Multimodel method of rubricating the unstructured electronic text documents

M I Dli[1], O V Bulygina[1] and P Yu Kozlov[1]

[1] The Branch of National Research University "Moscow Power Engineering Institute" in Smolensk, Smolensk, Russia

Abstract. Analysis of electronic text documents written in natural language is one of the most important tasks implemented in systems of automated analyzing the linguistic information. It is known that text documents can be characterized by different parameters: size, presence of a structure, frequency of references keywords, etc. Depending on these parameters, a procedure of selecting the type of the rubrication model is proposed. As a result, a multi-model approach to document classification, characterized by the combined use of artificial neural networks, growing pyramidal networks and probabilistic-statistical methods, is proposed. Its application will improve the accuracy of attributing the electronic text documents to concrete rubrics, taking into account their specificity and various objectives of practical application in the organization.

1. Introduction

The task of unstructured information processing is one of the priority problems, since its solution is necessary to achieve the goal of transition to a digital economy, formulated in Strategy of the Information Society Development in the Russian Federation for 2017-2030. The program "Digital Economy", approved by the Government in July 2017, is considered as one of the most important tasks to improve accessibility and quality of public services for citizens. To solve this task it is necessary to improve the procedures of electronic information exchange between citizens and public services. The effectiveness of this interaction is assessed using such indicators as operativeness of reaction to the citizen requests and satisfaction with the results of provided services. Both indicators directly depend on the processing speed of incoming information requests and the accuracy of message recognition.

The fact that information in the citizens' queries is presented in unstructured form does not allow using classical methods of processing relational, object, hierarchical and marked text data. The methods of semantic search can be recommended for the analysis of unstructured text information. But their application is complicated by the specifics of information messages containing citizens' requests and the peculiarities of functioning of the state services processing these messages.

The foregoing stipulates the need to develop methods of analyzing the electronic unstructured text documents (EUTDs), taking into account the specifics of its content and usage in the system of electronic public services.

2. Selection of an approach to the EUTD rubrication

2.1. The concept of an electronic unstructured text document

The main features of EUTD are the following:

- the text is written in natural language;
- there is no explicitly defined structure;
- there are the grammatical and syntactic errors in the text;
- automatic allocation of a structure can't be performed in an unambiguous way.

The examples of unstructured text document types can be: books; magazines; metadata; medical records; audio and video materials; analog data; Images; files based on unstructured text (e-mail messages, web pages, documents created using word processors) 0.

EUTDs can include documents in such formats as html, doc, rtf, etc. Usually the content of such EUTDs is presented to the user for examination by the Internet browser or any text editors (e.g. Microsoft Word).

2.2. Analysis of the EUTD rubrication methods

Currently, information systems can use a sufficiently developed methodological apparatus of the theory of classifying the objects of various types (in this case a document is regarded as an object of classification or rubrication) 0.

By the criterion of learning stage existence in the EUTD rubrication model, all methods can be conditionally divided into two groups: with and without training.

The EUTD rubrication methods with a teacher suggest the availability of complete information about the rubricating field (number and characteristics of columns). In this case, a "teacher" is a prepared sample of EUTDs (training set) with the concrete degree of belonging to one or another rubric.

Currently, the rubrication methods with a teacher are widely used in information systems and are suitable for solving a large number of emerging tasks. However, these methods require a large amount of preliminary information for teaching the rubrication model.

The rubrication methods without a teacher analyze the collection of EUTDs for classifying them in such a way that each rubric contains documents closest to the chosen metrics. In these methods there is no need for initial training of the models "with a teacher", since the characteristics of the rubric field of text documents are not known in advance.

In general, the rubrication algorithms without a teacher (clustering algorithms) group EUTDs in the feature space (document parameters) with the subsequent interpretation of the result. The EUTD classification is based on the hypothesis that text documents close in meaning are relevant to the same queries and selected rubrics.

The papers 0,0 consider a wide class of various modifications of the EUTD rubrication algorithms based on the decision rules, artificial neural networks, growing pyramidal networks and probabilistic-statistical methods.

Analysis of the prospects of using the automated ENDD rubrication methods has showed that a multimodel approach is required to solve this task. It involves the EUTD classification using various intellectual analysis algorithms and taking into account the specificity of each document.

2.3. Criteria of selecting the ENDD rubrication method

To select of the rubrication method, it is advisable to take into account the following features: the rubric field characteristics, the amount of accumulated statistical information and the rubric thesaurus characteristics.

The rubric field characteristic shows the degree of interconnection of their thesauri. The presence of interrelated rubrics is the most difficult condition for the rubrication.

The amount of accumulated statistical information determines the possibility of constructing the various rubrication models. Thus, if it is possible to correctly use the probabilistic-statistical methods, the amount of EUTD can be considered "sufficient".

