Application of Fuzzy Decision Trees for Rubricating Unstructured Electronic Text Documents

M. Dli, O. Bulygina, P. Kozlov

National Research University “MPEI”, Smolensk, Russia
e-mail: midli@imail.ru, baguzova_ov@mail.ru, originaldod@gmail.com

Abstract. Every day, a large number of electronic text messages (complaints, appeals, suggestions, etc.) presented in unstructured form arrives on the Internet portals of public authorities. The quality and speed of automated processing of such requests directly depend on the correctness of their rubrication (assignment to a specific subject area). The distinctive features of these messages (small size, presence of errors, lack of a certain structure, etc.) do not allow using the well-known methods of categorizing text documents. The authors have developed a model for rubricating electronic unstructured text documents taking into account syntactic links and word roles in the sentences on basis of fuzzy decision trees. The construction of a decision tree is based on analyzing the degree of rubric dictionary intersection and the distances between rubrics in the n-dimensional feature space. This model makes it possible to more precisely rubricate unstructured electronic text documents under conditions of interconnected rubrics and to increase the efficiency of document processing.

1 Introduction

Currently, public authorities are actively developing the information and communication technologies of interaction with citizens and organizations for the rapid exchange of electronic messages.

The main indicators used to assess the effectiveness of such electronic interaction are the responsiveness of public authorities to incoming requests and the satisfaction of citizens and organizations with the results of public services. Practice shows that these indicators are directly dependent on the speed of request processing and the accuracy of content recognition.

Automated processing of the requests received by the Internet portals of public authorities is based on the results of the classification (rubrication) of electronic text documents.

In: P. Sosnin, V. Maklaev, E. Sosnina (eds.): Proceedings of the IS-2019 Conference, Ulyanovsk, Russia, 24-27 September 2019, published at http://ceur-ws.org

In general, rubrication is the assignment of a document to one or several rubrics (thematic sections) according to its semantic content for subsequent analysis by experts in the given subject area for preparing a qualified answer [6, 9].

The articles [1, 23, 27] showed that the choice of a mathematical approach to the analysis and rubrication of electronic documents received by the Internet portals of public authorities directly depends on their characteristics.

So, the main distinctive features of electronic message send by citizens and organizations to public authorities are the following:

- small size of the message;
- presentation of the problem in free (unstructured) form;
- presence of grammatical and spelling errors in the text;
- simultaneous description of several issues related to different areas.

These features do not allow us to successfully use the classical methods for analyzing relational, object, hierarchical and labeled text data [3, 15].

In this regard, the analysis of unstructured documents can be carried out using data mining methods that allow us to correctly classify objects under conditions of lack of statistical data [10, 12, 20].

The foregoing makes it necessary to develop the method and algorithms for automated analysis of unstructured electronic text documents (UETD) taking into account the specifics of their content and use in the system of electronic public services.

2 Related works

Today, a large number of Russian and foreign publications are devoted to the problems of text document rubrication. Their authors propose a large variety of methods for binary and multi classification of texts.

The most popular machine learning methods used to classify texts are artificial neural networks and decision trees. The problems of using decision trees to solve this problem are considered by the authors [5, 7, 22, 30]. The features of applying various types of artificial neural networks for the text document classification are considered in the works [4, 8, 13, 24, 25].

However, it should be noted that these machine learning methods have particular limitations on its practical application due to the problems of training (size and quality of the training sample).

The article [14, 16, 21] showed that for automated analysis of UETDs it is advisable to use several rubrication models depending on the document characteristics (document size, degree of thesauri rubric intersection, amount of accumulated statistical information).

For example, we have proposed to use decision trees for analysis and rubrication of short and average-size UETDs under conditions of rubric intersection and lack of statistical data.

3 Statement of the UETD rubrication task

Today, decision trees are one of the popular methods for automatic data analysis built on example-based learning. The known models of decision trees assume a multi-level structure including [18]:

- nodes (vertices) - some attributes expressed in descriptive language;
- leaves (the lowest vertices) - selected classes characterized the individual subject area;
- links between nodes and leaves.

