Method for Documents Rubrication and Analysis Based on Fuzzy Relations of Difference between Their Syntactical Characteristics

V. Borisov, M. Dli, P. Kozlov

The Branch of National Research University "Moscow Power Engineering Institute" in Smolensk, Russia
e-mail: vbor67@mail.ru, midli@imail.ru, originaldod@gmail.com

Abstract. The paper states the formulation and proposes a method for rubrication and analysis of electronic nonstructural documents. The application of the proposed method results in forming a tree structure of a rubric field based on fuzzy relations of difference between syntactical characteristics of rubricated documents. The documents analysis is based on the determination of the fuzzy correspondence for these documents according to syntactical characteristics with the values of the centers for the detected clusters sequentially from the root to the leaves of the built fuzzy decision tree. The conducted computational experiments have shown that the proposed method allows reducing the number of erroneously rubricated documents (in comparison with probabilistic and neural network methods)

1 Introduction

The program "Electronic government" suggests the dynamic introduction of information and communication technologies in the activities of public authorities. The main program goal is to increase the efficiency of public administration and to develop partnerships with civil society and business.

A key task of program implementation is to develop Internet services, which provide information support and a variety of services in electronic form. Their use can improve the quality and accessibility of state and municipal services to citizens and businesses, reduce the cost of their provision and increase the labor productivity in institutions of government at various levels.

One of the ways to use information and communication technology to solve this task is to automate the process of analyzing electronic appeals (applications, complaints, suggestions) of individuals and legal entities arriving at official websites and portals of authorities and local self-government.

The text rubrication plays an important role in the process of automatic analysis of incoming electronic appeals. It consists of their distribution according to thematic rubrics that determine the areas of activity of the departments involved in their processing and preparation of the corresponding response.

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

Today, there are many methodological approaches to the classification of documents of various types. The choice of a specific method is directly determined by the characteristics of the rubrication objects (i.e. documents received by public authorities).

The analysis has revealed the following specific characteristics of electronic documents received on official websites and portals of public authorities, which must be taken into account when choosing a rubrication method:

- relatively small size of electronic documents that impedes their statistical analysis;
- absence of marking in these documents that complicates the procedures for highlighting the structure and extracting the information relevant to the analysis;
- presence of grammar and syntactical errors in electronic messages that entails the necessity for additional processing;
- nonstationarity of the thesaurus (the composition and relevance of the rubric words);
- dynamic changes of the legislative and regulatory framework that can change the distribution of tasks between departments;
- description of several problems in one message (answers can be prepared by several specialists or even several departments).

These features significantly limit the possibilities of application of the methods based on the probabilistic and statistical approach to the rubrics generation and electronic text analysis [1, 6, 27].

The aforementioned determines the urgency of the task of developing a new method of rubricating the electronic unstructured documents, taking into account the specific features of text messages received on official websites and portals of public authorities.

2 Related works

At present, there are a variety of methods, models and algorithms for the classification of text documents written in natural language. However, each of them has its applicability conditions determined by the statement of the rubrication problem.

It was shown in articles [10, 11, 12] that the choice of a specific classification (rubrication) method is determined by such characteristics as the size of the analyzed document, the degree of rubric thesaurus intersection and the amount of accumulated statistical information.

Machine learning is a well-known approach to classifying unstructured documents. It offers the use of artificial intelligence methods that can learn from a set of precedents.

One of the machine learning methods that have been successfully used to solve various classification problems is artificial neural networks. The classification of texts is devoted to the works of authors [5, 17, 20, 21, 26]. The main limitation of the application of this approach is the requirement for the presence of a large amount of statistical data necessary for training algorithms.

Another machine learning method that can be used to classify text documents is fuzzy decision trees. They are based on learning by examples, while the rules are presented in the form of a hierarchical sequential structure. The issues of using fuzzy decision trees are considered in the works [2, 9, 13, 15, 16, 22, 23, 25, 26].

3 Statement of the rubrication problem

Initial data

1. For the formalized presentation of electronic unstructured documents (EUD) "a unification" for a set of syntactical characteristics is performed in advance. These characteristics are selected by a classical analyzer (parser), for example, LinkGrammar [24]:

$$S = \{s_n \mid n \in 1..N\},$$

where for the typical case $N = 5$; s_1 – the root word or the predicate; s_2 – the subject; s_3 – the adverbial modifier; s_4 – the object under the action; s_5 – the predicate.

