VOL. 13, NO. 21, NOVEMBER 2018 ISSN 1819-6608
ARPN Journal of Engineering and Applied Sciences

www.arpnjournals.com

CASCADE NEURAL-FUZZY MODEL OF ANALYSIS OF SHORT ELECTRONIC UNSTRUCTURED TEXT DOCUMENTS USING EXPERT INFORMATION

Dmitry Tukaev[1], Olga Bulygina[1], Pavel Kozlov[1], Anatoly Morozov[2] and Margarita Chernovalova[1]
[1]Smolensk Branch of National Research University "MPEI", Russian Federation
[2]Smolensk Branch of Financial University under the Government of the Russian Federation, Russian Federation
E-Mail: baguzova_ov@mail.ru

ABSTRACT

The goal of this article is to increase the efficiency of analyzing small electronic unstructured text documents in conditions of a statistical data lack for using the probabilistic methods that should be based on the application of neural-fuzzy instruments. This paper suggests a cascaded neural-fuzzy model using expert information to determine the importance of meaningful words in the formalization and subsequent rubrication of text documents based on the neural-fuzzy classifier, which allows analyzing small documents based on their unified representation. The practical use of the results is expected in creating information systems of the automated analyzing electronic unstructured text documents used in state and municipal government.

Keywords: electronic unstructured text documents, text rubrication, neural-fuzzy classifier, cascade model, short document analysis.

INTRODUCTION

Currently, Internet-portals and web-applications for state and municipal governments, which provide two-way communication with the public and legal entities, are an actively developing IT-sector.

Generally each similar portal has the function of electronic reception that allows individuals or organizations to send appeals in electronic form. These appeals have a number of specific characteristics that allow attributing them to electronic unstructured text documents (EUTDs).

At the same time, these features do not allow using the well-known methods of text analysis without significant development (Kozlov, 2015). It determines the need to create new methods of processing and rubricating textual information.

The intellectual analysis methods, which allow receiving informed solutions in conditions of a static data lack, can be considered as the perspective way to solve this task (Gimarov *et al.*, 2004; Bulygina *et al.*, 2016).

The article suggests the use of neural-fuzzy methods which gives the possibility of reducing the degree of subjectivity inherent in the models based on expert assessments.

The goal of this article is to increase the efficiency of analyzing small electronic unstructured text documents in conditions of a statistical data lack for using the probabilistic methods that should be based on the application of neural-fuzzy instruments.

MATERIALS AND METHODS

Currently, world information environment contains a huge number of various types of EUTDs written in natural language, which are sources of data and knowledge in various areas of human activity (Bevainyte *et al.*, 2010; Dli M.I. *et al*, 2017). At the same time, the number of such EUTDs is constantly increasing. That fact determines the need to accelerate development of information systems of automated analysis of these documents (Sebastiani, 2002).

Often, the functional of such information systems is limited to separate subject areas, since the systems work with a certain group of concepts (Khapaeva, 2002). From this point of view they are the "closed" systems because of difficulty to make any changes (the number of rubrics, the composition of the thesaurus and the importance of words).

Based on analysis of modern information systems used by the authorities, it is possible to conclude that there are no effective tools to solve the problems of analyzing electronic text documents in conditions of temporary change of the rubrics.

One of the most common challenges faced by such systems is text mining of highly specialized text arrays (various reports, survey results, etc.). In large arrays of text documents, in which a set of vocabulary is limited, new information is accurately extracted on basis of statistics of using the meaningful words. A method of text document clustering based on the meaningful word analysis is considered in the paper (Shmulevich, 2009).

For the unstructured text documents, it is necessary to use the procedures of "understanding" of arbitrary texts written in natural language (Andreev *et al.*, 2003). This task is one of the "oldest" problems of artificial intelligence that can be solved by several approaches (Borisov *et al.*, 2016; Dli M.I. *et al*, 2017).