This method represents the rules in a hierarchical structure, where each object corresponds to single node that provides a solution. The classification process starts at the root node and moves to the leaves, checking the values of the vertex attributes.

However, there are situations when it is impossible to accurately classify an object by the specific attribute. These situations are resolved due to the capabilities of fuzzy logic: in this case, the degree of object belonging to a certain class is determined. In theory of fuzzy decision trees an object may have the properties of several attributes.

For each attribute it is necessary to determine several linguistic values and degree of example belonging to them. Instead of example set for a particular node, fuzzy decision tree groups their degree of ownership [2, 11].

The feasibility of constructing and using fuzzy decision tree for UETD rubrication is due to the simultaneous execution of the following conditions:

- the rubrics are linked by the subject, i.e. they are characterized by high degree of thesaurus intersection;
- the number of layers of fuzzy decision tree is greater than the number of rubrics;
- it is not enough data to train probabilistic models or neuro-fuzzy classifiers.

To apply the fuzzy decision tree for analyzing UETDs of considered types it is necessary to increase the number of levels for improving rubrication accuracy. It is also advisable to use fuzzy transition rules and develop a way to formalize text documents for adaptive accounting of the rubric field changes.

4 Algorithm of constructing the fuzzy decision tree

We have proposed an algorithm of constructing the fuzzy decision tree for UETD analysis which includes the following steps.

Step 1. For all thesauri of rubrics *R*, root vertices are formed. The number of child nodes *d* is equal to 2 by default.

Step 2. If the selected node is the final rubric (leaf), then go to step 6. Otherwise, thesaurus rubric intersections *d*, which absorb all other dictionaries with a coverage ratio of 0.9, are determined (sets of thesaurus rubrics *d,* maximally different from each other, are found, i.e. their dictionaries should not coincide by 90%).

Step 3. If dictionaries coincide by 90% or more, it is necessary to estimate the proximity of the rubric data dictionaries (i.e. compare the corresponding weights) in

n-dimensional space in order to determine the need to build additional levels in the tree.

So, if the distance ρ between dictionaries is less, than the threshold value ρ_{tv} then the creation of additional levels in the tree is not required and the selected rubrics become leaves. Otherwise, the first tree node is created for rubrics, the distance between which is greater than ρ_{tv}, and the second tree node is created for those whose distance is less than ρ_{tv}.

For each rubric R_j, J coordinates are calculated for each word w_{m_j} of the following form:

$$KD^{(R_j)} = \{(w_{m_j}^{(j)}, u_{m_j}^{(j,1)}, u_{m_j}^{(j,2)}, ..., u_{m_j}^{(j,j)}, ..., u_{m_j}^{(j,J)})\},\ j = 1, ..., J, m_j = 1, ..., M_j,$$

where $w_{m_j}^{(j)}$ – the word m_j in the rubric R_j, $u_{m_j}^{(j,J)} = r_p^{(J)} \mid w_{m_j}^{(j)} = w_p^{(J)}$ – the weighting factors of the word m_j of the rubric R_j in context of J, $r_{m_j}^{(j)} \in [0,1]$ – the degree of compliance of the word m_j with the rubric R_j.

To calculate the distance ρ between rubrics, it is necessary to find the centers of their cluster fields:

$$C^{(R_j)} = \{U_1^{(j)}, ..., U_j^{(j)}, ..., U_J^{(j)}\},$$

$$U_1^{(j)} = \frac{\sum_{m_j=1}^{M_j} u_{m_j}^{(j,1)}}{M_j}, ..., U_J^{(j)} = \frac{\sum_{m_j=1}^{M_j} u_{m_j}^{(j,J)}}{M_j},$$

where $U_J^{(j)}$ – coordinates of the center of the cluster field of the rubric R_j.