The rubric thesaurus characteristic determines whether the composition and influence degree of significant words of their dictionaries change. Thus, some rubrication models can't be used in the

conditions of dynamically changing thesauri because of the complexity of retraining and restructuring of the rubrication model structure.

The degree of rubric thesaurus intersection $K_{\cap}$ depends on the number of unique words for all the rubrics and defined as:

$$K_{\mathrm{I}} = \frac{1}{J} \cdot \sum_{j=1}^{J} \sum_{i=1}^{J} \frac{Count(R_i \mathrm{I} R_j)}{M_j},$$

where J – total number of rubrics, $Count(R_i \mathrm{I} R_j)$ – number of matching significant words in the thesauri of the rubrics R_i and R_j, M_j – total number of significant words in the thesaurus of the rubric R_j.

The following criteria levels of the thesaurus intersection have been taken on the basis of research results: $K_{\cap} < 0{,}15$ – insignificant level; $0{,}15 \le K_{\cap} \le 0{,}4$ – medium level; $K_{\cap} > 0{,}4$ – significant level.

In addition, it is advisable to take into account the EUTD size: short (contain less than 10% of unique significant words), medium (10-20% of unique significant words), large (more than 20% of unique significant words). Concrete values of the EUTD sizes are specified by experts when setting up the automated rubrication system.

3. Multimodel method of the EUTD rubrication

3.1. Typical situations of the EUTD rubrication

In accordance with the indicated EUTD characteristics, the typical situations of their rubrication have been outlined and methods of their conducting have been provided.

Typical situation 1. The analysis of short or medium-size EUTDs is performed in conditions of medium or significant degree of rubric intersection and in case of statistical data insufficiency for the effective use of traditional mathematical models of the EUTD rubrication. In this situation, it is advisable to use fuzzy pyramidal networks that allow classifying objects without a teacher in the absence of statistical data 0. In contrast to the known 0, this apparatus allows taking into account the degree of rubric cohesiveness based on the fuzzy-logic algorithm use that increases the accuracy of the EUTD rubrication.

Typical situation 2. The analysis of medium-size EUTDs is performed in conditions of insignificant or medium degree of rubric intersection and in case of statistical data insufficiency for the effective use of traditional mathematical models of the EUTD rubrication. In this case it is advisable to apply the model using weight coefficients that takes into account, in addition to statistical characteristics, expert information in the processing of the EUTD rubrication. In contrast to the known 0, this model should take into account the degree of significance of words in the EUTDs, depending on the appearance of new significant events affecting on the rubric thesauri.

Typical situation 3. The analysis of medium-size EUTDs is performed in conditions of medium or significant degree of the rubric intersection and in case of statistical data insufficiency for the effective use of traditional mathematical models of the EUTD rubrication. This situation determines the expediency of applying the rubric selection model based on a fuzzy decision tree that will improve the accuracy and efficiency of the rubrication with the use of additional expert information. In contrast to the known 0, this model should use syntax links and word roles, the fuzzy evaluation of document differences in the n-dimensional space of the text attributes in the construction and application of the fuzzy decision tree for referring the document to specific rubric in conditions of the interdependence of their thesauri.

Typical situation 4. The analysis of short-size EUTDs is performed in conditions of insignificant degree of rubric intersection and in case of sufficient amount and statistical data quality of this document type (for training the hybrid fuzzy models 0,0). For this situation, it is advisable to use a neural-fuzzy classifier based on a well-known approach 0, but using expert information to determine the keyword relevance in the formalization and the subsequent rubrication of the EUTD.

Typical situation 5 is characterized by the presence of large-size EUTDs, sufficient amount of statistical data on this document type and small degree of rubric intersection. In this situation, it is advisable to use the known probabilistic methods of the textual information analysis 0,0.

Typical situation 6. The analysis of large-size EUTDs is performed in conditions of medium or significant degree of rubric intersection and in case of sufficient amount and statistical data quality about this document type. In this case, it is advisable to use the voting method taking into account all models that will reduce the error probability in the EUTD rubrication.

Table 1 contains a brief description of the typical situations for the EUTD rubrication, based on the EUTD size, the degree of rubric intersection, the statistical data sufficiency for the application of probabilistic analysis methods and the corresponding models for the effective rubrication.

Table 1. Choice of the EUTD rubrication model.

Situation	EUTD size	Degree of rubric intersection	Statistical data sufficiency	EUTD rubrication model
1	short, medium	medium, significant	insignificant	Model based on fuzzy pyramidal networks
2	medium	insignificant, medium	insignificant	Model using the weighting coefficients
3	medium	medium, significant	insignificant	Model based on fuzzy decision tree
4	short	insignificant	significant	Cascade neural-fuzzy model
5	large	insignificant	significant	Probabilistic classifier
6	large	medium, significant	significant	Classifier voting method

3.2. Algorithm of implementing the multimodel method of the EUTD rubrication

The analysis of typical situations shows the impossibility of applying a unified model of the automated EUTD rubrication and allows justifying the advisability of creating a multimodel method for solving this task. The choice of a particular model of the EUTD rubrication should be based on the results of identifying the corresponding typical situation.