Examples of these approaches can be methods of data processing in natural language - NLP (Natural Language Processing), neural network, etc. Papers (Shemenkov, 2009; Meshkova, 2009; Korzh, 2000) suggest the models of neural network classifiers with methods of formalizing the text documents and representing the results of the classifier in the form of semantic images.

Due to the feature of EUTDs (namely, complaints, appeals, proposals, etc.) entering the Internet-

portals of the authorities and the need to rubricating them under special conditions, when developing algorithmic and software of information systems, it is advisable to use several rubrication models depending on characteristics of specific EUTD.

For example, when rubricating short EUTDS in the presence of a sufficient amount of statistical data and an insignificant degree of rubric thesaurus intersection, it is advisable to use neural network algorithms, in particular the neural-fuzzy classifier (Kruglov *et al.*, 2001).

A short EUTD is a text document written in natural language and containing information in linguistic or digital form. Its volume does not allow applying the well-known procedures of statistical text analysis, but permits the use of expert information obtained as a result of combining the knowledge of linguists and specialists in the considered subject areas (Kozlov, 2017).

When using a neural-fuzzy classifier, the EUTD is represented as a huge array of binary values, each of which corresponds to the presence or absence of all the words from the thesaurus of the entire rubric field. Such representation of EUTD makes it irrational to apply this rubrication model under the conditions of dynamically changing rubric thesauri because of the complexity of rebuilding the neural-fuzzy network and the approach of formalizing the EUTD each time when changing rubric composition.

Therefore, it is necessary to develop a method of EUTD formalization that will make it possible to use the neural-fuzzy classifier under the dynamic rubric thesaurus and also to present the model as a cascade to convenient rebuild the entire rubrication model when the rubric field changes.

RESULTS AND DISCUSSIONS

Taking into account the EUTD features, a cascaded neural-fuzzy model of the document rubrication that allows analyzing small documents on the basis of their unified representation is developed. Figure-1 shows the proposed cascade neural-fuzzy model of rubricating the short EUTDs.

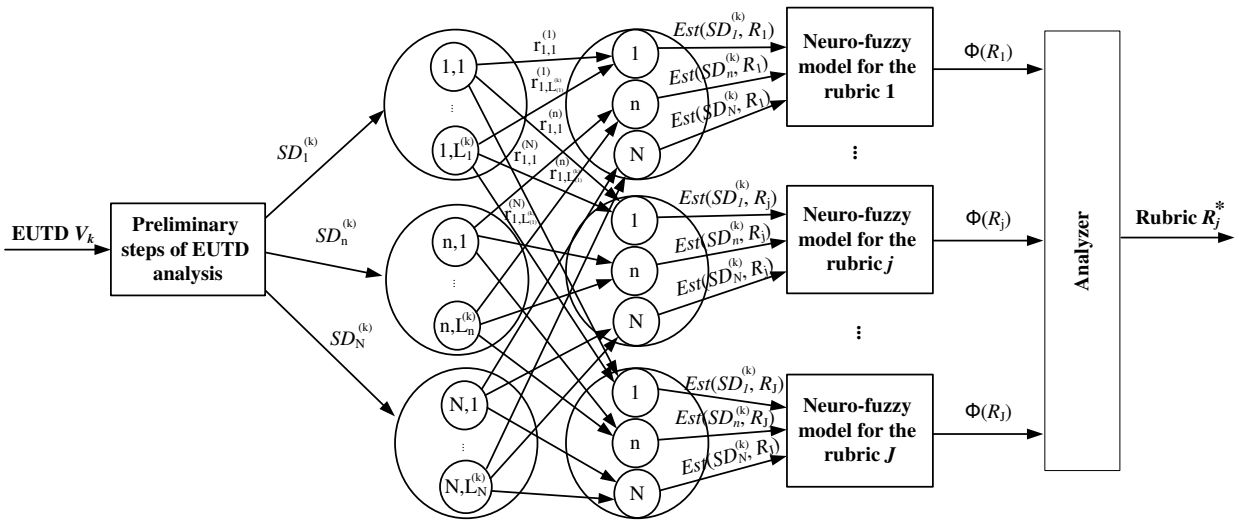

Figure-1. The structure of the cascade neural-fuzzy classifier for rubricating the short EUTDs.