Further, distance ρ is calculated using the following formula:

$$\rho(R_j, R_i) = \rho(C^{R_j}, C^{R_i}) = 1 - \frac{1}{\sqrt{J}} \cdot \sqrt{\sum_{p=1}^{J} (U_p^{(j)} - U_p^{(i)})^2}.$$

Step 4. If it is impossible to find d intersections in step 3, then the number of intersections for the search is increased by 1 and step 3 is repeated, otherwise, go to step 5.

Step 5. Go to the first child node and then go to step 2.

Step 6. If there is the next child node, then select it and go to step 2, otherwise, the tree is built.

The created fuzzy decision tree is binary, although a large arity of the tree is allowed.

Figure 1 shows a flow chart of the algorithm for constructing a model for analyzing and rubricating UETDs based on the fuzzy decision tree.

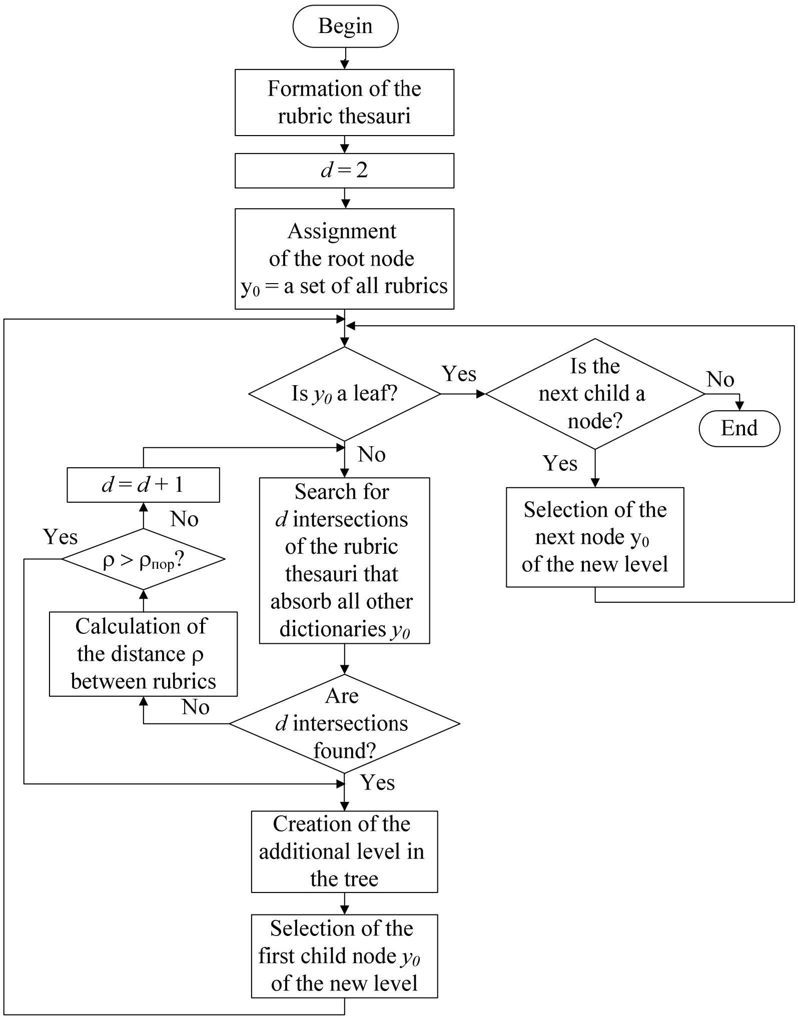
Begin
Formation of the rubric thesauri
$d=2$
Assignment of the root node y0 = a set of all rubrics
Is y_0 a leaf?
Yes
Is the next child a node?
No
End
Yes
Selection of the next node y0 of the new level
No
$d=d+1$
No
Yes
$\rho > \rho_{пор}$?
Search for d intersections of the rubric thesauri that absorb all other dictionaries y_0
Calculation of the distance ρ between rubrics
No
Are d intersections found?
Yes
Creation of the additional level in the tree
Selection of the first child node y_0 of the new level

Figure 1. A flow chart of the algorithm for constructing the fuzzy decision tree

5 Algorithm of applying fuzzy decision trees

We have proposed an algorithm for rubricating the short and average-size UETDs under conditions of rubric intersection and lack of statistical data using the fuzzy decision trees. It includes the following steps.