2. There is a set of EUD

$$V = \{V_k \mid k \in 1..K\},$$

in which every document V_k is presented by a set of its relevant words:

$$\forall k = 1..K \;\; V_k = \{v_{l_k}^{(k)} \mid l_k \in 1..L_k\},$$

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

3. The set of EUD V is presented as a set of SD formalized documents:

$$SD = \{SD_k \mid k \in 1..K\},$$

in which the formalized document SD_k corresponds to each EUD:

$$\forall k \in 1..K \;\; SD_k = \{ SD_n^{(k)} \mid n \in 1..N\},$$

where $SD_n^{(k)}$ – the set of words from EUD V_k, corresponding to the syntactical parameter s_n [7].

Required

To propose a method for rubrication and EUD analysis based on the hierarchical clustering which uses fuzzy relations between syntactical characteristics of rubricating documents.

4 Method description

The proposed method for rubrication and analysis of EUD includes the steps discussed below.

Step 1. To give the parameters to determine the degree of correspondence for formalized documents according to the syntactical characteristics.

For each formalized document SD_k ($k \in 1..K$) a set of values for parameters $\widetilde{SD}_k = \left\{ \left(\widetilde{SD}_n^{(k)} / s_n \right) \mid n \in 1..N \right\}$ is given to assess the degree of its correspondence according to all syntactical characteristics.

Step 2. To determine the degree of difference between all pairs of formalized documents according to all syntactical characteristics.

Consider a pair of documents SD_k and SD_l, $k, l \in 1..K$:

$$SD_k = \{ SD_n^{(k)} \mid n \in 1..N \} \text{ и } SD_l = \{ SD_n^{(l)} \mid n \in 1..N \}.$$

To compare these documents sets of parameters values are given for all syntactical characteristics:

$$\widetilde{SD}_k = \left\{ \left(\widetilde{SD}_n^{(k)} / s_n \right) \mid n \in 1..N \right\} \text{ и } \widetilde{SD}_l = \left\{ \left(\widetilde{SD}_n^{(l)} / s_n \right) \mid n \in 1..N \right\}.$$

As a result, sets of parameter values are formed. These parameters characterize the degrees of difference for documents SD_k and SD_l according to all syntactical characteristics:

$$\tilde{d}(\widetilde{SD}_k, \widetilde{SD}_l) = \left\{ \left(d\left(\widetilde{SD}_n^{(k)}, \widetilde{SD}_n^{(l)} \right) / s_n \right) \mid n \in 1..N \right\},$$

where, for example, $d\left(\widetilde{SD}_n^{(k)}, \widetilde{SD}_n^{(l)} \right) = \left| \widetilde{SD}_n^{(k)} - \widetilde{SD}_n^{(l)} \right|$.

<u>Note</u>. The obtained set of values $\tilde{d}(\widetilde{SD}_k, \widetilde{SD}_l)$ can be presented in the form of a fuzzy set and interpreted as *a fuzzy difference* between fuzzy sets $\widetilde{SD}_k = \left\{ \left(\widetilde{SD}_n^{(k)} / s_n \right) \mid n \in 1..N \right\}$ and $\widetilde{SD}_l = \left\{ \left(\widetilde{SD}_n^{(l)} / s_n \right) \mid n \in 1..N \right\}$, syntactical characteristics from $S = \{ s_n \mid n \in 1..N \}$ are their carriers, and the documents degrees of correspondence to these characteristics $\widetilde{SD}_n^{(k)}$ and $\widetilde{SD}_n^{(l)}$ are the degrees of membership for the fuzzy set $\tilde{d}(\widetilde{SD}_k, \widetilde{SD}_l)$.