The proposed cascade neural-fuzzy model of the EUTD rubrication includes the following submodels:

a) A model of the preliminary EUTD analysis using the syntactic parser

The preliminary analysis includes the following procedures:

- lexical analysis (dividing words, punctuation marks, numbers and other text units);
- morphological analysis (determining grammatical characteristics of lexemes and basic word forms);
- syntactic analysis (identifying the sentence structure).

In the process of using known software products to carry out additional stages of analysis, developers will have to face the problem of the diversity of linguistic markings. For example, most of the syntactic parsers represent each sentence of the text in the form of dependency trees that are described by linguistic markup. Linguistic markings must be modified for further classification and assignment of weight coefficients, as a result of which the metric dimension will be increased.

This model is intended for forming the sets of meaningful words of the EUTD, characterized by the same syntactic role in the sentences.

The EUTD V_k arrives at the input, and a set of syntactic groups is generated at the output:

$$SD_k = \left\{ SD_n^{(k)} \mid n = 1..N \right\},$$

where $SD_n^{(k)}$ - a set of the words corresponding to the syntactic parameter n, N - number of syntactic groups.

b) A model of formalizing the EUTD using weighting coefficients includes two procedures

a) comparing meaningful words $v_p^{(k)}$ of each syntactic group $SD_n^{(k)}$ with the database of weighting coefficients (the degree of influence of meaningful words relative to each rubrics is formed at the output);

b) accumulating and normalizing the weighting coefficients (estimates of the degree of belonging of the syntactic groups $SD_n^{(k)}$ to all rubrics are determined at the output).

The procedure of constructing the model using weight coefficients includes the following steps:

Step 1. The initial set of rubrics is determined:

$$R = \left\{R_j \middle| j = 1..J\right\},$$
$$R_j = \left\{\left(w_{m_j}^{(j)}, c_{m_j}^{(j)}, f_{m_j}^{(j)}, p_{m_j}^{(j)}\right) | \, m_j = 1..M_j\right\}.$$

where $w_{m_j}^{(j)}$ - the word m_j in the rubric R_j, $c_{m_j}^{(j)} \in [0,1]$ - the degree of compliance of the word m_j with the rubric R_j, $f_{m_j}^{(j)}$ - the frequency of occurrence of the word m_j in the rubric R_j $p_{m_j}^{(j)}$ - the threshold of using the word m_j in the rubric R_j, J - the number of rubrics, M_j – the total number of meaningful words in the rubric R_j.

Step 2. A set of EUTDs with predefined rubrics is defined:

$$V^{(tr)} = \left\{\left(V_b^{(tr)}, RR_b\right) | \, RR_b \in R\right\},$$

where $V^{(tr)}$ - the training sample, $V_b^{(tr)}$ - the EUTD from the training sample, RR_b - the rubric corresponding to the EUTD $V_b^{(tr)}$ from the training sample.

The meaningful words $v_{l_b}^{(tr)}$ that are longer than three characters are searched in these EUTDs.

As a result, the EUTD $V_b^{(tr)}$ can be represented in the following form:

$$V_b^{(tr)} = \left\{v_{l_b}^{(b)} \middle| l_b = 1..L_b\right\},$$

where L_b – the number of meaningful words of EUTD $V_b^{(tr)}$.

Step 3. Each word $v_{l_b}^{(b)}$ of the EUTD receives an initial weighting coefficient $u_{l_b}^{(b)} = 0{,}5$ that indicates degree of its compliance with rubric R_j which is related to the EUTD $V_b^{(tr)}$. Thus, we obtain a set of pairs of the following form:

$$V_b^{(tr)} = \left\{\left(v_{l_b}^{(b)}, u_{l_b}^{(b)}\right) \middle| l_b = 1..L_b\right\}$$

Step 4. The adjustment of the model's weighting coefficients is carried out using the training sample $V^{(tr)}$. As a result, the weights $u_{l_b}^{(b)}$ of the meaning words are changed according to the degree of their compliance with the particular rubric R_j.