Step 1. The input document is presented as:

$$SD' = \{SD'_1, ..., SD'_k, ..., SD'_K\},\ SD'_k = \{v^{(k)}_{l_k}\},\ l_k = 1, ..., L_k,$$

where L_k – the number of words in UETD k.

Step 2. To determine whether UETD k belongs to rubrics, it is required to pass the fuzzy decision tree from the root node to the leaf, matching the set $SD_k^{'}$ with the nodes of the tree:

$$SD_k^{'} \leftrightarrow \{R_{\sum_1}^{(h)}, ..., R_{\sum_g}^{(h)}, ..., R_{\sum G}^{(h)}\},$$

where $R_{\sum g}^{(h)}$ – the sum of rubrics related to node g of the fuzzy decision tree at level h, G – the number of nodes at level h.

To do this, many assessments are introduced in the following form:

$$\forall j \in J : Est(SD_k^{'}, R_{\sum g}^{(h)}) = \{Est(SD_n^{(k)'}, R_{\sum g}^{(h)})\}, n = 1, ..., N,$$

$$Est(SD_n^{(k)'}, R_{\sum g}^{(h)}) = \{(v_p^{(k)}, u_p^{(k)})\}, (v_p^{(k)}, u_p^{(k)}) : u_p^{(k)} = r_{m_{R_{\Sigma g}^{(h)}}}^{(R_{\Sigma g}^{(h)})} \mid w_{m_{R_{\Sigma g}^{(h)}}}^{(R_{\Sigma g}^{(h)})} = v_p^{(k)},$$

where $u_p^{(k)}$ - the degree of compliance of the word p of UETD k with the sum of rubrics $R_{\sum g}^{(h)}$ corresponding to the node g on the layer h of the fuzzy decision tree; $r_{m_{R_{\Sigma g}^{(h)}}}^{(R_{\Sigma g}^{(h)})}$ - the average value of the degree of compliance $v_p^{(k)}$ of the word p of UETD k with rubrics $R_{\sum g}^{(h)}$.

Figure 2 shows an example of the fuzzy decision tree for rubricating UETD k.

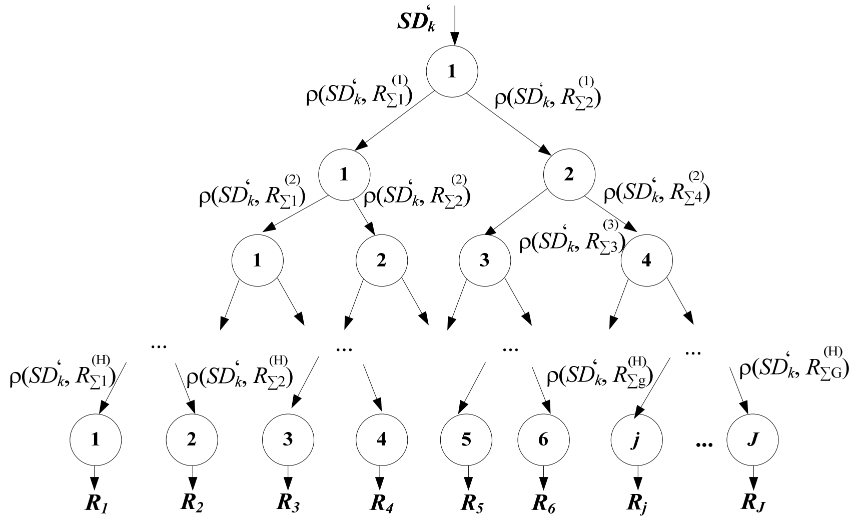

Figure 2. The fuzzy decision tree for rubricating UETD k

Comparison of UETD k with rubrics $R^{(h)}_{\sum g}$ is performed using the following fuzzy sets:

