Development of Neural Network Information Retrieval System from Text Documents

Igor Mokriš, Lenka Skovajsová

Institute of informatics, Slovac academy of sciences, Bratislava
Office Liptovský Mikuláš
03104 Liptovský Mikuláš
mokris@valm.sk, skovajsova@valm.sk

Abstract: The aim of the paper is to describe the information retrieval model which retrieves information from text documents in natural language by neural networks. This model comes from the linguistic and conceptual approach for the analysis of text documents. The neural network model accepts the structure of conceptual, lingvistic oriented model, where the problem of document database creation and document indexing for keyword determining is solved. Query entering uses the same mechanisms as document formalization for example document database creating method. Proposed structure of neural network model loses the problem of document retrieval on the base of user question. However, learning algorithms and neural network invariancy is used by usage of neural networks, it is possible to decrease the complexity of language analysis algorithm computatation.

Keywords: information retrieval, queries, keywords, text documents, neural networks, natural language

1 Introduction

The aim of this paper is to describe information retrieval system which retrieves information from the text documents in natural language and comes from the information retrieval system using statistical, conceptual and linguistic model. The aim of contribution is to develope this system in natural language by neural networks.

With growing the number of information in the document space, there is also growing the request of effective information retrieval in this space. Therefore it was developed many different information retrieval systems. This paper shows the possibility to simplify the computational complexity of information natural language retrieval systems with neural networks.

Slovak language is, despite many languages, hard to process on computer. Nouns, adjectives, pronouns, numerals are differently inflected and verbs are differently

timed. It is therefore difficult to recognize keywords in slovak text. Keywords are words, which take part on document indexing. Next problem arises with synonyms, the words with different shape. Another problem arises with homonyms with different meaning and another problem arises with phrases with more than one word and so on. Situation is also more complicated with varied text structures which aren't based on words, for example dates or various numberings, dates, etc..

For the slovak text analysis various approaches are used. Most often used of them is statistical approach, linguistic approach and knowledge based approach. Statistical approach analyses words in the text by comparing them with keywords. The keyword set can be made before indexing, or is created directly from documents [11].

Lingvistic approach extracts linguistic units from text documents [8]. Lingvistic unit can be phoneme, morpheme, lexeme and so on. Linguistic units are extracted from words in texts. Linguistic text analysis consists of partial analyses meanly language plains, phonological, morphological, syntactical, semantical and pragmatical analysis. The result is the thorough representation of texts, where every relevant data from the point of content are signed. There are clearly identified language units and relationships between them.

Knowledge based approach uses concepts or parts of text associable with context for document indexing [6, 15]. Knowledge model can be defined as a formal description of documents, concepts and relationship between them with accent on their semantics. Concepts create a structure with related documents in the given domain. This model is named as a domain model. Knowledge based approach uses mainly semantic nets, existencial graphs, conceptual graphs and ontologies. More detailed description of ontologies can be found in [1,16] and description of basic systems used for ontology creation can be found in [3,9,12,13]. Aggregated despription of ontological languages can be found on the world wide web [10, 17].

From the upper reasons it is advantageous to use neural networks for information retrieval system [14,19,20,21,22]. Well trained neural network is able to better determine stems in the words so there aren't problems with text language analysis. Besides this, the neural network is in their life phase faster as a systems not using the neural networks. The system proposed with neural networks can also have some disadvantages and they are the time of training the neural networks, when new documents are added to the document base.

Information retrieval system model with neural networks comes from the model based on statistical, linguistic and knowledge based approach.

2 Statistic, Linguistic and Knowledge Information Retrieval System

Proposed information retrieval system serves for information extraction from text documents in natural Slovak language. These documents are stored in document base. Each of them is marked with its index, which expresses the document content and the document relevance. User enters the question for that system and system returns him a document subset relevant to his query.