At the output, the dictionaries of the rubric R_j are formed:

$$R_j = \left\{\left(w_{m_j}^{(j)}, r_{m_j}^{(j)}\right) | \, m_j = 1..M_j\right\},$$

where $w_{m_j}^{(j)}$ – the word m_j in the rubric R_j, $r_{m_j}^{(j)} \in [0,1]$ – the weighting coefficient of the word m_j in the rubric R_j, M_j - the total number of meaningful words in the rubric R_j.

Step 5. Since the weighting coefficients for the neural-fuzzy classifier are taken into account in the absence of a large amount of the training sample (it is due to dynamics of the rubric field), correction of the weight coefficients $r_{m_j}^{(j)}$ is carried out at all stages by experts.

The procedure of applying the described model to construct the neural-fuzzy rubrication model includes the following steps.

Step 1. The unification of the set of the EUTD parameters is carried out:

$$S = \left\{s_n | \, n = 1..N\right\}.$$

Using the syntactic parameters, each EUTD V_k is represented as:

$$V_k = \left\{\left(v_{l_k}^{(k)}, h_{l_k}^{(k)}\right) | \, h_{l_k}^{(k)} = s_n, l_k = 1..L_k\right\},$$

where $v_{l_k}^{(k)}$ - the word l_k of the EUTD V_k, $h_{l_k}^{(k)}$ - syntactic parameter characterizing the word l_k, L_k - number of the words in the EUTD V_k.

Each EUTD V_k is assigned a set of the syntactic groups SD_k:

$$SD_n^{(k)} = \left\{v_p^{(k)} | \, \forall p = 1..L_n^{(k)}, h_{l_k}^{(k)} = s_n\right\},$$

where $L_n^{(k)}$ - the number of the words of the set n in the EUTD V_k.

Step 2. Matching of the set SD_k with the rubric R_j is carried out:

$$SD_k \leftrightarrow \left\{R_1, ..R_j, ..R_J\right\}.$$

To do this, many assessments are carried out:

$$\forall j \in J : Est(SD_k, R_j) = \left\{Est(SD_n^{(k)}, R_j) | \, n = 1..N\right\},$$

$$Est(SD_n^{(k)}, R_j) = \frac{1}{L_n^{(k)}} \cdot \sum_{p=1}^{L_n^{(k)}} u_p^{(k)},$$

$$u_p^{(k)} = r_{m_j}^{(j)} \mid w_{m_j}^{(j)} = v_p^{(k)},$$

where $u_p^{(k)}$ – the weighting coefficient of the meaningful word $v_p^{(k)}$ of the EUTD V_k for the rubric R_j, $r_{m_j}^{(j)}$ - the weighting coefficients of rubric's meaningful words configured for the model using weighting coefficients.

As a result, the set $Est(SD_k, R_j)$ input to the neural-fuzzy classifier for the rubric R_j.

The effectiveness of the proposed approach to the EUTD formalization insignificantly depends on the number of meaningful words contained in it. This fact makes it possible to use the neural-fuzzy classifier to rubricate documents of different volumes without changing its structure.

c) A set of the neural-fuzzy models of assessment of belonging to particular rubrics.

Each model is designed to form the degree of belonging of the EUTD to the particular rubric R_j.

Particular models are three-layer hybrid neural-fuzzy networks.

The values of the EUTD parameters in form $Est(SD_k, R_j)$ are input into elements of the first layers.

The elements of the model's second layers realize the fuzzy activation functions for the output rules that evaluate the effect of the analyzed word on the rubric definition and represent the term sets corresponding to the values ("weak", "medium" and "high" influence).