$$\forall j \in J, FS(SD^{'}_{k}, R^{(h)}_{\sum g}) = \left\{\left(\mu_{FS(SD^{'}_{k}, R^{(h)}_{\Sigma g})}(SD^{(k)'}_{n}) / s_{n}\right)\right\}, n = 1, ..., N,$$

where $\mu_{FS(SD^{'}_{k}, R^{(h)}_{\Sigma g})}(SD^{(k)'}_{n})$ – the degree of belonging of UETD k to rubrics $R^{(h)}_{\sum g}$ by syntactic parameter s_n:

$$\mu_{FS(SD^{'}_{k}, R^{(h)}_{\Sigma g})}(SD^{(k)'}_{n}) = \frac{1}{L^{(k)}_{n}} \sum_{p=1}^{L^{(k)}_{n}} u^{(k)}_{p}, n = 1, ..., N.$$

We introduce an indicator $\rho(SD^{'}_{k}, R^{(h)}_{\sum g})$ characterizing the degree of compliance UETD k with the rubrics $R^{(h)}_{\sum g}$.

Various methods can be applied to determine the degree of compliance [17, 19, 28]. The most appropriate method for solving this problem is to use the complement of the relative Euclidean distance between fuzzy sets:

$$\forall j \in J, \rho(SD^{'}_{k}, R^{(h)}_{\sum g}) = 1 - \frac{1}{\sqrt{N}} \sqrt{\sum_{n=1}^{N} \left(1 - \mu_{FS(SD^{'}_{k}, R^{(h)}_{\Sigma g})}(SD^{(k)'}_{n})\right)^{2}},$$

where n – the number of elements in fuzzy sets.

It is considered that UETD k applies mainly to the rubric $R^{*}_{\sum l}$, the degree of belonging to which is the maximum:

$$R^{*}_{\sum l} : \max_{j=1, ..., J} \rho(SD^{'}_{k}, R^{(h)}_{\Sigma g}).$$

Thus, on each layer h for the node g, the values $\rho(SD^{'}_{k}, R^{(h)}_{\Sigma g})$ are calculated for as many characteristics as this node has child nodes. The node with maximum value of the belonging characteristic is selected among them. Then the conditions are checked for it (threshold value or activation function).

Step 3. Step 2 is repeated until reaching the bottommost layer of the fuzzy decision tree, which defines the rubric of the input text document.

Figure 3 shows a flow chart of the algorithm for UETD rubrication based on the constructed fuzzy decision tree. The result of this algorithm is the maximum degree of document belonging to the closest rubric.