The algorithm of multimodel method of the EUTD rubrication based on the combined use of intellectual analysis models is given in Figure 1. Directly choosing a particular model or their ensemble is carried out using the base of fuzzy product rules (R1-R6).

The method involves analysis of the set of EUTDs $V=\{V_1,...,V_k,...V_K\}$, where each document V_k is represented as a set of meaningful words:

$$V_k = \{v_1^{(k)},...,v_{l_k}^{(k)},...,v_K^{(k)}\},\ k=1..K,\ l_k=1..L_k,$$

where $v_{l_k}^{(k)}$ – the word, L_k – number of words in EUTD k.

At the preliminary analysis stage, each document V_k is registered and presented as XML linguistic markup using a set of tags (descriptors).

After registration, the modified text document is analyzed:

$$V_k^r = V_k \ \Upsilon\, h^{(k)},$$

where $h^{(k)}$ – header information of document k that described by the rules.

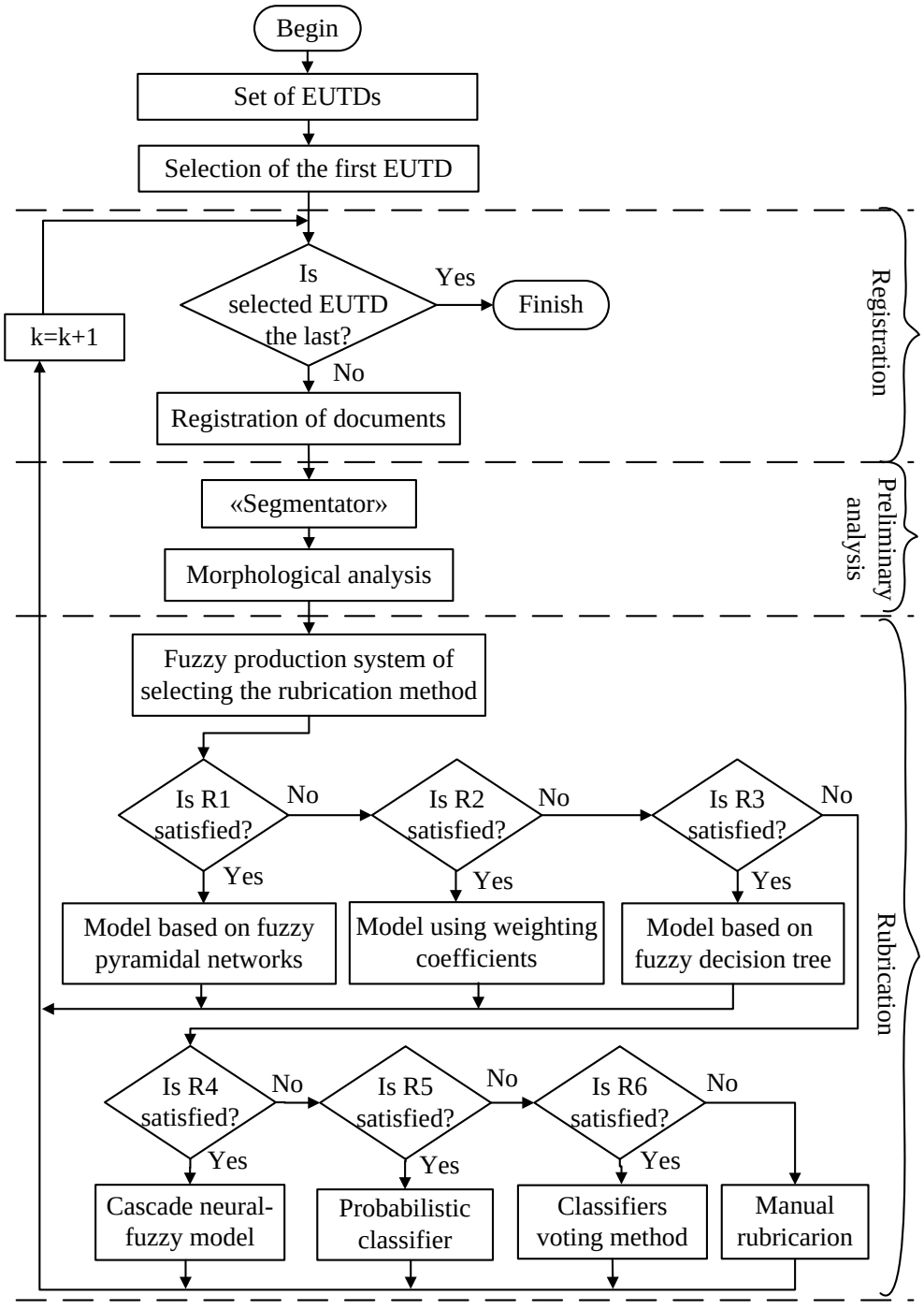

Figure 1. Algorithm of multimodel method of EUTD rubrication.