The elements of the model's third layers realize the calculation of the minimum functions over all input values while the number of neurons of these layers is 3^N.

The fourth layers consist of J elements, each of which realizes the maximum function.

As a result, the degree of belonging of the EUTD to the appropriate rubric R_j is determined at the output of each particular model.

d) A model of selecting the rubrics that are the most appropriate for the analyzed EUTD.

This model is designed for the final selection of the rubrics which the EUTD belongs.

Outputs of all neural-fuzzy models are fed to the analyzer that allows determining the rubric R_j^* which the EUTD belongs:

$$R_j^* : \max_{j=1..J} \Phi(R_j),$$

where $\Phi(R_j)$ - non-linear transformation (for example, sigmoidal form) to determine the degree of belonging of the EUTD to the rubric R_j.

Figure-2 shows the generalized procedure of using the neural-fuzzy classifier for the EUTD rubrication in the information system.

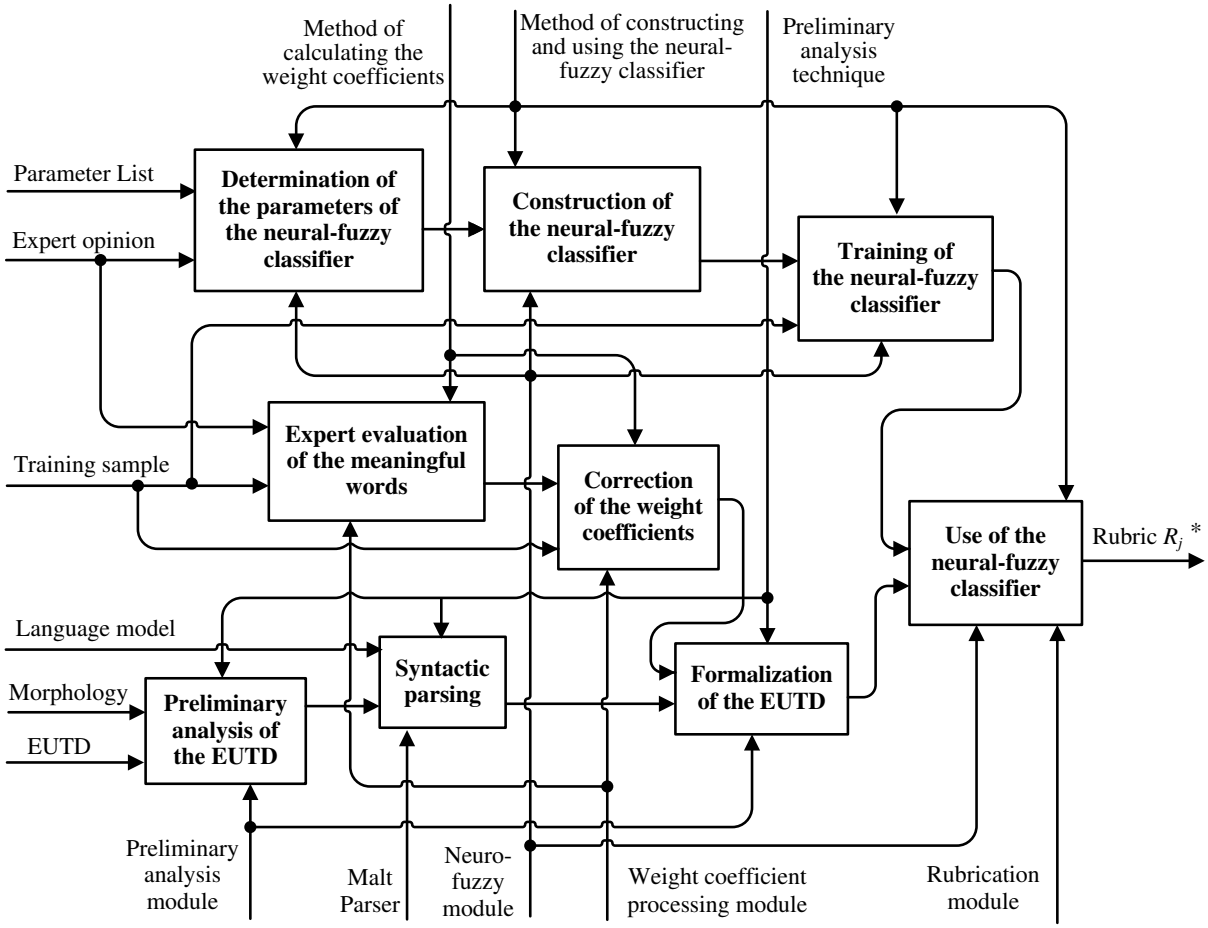

Figure-2. The generalized procedure of using the neural-fuzzy classifier for the EUTD rubrication.

CONCLUSIONS

The article suggests the cascade neural-fuzzy model of rubricating the short electronic unstructured text documents taking into account the determination of significance of the meaningful words during their formalization for subsequent analysis based on the neural-fuzzy classifier.

This model allows rubricating the short electronic unstructured text documents in conditions of a statistical data lack for using probabilistic classifiers.

The foregoing leads to the conclusion about the topic relevance and the prospects of practical application of the research results in developing the information systems used in state and municipal government.

ACKNOWLEDGEMENTS

The reported study was funded by RFBR according to the research project № 18-01-00558.

REFERENCES

Andreev A.M., Berezkin D.V., Suzev V.V., Shabanov V.I. 2003. Models and methods of automatic classification of text documents. Herald of the Bauman Moscow State Technical University. Series Instrument Engineering, no.3.