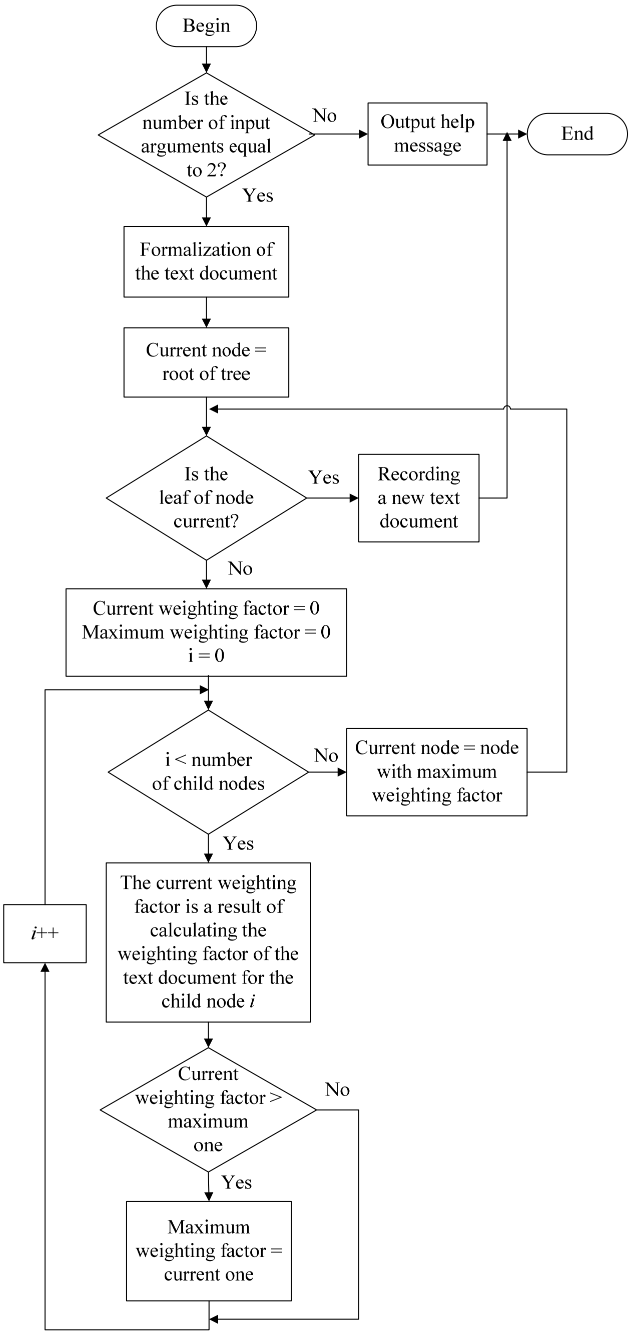
Begin
Is the number of input arguments equal to 2?
No
Yes
Output help message
End
Formalization of the text document
Current node = root of tree
Is the leaf of node current?
Yes
No
Recording a new text document
Current weighting factor = 0
Maximum weighting factor = 0
i = 0
i < number of child nodes
No
Yes
Current node = node with maximum weighting factor
The current weighting factor is a result of calculating the weighting factor of the text document for the child node *i*
i++
Current weighting factor > maximum one
No
Yes
Maximum weighting factor = current one

Figure 3. A flow chart of the algorithm for applying the fuzzy decision tree for rubricating UETDs

6 Procedure of UETD rubricating using a model based on the fuzzy decision trees

The proposed procedure of rubricating short and average-size UETDs under conditions of rubric intersection and lack of statistical data includes the following steps:

1. The preliminary analysis of the UETD includes [26]:

- the lexical analysis (highlighting words, numbers, punctuation, etc.);
- the morphological analysis (determining grammatical characteristics of lexemes and basic word forms).

2. The syntactic analysis using the parser is intended to forming the sets of significant words characterized by the same syntactic role in sentences.

3. The formalization of the UETD using weighting factors includes [29]:

- determining the degree of influence of significant words in relation to each rubric;
- accumulation and normalization of weighting factors.

4. The UETD rubrication using a model based on the fuzzy decision trees.

Figure 4 shows the process of rubricating UETDs in the form of IDEF0-diagram. The diagram clearly shows the stages of analysis and their participants: information system modules and help subsystems, regulatory documentation, specifications.