<u>Example</u>. Consider an example of documents SD_k and SD_l comparison taking into account the below-mentioned parameters:

$$\widetilde{SD}_k = \left\{ (0.7 / s_1), (0.5 / s_2), (0.3 / s_3), (0.3 / s_4), (0.8 / s_5) \right\} \text{ and}$$

$$\widetilde{SD}_l = \left\{ (0.1 / s_1), (0.9 / s_2), (0.2 / s_3), (0.6 / s_4), (0.4 / s_5) \right\}.$$

As a result, the following set of parameters values, characterizing the degree of difference between the documents according to the syntactical characteristics, is formed:

$$\tilde{d}(\widetilde{SD}_k, \widetilde{SD}_l) = \{(0.6/s_1), (0.4/s_2), (0.1/s_3), (0.3/s_4), (0.4/s_5)\}.$$

The calculation for the degree of differences according to all syntactical characteristics is performed for all pairs of formalized documents SD_k and SD_l, $k, l \in 1..K$.

Step 3. To form a matrix of difference between all pairs of the formalized documents.

The results of the previous step allow forming a compose matrix of difference between all pairs of documents.

Figure 1 shows such type of a matrix.

$$\begin{array}{c|ccccc|}
 & SD_1 & & SD_l & & SD_K \\
\hline
SD_1 & \tilde{d}(\widetilde{SD}_1, \widetilde{SD}_1) & \dots & \tilde{d}(\widetilde{SD}_1, \widetilde{SD}_l) & \dots & \tilde{d}(\widetilde{SD}_1, \widetilde{SD}_K) \\
 & \dots & & \dots & & \dots \\
SD_k & \tilde{d}(\widetilde{SD}_k, \widetilde{SD}_1) & \dots & \tilde{d}(\widetilde{SD}_k, \widetilde{SD}_l) & \dots & \tilde{d}(\widetilde{SD}_k, \widetilde{SD}_K) \\
 & \dots & & \dots & & \dots \\
SD_K & \tilde{d}(\widetilde{SD}_K, \widetilde{SD}_1) & \dots & \tilde{d}(\widetilde{SD}_K, \widetilde{SD}_l) & \dots & \tilde{d}(\widetilde{SD}_K, \widetilde{SD}_K) \\
\hline
\end{array}$$

Figure 1. The compose matrix of differences between all pairs of documents

Step 4. Fuzzy hierarchical clustering of documents based on the fuzzy relations of difference between all pairs of formalized documents according to all syntactical characteristics.

Parameters $d\left(\widetilde{SD}_n^{(k)}, \widetilde{SD}_n^{(l)}\right)$ are used as the parameters for fuzzy hierarchical clustering of formalized documents, their values characterize the results of pairwise comparison $\widetilde{SD}_n^{(k)}$ and $\widetilde{SD}_n^{(l)}$ separately according to all syntactical characteristics $\{s_n \mid n \in 1..N\}$.

It is reasonable to use well-known agglomerative methods as a base for the hierarchical clustering procedure [14].

Clusters $Cl = \{Cl_i \mid i \in 1..I\}$ are detected as a result of hierarchical clustering. Let the centers of these clusters be $\{\widetilde{Cl}_i \mid i \in 1..I\}$, where $\widetilde{Cl}_i = \left\{\left(\widetilde{Cl}_n^{(i)} / s_n\right) \mid n \in 1..N\right\}$.

The detected clusters $Cl = \{Cl_i \mid i \in 1..I\}$ correspond to the rubrics:

$$R = \{R_i \mid i \in 1..I\},$$

where for all $i \in 1..I$ $\quad R_i = \left\{ \left\langle t_{ji}, \left\{ (w_{ji_n} / s_n) \mid n = 1..N \right\} \right\rangle \mid j \in 1..J_i \right\}$, t_{ji} – j-th relevant word in the rubric R_i, $w_{ji_n} \in [0,1]$ – the degree of correspondence for the word t_{ji} to the syntactical characteristic s_n in the rubric R_i.

Thus, the result of the hierarchical clustering for documents is a tree-type structure of the formed rubric field based on the fuzzy relations between syntactical characteristics of the rubricating documents.

Step 5. Documents analysis.

The proposed procedure of analysis is based on the comparison of the correspondence degrees $\tilde{SD}_k$ for the analyzing document SD_k according to the syntactical characteristics with the values for the clusters centers $\tilde{SD}_k$ sequentially from the root to the leaves of the built decision tree. In this case, the analysis procedure takes into account the specificity of the detected clusters.