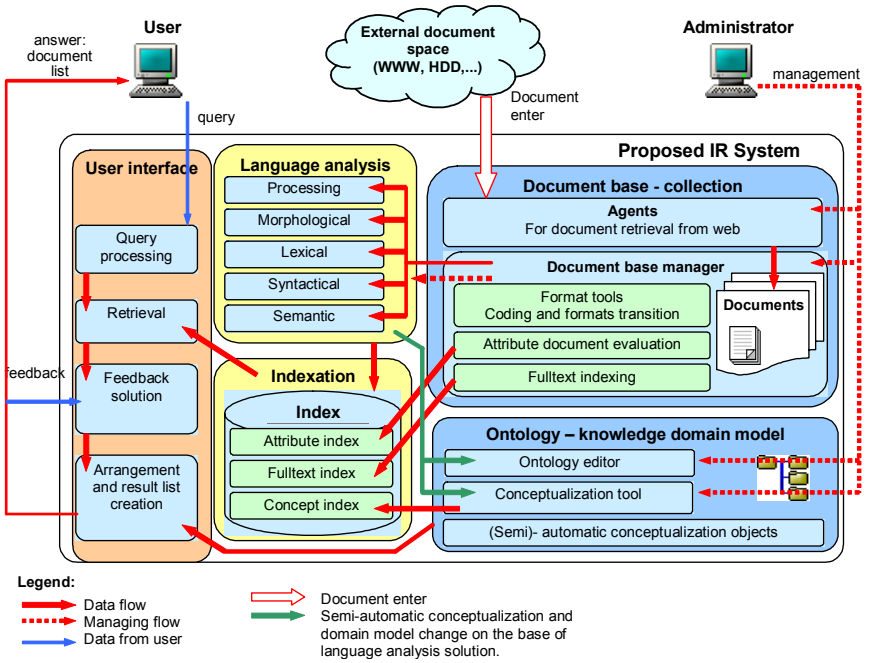

Fig. 1. Modular structure of information retrieval system

In fig. 1, there is the proposed information retrieval system structure based on the statistical, linguistic and knowledge model. The structure of the system is modular and has separated document space from question and keyword space. There is document base module, language analysis module, user interface module, indexation module and knowledge based domain model.

In the document base there are the documents stored and each of them is represented by index. This index consists of the concept vector where the weights are made also by a structure of domain knowledge model. Besides that, each document is represented by attribute set, which also characterizes given document. Attributes can be document author , date of creation and so on.

Language analysis of texts, if proposed modularly, consists of data modules, which are marked also as a dictionaries. Then there are modules marked as partial

analysis (phonological, morphological, lexical, syntactical, semantical and pragmatical) of text documents and it influences also its structure of a domain model.

User interface module serves for entering question by user, its processing by indexation module, knowledge domain model, and the result returning model. It consists of query formulation, query processing module, searching module, module of relevance feedback, and relevant document list module creation. Then the searched documents are given to user as a result.

Indexation module serves for searching of relevant documents by query. Searching in given system is possible by three ways:

- conceptual retrieving - uses concept and domain model,
- attribute searching – attributes are indexed in attribute index in relational database,
- fulltext searching – using statistical index list (term vector)

Conceptual searching is searching with concept vector which belong to each document as an index.

Fulltext searching is searching with keywords using keyword and document matrix. Each document in this matrix is represented by the keyword vector. Each keyword belonging to given document is expressed by the relative frequency of a keyword in the document [5]. This representation is called the vector space model.

3 Neural network information retrieval system

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

Administrator subsystem guaranties document set operations. Administrator determines document set due to creation of document base from them. Document base manager then provides the system representation of documents. He also determines suitable model of document storing and creates document base of system representation. Indexation subsystem solves two tasks. Firstly it is creation of index and secondly it is creation of question representation that is comparable with document index.