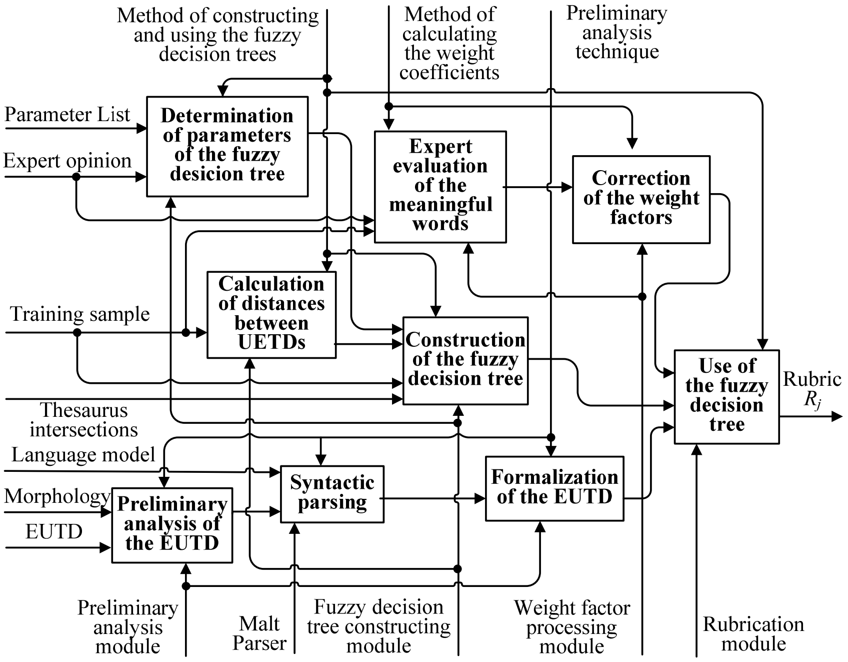
Method of constructing and using the fuzzy decision trees
Method of calculating the weight coefficients
Preliminary analysis technique
Parameter List
Expert opinion
Determination of parameters of the fuzzy desicion tree
Expert evaluation of the meaningful words
Correction of the weight factors
Calculation of distances between UETDs
Training sample
Construction of the fuzzy decision tree
Use of the fuzzy decision tree
Rubric R_j
Thesaurus intersections
Language model
Morphology
EUTD
Preliminary analysis of the EUTD
Syntactic parsing
Formalization of the EUTD
Preliminary analysis module
Malt Parser
Fuzzy decision tree constructing module
Weight factor processing module
Rubrication module

Figure 4 A diagram of the process of rubricating UETDs using the model based on the fuzzy decision trees

7 Conclusion

This paper proposed the model based on the fuzzy decision trees for analyze short and average-size unstructured electronic text documents under conditions of medium or significant degree of rubric thesaurus intersection and insufficient amount of accumulated statistical information.

The key advantages of the UETD rubrication model based on the fuzzy decision trees are follows:

- high accuracy of the UETD rubrication under conditions of rubric intersection and lack of statistical data due to a lower probability of random errors on the upper tiers of the fuzzy decision tree;
- low laboriousness of the UETD rubrication as a result of the directional (and not repetitive) analysis of a separate branch of the fuzzy decision tree;
- high scalability of the UETD rubrication model.

8 Acknowledgment

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

References:

1. Dli, M., Bulygina, O., Kozlov, P.: Development of multimethod approach to rubrication of unstructed electronic text documents in various conditions. Proceedings of the International Russian Automation Conference (RusAutoCon), Sochi (2018).

2. Janikow, C.: Fuzzy Decision Trees: Issues and Methods. IEEE Transactions of Man, Systems, Cybernetics, vol 28(1), pp. 1-14 (1998).

3. Khapaeva, T.: Automatic classification of documents. Softerra, no.2 (2002).

4. Nakagawa, T., Inui, K., Kurohashi, S.: Dependency tree-based sentiment classification using CRFs with hidden variables. Proceedings of ACL 2010 (2010).

5. Kaftannikov, I.L., Parasich, A.V.: Decision Tree's Features of Application in Classification Problems. Bulletin of the South Ural State University. Ser. Computer Technologies, Automatic Control, Radio Electronics, vol. 15, no. 3, pp. 26-32 (2015).

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

7. Shevelyov, O.G., Petrakov, A.V.: Text classification with decision trees and feed-forward neural networks. Tomsk State University Journal,vol.290, pp. 300-307 (2006).