Bevainyte A., Butenas L. 2010. Document classification using weighted ontology. Materials Physics and Mechanics, no. 9.

Borisov V.V., Dli M.I., Zernov M.M., Fedulov A.S. 2016. Method of time series analysis using scenarios. International Journal of Applied Engineering Research, no. 11(21): 10536-10539.

Bulygina O.V., Okunev B.V. 2016. Creating fuzzy network tools to analyze prospects of projects of information and telecommunication infrastructure development. Neyrokomp'yutery. (7): 15-20.

Dli M.I., Zaenchkovski A.E., Tukaev D.A., Kakatunova T.V. 2017. Optimization algorithms of the industrial clusters' innovative development programs. International Journal of Applied Engineering Research, no. 12(12): 3455-3460.

Dli M.I., Ofitserov A.V., Stoianova O.V., Fedulov A.S. 2016. Complex model for project dynamics prediction. International Journal of Applied Engineering Research, no. 11(22): 11046-11049.

Gimarov V.A., Dli M.I. 2004. Neural network algorithm of complex object classification. Programmnye produkty i sistemy. (4): 51-56.

Khapaeva T. 2002. Automatic classification of documents. Softerra, No. 2.

Korzh V.V. 2000. Methods of coding textual information for building neural network document classifiers: PhD thesis.

Kozlov P.Yu. 2015. Comparison of frequency and weight algorithms of automatic document analysis. Nauchnoye obozreniye. (14): 245-250.

Kozlov P.Yu. (2017). Automated analysis method of short unstructured text documents. Programmnye produkty i sistemy, no. 1, pp. 100-105.

Kruglov V.V., Dli M.I., Golunov R.Yu. 2001. Fuzzy logic and artificial neural networks. Moscow: Nauka, Fizmatlit.

Meshkova E.V. 2009. Development and research of hybrid neural network models for the automatic classification of text documents: PhD thesis.

Sebastiani F. 2002. Machine learning in automated text categorization. ACM Computing Surveys. 34(1): 1-47.

Shemenkov P.S. 2009. Development and research of model of neural network method of text document analysis: PhD thesis.

Shmulevich M.M. 2009. Methods of automatic text clustering based on extracting the objects' names from the texts and subsequent constructing the graphs of the joint occurrence of key terms: PhD thesis.