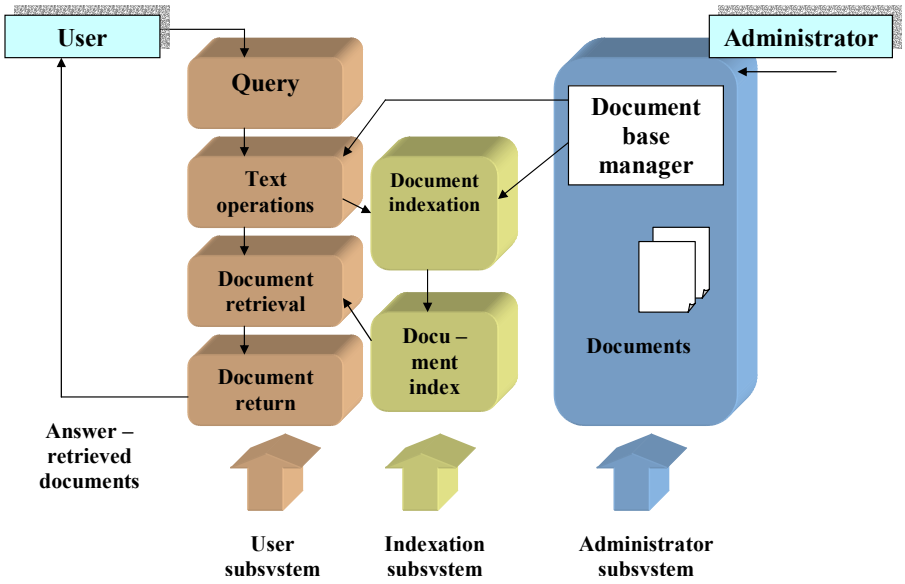

Fig. 2. Simplified information retrieval system structure

User subsystem processes user query and searches for the relevant documents. Firstly, user defines a question. User subsystem processes this query and assigns it a keyword as a result. Then indexation subsystem indexes a query which is compared with document index and, on the base of this comparison, administrator subsystem retrieves relevant documents and send them to user as a result. The user can use relevance feedback where user marks the most relevant documents from result and sends it as a new query. System creates a new question from these documents and searches them again.

These three subsystems of information retrieval system can be representented as a three layer model on the fig. 3.

The first sublayer of this system is a query sublayer, the second sublayer is the keyword sublayer and the third sublayer is the document sublayer. User puts a query, this is transformed into a keyword, on the base of which a relevant documents are retrieved from document sublayer.

Transition from the query sublayer into keyword sublayer and transition from the keyword sublayer to document sublayer can be substituted by neural networks as is depicted in fig. 4 [2, 1, 4].

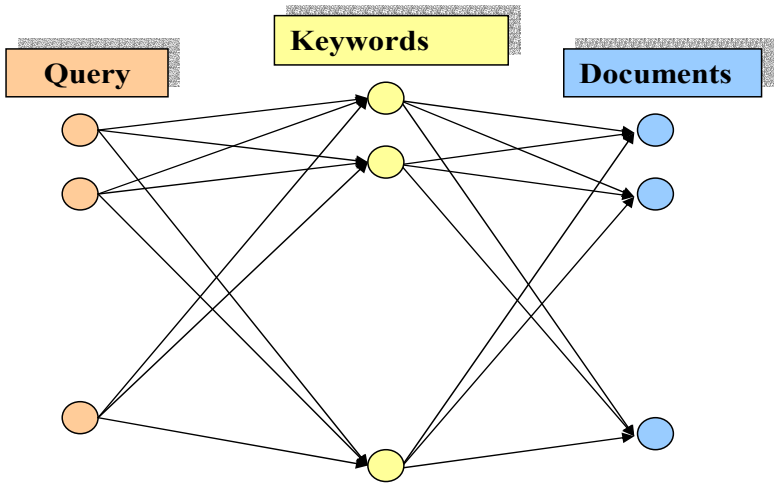

Fig. 3 Layer information retrieval system structure

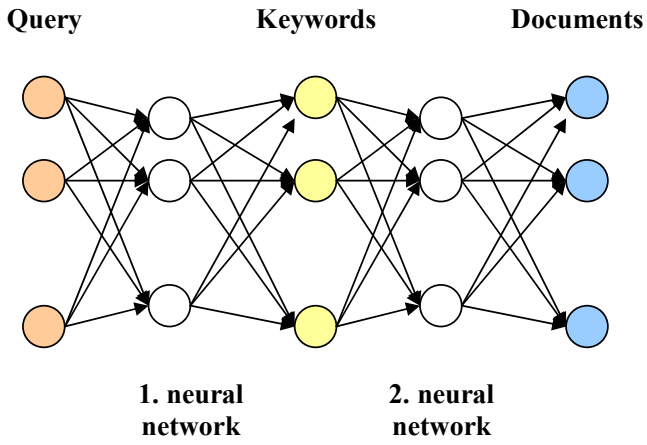

Fig. 4. Neural network information retrieval model

3 Description of Developed Information Retrieval System

Retrieved documents are encoded as a word by word and brought on input of first neural network[18]. Then it is determined by this neural network whether the given word in the document is the keyword or not. If it is, his occurrence is added

into vector space model matrix[5]. The result of that is normalized vector space model matrix. The weights in this matrix are matched with weights of second neural network, which has as an input some keywords and as an output the base of documents. Joining these two neural networks, the information retrieval system is developed as is depicted in fig. 5.

Then the input interface is created, which enables the user to enter a query. And then the information retrieval system enables to find the relevant documents. At the end the output interface is created, which sorts the relevant documents and sends them to user as a result.