8. Socher, R., Perelygin, A., Wu, J.Y., Chuang, J., Manning C.D., Ng A.Y., Potts C.: Recursive deep models for semantic compositionality over a sentiment treebank. EMNLP, 1631-1642 (2013).

9. Sulima, E., Milenin, V.: Method of using the means of the educational material structuring. International Journal "Information Theories and Applications", vol. 17, no. 4, pp.387-395 (2010).

10. Cao, Jian-fang, Wang, Hong-bin: Text categorization algorithms representations based on inductive learning. Information Management and Engineering (2010).

11. Faifer, M., Janikow, C.: Bottom-up Partitioning in Fuzzy Decision Trees. Proceedings of the 19th International Conference of the North American Fuzzy Information Society. IEEE, pp. 326-330 (2000).

12. Shmulevich, M.: 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, Moscow (2009).

13. Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J.: Distributed Representations of Words and Phrases and their Compositionality. Proceedings of NIPS 2013 (2013).

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

15. Aleksandrov, M.: Methods of automatic classification and statistical analysis of the input stream of text information in information systems: PhD thesis, Moscow (2008).

16. Dli, M., Bulygina, O., Kozlov, P., Ross, G.: Developing the economic information system for automated analysis of unstructured text documents. Journal of Applied Informatics, vol. 13, no. 5 (77), pp. 51-57 (2018).

17. Bauman, E., Dorofeyuk, A.: Types of fuzzines in clustering. Intelligent Techniques and Soft Computing. Verlag Mainz, Aachen (1997).

18. Quinlan, J.: Induction of decision trees. Machine Learning, vol. 1, no. 1, pp. 81-106 (1998).

19. Dli, M., Salov, N., Kakatunova, T., Tukaev, D.: An economic and mathematical model of it service provider selection on the basis of analysis of non-structured text documents. Journal of Engineering and Applied Sciences, vol. 14, no. 5, pp. 1662-1667 (2019).

20. Chugreev, V.: Model of the structural representation of textual information and methods of its thematic analysis on the basis of frequency-context classification: PhD thesis. S. Petersburg (2003).

21. Borisov, V., Dli, M., Kozlov, P.: The method of fuzzy analysis of texts and their rubrics actualization. Proceedings of the II International Scientific and Practical Conference. Ulyanovsk, pp. 259-263 (2018).

22. Petrov, S., Barrett, L., Thibaux, R., Klein, D. Learning accurate, compact, and interpretable tree annotation. Coling-ACL, pp. 433-440 (2006).

23. Dli, M., Bulygina, O., Kozlov, P.: Multimodel method of rubricating the unstructured electronic text documents. Fuzzy Technologies in the Industry – FTI 2018: Proceedings of the II International Scientific and Practical Conference. Ulyanovsk, pp. 366-372 (2018).

24. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A Neural Probabilistic Language Model. JMLR 3, pp.1137–1155 (2003).

25. Iyyer, M., Enns, P., Boyd-Graber, J., Resnik, P.: Political Ideology Detection Using Recursive Neural Networks. Proceedings of ACL 2014 (2014).

26. Kozlov, P.: Comparison of frequency and weight algorithms of automatic document analysis., Nauchnoye obozreniye, no. 14, pp. 245-250 (2015).

27. Dli, M., Bulygina, O., Kozlov, P.: Formation of the structure of the intellectual system of analyzing and rubricating unstructured text information in different situations. Journal of Applied Informatics, vol. 13, no. 4 (76), pp. 111-123 (2018).

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

29. Tukaev, D., Bulygina, O., Kozlov, P., Morozov, A., Chernovalova, M.: Cascade neural-fuzzy model of analysis of short electronic unstructured text documents using expert information. ARPN Journal of Engineering and Applied Sciences, vol. 13, no. 21, pp. 8531-8536 (2018).

30. Avdeenko, T., Makarova, E.: Acquisition of knowledge in the form of fuzzy rules for cases classification. Lecture Notes in Computer Science. Data Mining and Big Data, vol. 10387, pp. 536-544 (2017).