980 and he became a professor in technical cybernetics from Military Academy Liptovský Mikuláš in 1997. He is scientist in Institute of Informatics, Slovak Academy of Sciences Bratislava, Slovakia. His current research interest is oriented into the knowledge systems and neural networks.

Lenka Skovajsová was born in 1979. She received MSc degree in telecommunications from the Military Academy in Liptovský Mikuláš in 2003. Since 2004 she is studying her PhD. study in the Institute of Informatics, Slovak Academy of Sciences Bratislava, Slovakia in the field of information retrieval and neural networks.