In case when it is necessary to give as a query which consists of two or more words, it is necessary to add more input neurons groups into first neural network where each group represents one word. Then it is each word separately created and there are created coherent keywords to each query. This structure enables giving more keywords as an input and also enables giving as input the word connections into neural network.

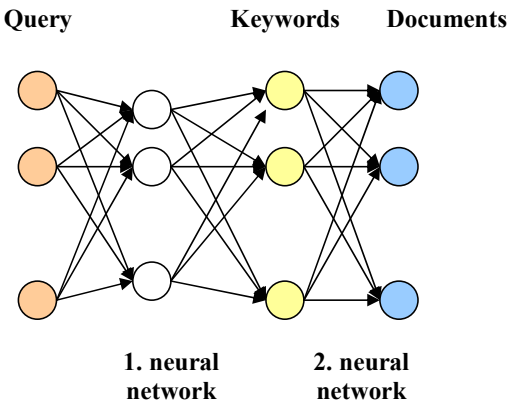

Fig.5 Developed information retrieval model

Conclusion

For development of the neural information retrieval system were used multilayer perceptrons. First of them was trained by a set of keywords and second one was associated with a set of documents. On the base of this model it was developed MATLAB program which uses word length of 12, 13 keywords and as the document base 90 documents were used. In each experiment in this model the relevant documents were found in the case of correct keyword and no documents were found in the case of incorrect keyword. However, used neural networks are invariant against the word changes, this property was expressed in document retrieval.

References

[1] Back, H., Pinto, H., S.: Overview for Approach, Methodologies, Standards, and Tools for Ontologies. The Agricultural Ontology Service UN FAO. p. 3-5, 2001 – 2002.

[2] Cunnigam, S., J. A kol.: Applying Connectionist Models to Information Retrieval. The University of Waikato, Hamilton, New Zeland, p. 3-6.

[3] Decker, S.et al.: The Semantic Web on the Respective Roles of XML and RDF. Stanford University, USA, p. 4-7, IEEE00.pdf

[4] Furdík, K.: Information Retrieval from Natural Language by Hypertext Structures. [PhD Thesis], FEI TU Košice, 2003, (in Slovak).

[5] Intelligent Signal Processing Group. http://isp.imm.dtu.dk/thor/projects/multimedia/textmining/node5.html

[6] Mokriš, I., Skovajsová L.: Possible Approach for Utilization of Neural Networks for Information Retrieval from Natural Language. Proc. of Conference Informatics and Information Technologies, UMB Banská Bystrica, 2004, 6 p, (in Slovak).

[7] Niles, I., Pease, A.: Towards a Standard Upper Ontology. Teknowledge Corporation, 2001.

[8] Páleš, E., et al.: Slovak Para – Phraser. 1. vyd. Bratislava: VEDA, 1994, ISBN 80-224-0109-9, (in Slovak).

[9] Phytila, C.: An Analysis of the SUMO and Description in Unified Modeling Language.Phytila-SUMO.htm 2002.

[10] Protégé. www.protege.stanford.edu

[11] Raghavan, V. V. and 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, p. 279-87.

[12] Sklenák, V.: Retrieval Tools in Internet Space – beyond Towards? VŠE Praha. 2003, (in Czech).

[13] Sklenák, V.: Semantic Web. VŠE Praha. 2003 , (in Czech).

[14] Mandl, T.: Tolerant and Adaptive Information Retrieval with Neural Networks. EXPO 2000. University of Hildesheim – Germany.1999, p. 2-4.

[15] Sowa, F.: Knowledge Representation: Logical, Philosophical, and Computational Foundations. Brooks Cole Publishing Co., Pacific Grove, CA, 1999.

[16] Svátek, V.: Ontology and www. VŠE Praha. 2002, (in Czech).

[17] W3C. www.w3c.org

[18] Gurney K.: An Introduction to Neural Networks, UCL Press, 1997, ISBN-1 -85728-503-4.

[19] Hsinchun C.: Machine Learning for Information Retrieval: Neural Networks, Symbolic Learning and Genetic Algorithms. University of Arizona, 2001, p.5-6

[20] Skopal T.: Neural Networks and Information Retrieval. TU Ostrava, p. 2-9.

[21] Crestani F.: Neural Relevance Feedback for Information Retrieval. International Computer Science Institute, Berkeley, 2000, p.2,3.

[22] Jan van den Berg, Schuernie M.: Information Retrieval Systems using Associative Conceptual Space, Europian Symposium on Artificial Neural Networks, 1999, ISBN 2-600049-9-X, pp. 351-356