The analyzing document SD_k is the most relevant to the rubric R_l^*, the degree of fuzzy correspondence to which is the maximum:

$$R_l^* : \max_{i \in 1..I} \rho(\tilde{SD}_k, \tilde{Cl}_i).$$

To calculate a parameter characterizing the degree of fuzzy correspondence of formalized documents SD_k to the rubric R_i, it is reasonable to use the following [3, 4]:

$$\rho(\tilde{SD}_k, \tilde{R}_i) = 1 - \frac{1}{\sqrt{N}} \sqrt{\sum_{n=1}^{N} \left(\tilde{SD}_n^{(k)} - \tilde{Cl}_n^{(i)} \right)^2}.$$

5 The results of the proposed method application

The proposed rubrication method was programmatically implemented as a component of the comprehensive information system for the automatic processing of electronic unstructured documents arriving at official websites and portals of public authorities.

This method was tested in the automated processing and analysis of appeals (applications, complaints or suggestions) of citizens and organizations receiving by Administration of Smolensk region in 2018-2019.

To carry out the classification of incoming electronic appeals, the experts have identified 17 interconnected rubrics reflecting the urgent civic problems: general issues of society and politics (R_1), separation of powers and functions in the Administration (R_2), social sphere (R_3), education (R_4), suggestions for improving the city of Smolensk (R_5), family (R_6), culture (R_7), physical education and sport (R_8), housing and communal services (R_9), maintenance and utilities (R_{10}), housing stock (R_{11}), non-residential fund (R_{12}), securing the right to housing (R_{13}), economy (R_{14}), business activities (R_{15}), natural resources (R_{16}) and environmental protection (R_{17}).

Two well-known methods (probabilistic and neural network) successfully used to classify unstructured text documents have been practically implemented for comparative text analysis.

The Bayes classification was chosen as the first alternative method because of its ease of implementation and minimal human and financial costs for software implementation. It uses the procedure for classifying documents based on Bayes formula for conditional probability.

The input text document is presented as a sequence of terms $\{w_n\}$. Each rubric R_i is characterized by the unconditional probability $P(R_i)$ of the assignment of document V to it and the conditional probability $P(w|R_i)$ to meet the term w in document V, subject to the choice of rubric R_i. Then the probability $P(V|R_i)$ is understood as the probability that the text document will be classified subject to the selection of rubric R_i.

The procedure for document rubrication consists in calculating the probabilities $P(R_i|V)$ for all rubrics R_i and choosing the rubric for which this probability is maximal. Classifier training consists of compiling a vocabulary of probabilities of various terms $\{w_n\}$ for each rubric.

The methods of using probabilistic algorithms for the classification of text documents are considered in more detail in [8].

Convolutional neural networks were used as the second alternative method for document rubrication.

Convolutional networks are artificial neural networks of feedforward type when a signal travels sequentially along the neurons (from the first layer to the last). These networks were originally developed for image analysis. Good results in this area have led to their application for solving other classification tasks, including unstructured documents.

This neural network is an alternation of convolutional, subsampling and fully-connected layers. A text document arrives at the network input wherein each word is determined by the vector (e.g., may use the algorithm *word2vec*). The Softmax function which implements multiclassification is used for the output layer of the neural network.

Convolutional neural networks for the classification of text documents are considered in more detail in [18, 19, 28].

During the preliminary analysis, the authors have identified 4 typical situations, identified depending on three indicators: the size of the received document, the degree of intersection of the headings, and the amount of accumulated statistics for training the models.

Depending on these typical situations, Table 1 shows the results of comparative assessment for the correct rubrication and analysis based on the example of more than 10 thousand messages.

For the mentioned typical situations the proposed classification method has allowed reducing the number of erroneously rubricated text documents by 7% on average compared with the probabilistic method and by 6.3% compared with the neural network method.

Table 1. The results of the comparative assessment for the correct rubrication and analysis of EUD received by Administration of Smolensk region

Typical situation for analysis and rubrication of EUD			Results for rubrication and analysis of EUD, %		
EUD size	Degree of rubrics overlay	Sufficient statistics	Probabilistic method	Neural network method	*Proposed method*
up to 150 words	≤ 0.4	not enough	65	60	65
up to 150 words	> 0.15	not enough	62	66	79
up to 50 words	< 0.15	enough	69	87	90
more than 150 words	< 0.15	enough	89	85	89

6 Conclusion

As a result of the implemented method a tree structure of a rubric field is formed, this structure is based on the fuzzy relations between the syntactical characteristics of the rubricated documents. The document analysis is based on the detection of the fuzzy correspondence for these documents according to the syntactical characteristics with the values of the determined clusters sequentially from the root to the leaves of the built decision tree.