Document V_k^r is analyzed to extract the words, sentences, paragraphs in the markup language form in a text part (block "Segmentator").

As a result, the modified EUTD is formed at the output of the block "Segmentator":

$V_k^S = V_k \, Y \, h^{(k)} \, Y \, Sen^{(k)} \, Y \, Abz^{(k)}$,

where $Sen^{(k)}$ – set of the sentences, $Abz^{(k)}$ – set of the paragraphs.

It should be noted the complexity of implementing the segmentation procedure, the high degree of error influence on the final results of the automated EUTD rubrication. First of all, this is typical for procedures of the rubrication of short EUTD.

After linguistic marking the morphological analysis of the EUTD V_k^S is carried out, with the separation of lexical characteristics of words and morphemes (the least meaningful linguistic units).

As a result of the morphological analysis the modified EUTD is formed:

$V_k^M = V_k \, \mathrm{Y} \, h^{(k)} \, \mathrm{Y} \, M^{(k)}$,

where $M^{(k)}$ – the results of decomposing the document k into words, sentences, paragraphs and allocating the units and morphological characteristics attributed to the words.

The standard morphological characteristics of the thesaurus are stored in a morphological database, which is filled up from several sources. The most important source is the set of text documents of the Russian National Corpus containing the labeled texts 0.

The results of the morphological analysis of the EUTD are stored as a separate linguistic group in the original document. It means that in addition to the morphological information attributed to each word of the text, the syntactic structure is given for each sentence.

The next stage is the selection of a particular EUTD rubrication model in accordance with the highlighted typical situations given in the table 1.

Since the values of the EUTD characteristics given the table 1 (EUTD size L_k, degree of rubric thesaurus intersection $K_\cap$, statistical data sufficiency V_{st}) are determined by expert information, to implement the procedure of selecting a specific model of the automated EUTD rubrication M_j from the model set, it is proposed to use the base of fuzzy product rules of the following form:

$$IF(V_{st} \text{ is } V_{st}^a) and (K_\cap \text{ is } K_\cap^b) and (L_k \text{ is } L_k^c >) THEN (M \text{ is } M_j),$$

$$a \in \{insignificant, significant\},$$

$$b \in \{insignificant, medium, significant\},$$

$$c \in \{short, meduim, large\}, j \in \{1,2,3,4,5,6\}.$$

The proposed multimodel method of analyzing the electronic unstructured text documents makes it possible to increase the accuracy of attributing text documents to specific rubric, taking into account their specific subject area and different amount of statistical data.

4. References

[1] Zaboleyeva-Zotova A, Petrovskiy A, Orlova Yu and Shitova T 2016 Automated analysis of news texts International Journal “InformationContentandProcessing vol. 3 no. 3.

[2] Sebastiani F 2002 Machine learning in automated text categorization ACM Computing Surveys vol. 34 no. 1 pp 1-47.

[3] Gulin V 2013 Research and development of methods and software of text document classification: PhD thesis.

[4] Shabanov V 2003 Models and methods of automatic text document classification: PhD thesis.

[5] Bulygina O and Okunev B 2016 Creating fuzzy network tools to analyze prospects of projects of information and telecommunication infrastructure development Neyrokomp'yutery no. 7 pp 15-20.

[6] Sulima E and Milenin V 2010 Method of using the means of the educational material structuring International Journal “Information Theories and Applications vol. 17 no. 4 pp 387-395.

[7] Aleksandrov M 2008 Methods of automatic classification and statistical analysis of the input stream of text information in information systems: PhD thesis.

[8] Lewis D 1992 Representation and Learning in Information Retrieval: PhD thesis.

[9] Gimarov V and Dli M 2004 Neural network algorithm of complex object classification Programmnye produkty i sistemy no.4 pp 51-56.

[10] Kruglov V, Dli M and Golunov R 2001 Fuzzy logic and artificial neural networks (Moscow: Nauka, Fizmatlit).

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

[12] Chugreev V 2003 Model of the structural representation of textual information and methods of its thematic analysis on the basis of frequency-context classification: PhD thesis.

[13] Kozlov P 2015 Comparison of frequency and weight algorithms of automatic document analysis Nauchnoye obozreniye no. 14 pp 245-250.
[14] Kozlov P 2017 Automated analysis method of short unstructured text documents Programmnye produkty i sistemy no. 1 pp 100-105.

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