The proposed method for rubrication and analysis of electronic unstructured text documents was implemented by the software and tested during automated processing of appeals (applications, complaints or suggestions) of citizens and organizations receiving by Administration of Smolensk region. It has made possible to ensure efficient and high-quality actualization for the rubrics and document analysis under the conditions of nonstationary composition of the thesaurus and the relevance of the words in rubrics.

7 Acknowledgment

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

References:

1. Analytical report on the work of Administration of Smolensk region with citizens' appeals. URL: https://www.adminsmolensk.ru/obrascheniya_grazhdan/obzori_obrascheniy/news_16096.html.
2. 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).
3. Batyrshin, I.: On definition and construction of association measures. Journal of Intelligent & Fuzzy Systems, vol. 29, pp. 2319-2326 (2015).
4. Batyrshin, I.: Towards a general theory of similarity and association measures: Similarity, dissimilarity and correlation functions. Journal of Intelligent & Fuzzy Systems, vol.36, pp. 2977-3004 (2019).
5. Bengio, Y., Ducharme, R., Vincent, P., Jauvin, C.: A Neural Probabilistic Language Model. JMLR 3, pp.1137–1155 (2003).
6. Borisov, V., Dli, M., Kozlov P.: Analysis and monitoring of electronic text documents rubrication. MPIE Bulletin, vol. 4, pp.121-127 (2018).
7. Borisov, V., Dli, M., Kozlov, P.: The method of fuzzy analysis of texts and their rubrics actualization. Fuzzy Technologies in the Industry – FTI 2018: Proceedings of the II International Scientific and Practical Conference. Ulyanovsk, pp. 259-263 (2018).
8. Burlakov, M.E.: Using optimize naïve bayes classifier in problem of sms classification. Izvestia of Samara Scientific Center of the Russian Academy of Sciences, vol. 18, no. 4, pp. 705-709 (2016).
9. Cordon, O., Herrera, F., Hoffmann, F., Magdalena, L.: Genetic Fuzzy Systems: Evolutionary Tuning and learning of Fuzzy Knowledge Bases. Sinqgapore, New Jersey, London, Hong Kong, World Scientific Publishing, 462 p. (2001).
10. 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).
11. 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).
12. 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).
13. 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).
14. Jambu, M.: Hierarchical cluster analysis and correspondences. Moscow: Finance and statistics (1988).
15. Janikow, C.: Fuzzy Decision Trees: Issues and Methods. IEEE Transactions of Man, Systems, Cybernetics, vol 28(1), pp. 1-14 (1998).
16. 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).
17. Kalchbrenner, N., Blunsom, P.: Recurrent convolutional neural networks for discourse compositionality. Workshop on CVSC, pp. 119-126 (2013).
18. Kim, Y.: Convolutional neural networks for sentence classification. IEMNLP, September, pp. 1746 -1751 (2014).
19. Krizhevsky, A. Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. NIPS, pp. 1106 -1114 (2012).
20. Kruglov, V., Dli, M., Golunov, R.: Fuzzy logic and artificial neural networks. Moscow: Nauka, Fizmatlit (2001).

21. LeCun, Y. Text understanding from scratch. Computer Science Department (2016).
22. Nakagawa, T., Inui, K., Kurohashi, S.: Dependency tree-based sentiment classification using CRFs with hidden variables. Proceedings of ACL 2010 (2010).
23. Passino, K., Yurkovich, S.: Fuzzy Control. Addison-Wesley, NJ, 522 p. (1998).
24. Protasov, S.: LinkGrammar. URL: http://sz.ru/parser/doc/
25. Quinlan, J.: Induction of decision trees. Machine Learning, vol. 1, no. 1, pp. 81-106 (1998).
26. 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).
27. Uchitelev, N.: Classification of text information with the use of SVM. Information technologies and system, no.1, pp.335-340 (2013).
28. Zhang, X., Zhao, J., LeCun, Y.: Character-level convolutional networks for text classification. Advances in Neural Information Processing Systems, Febrary, pp. 649-657 (2015).