The Method of fuzzy analysis of texts and their rubrics actualization

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

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

Abstract. The work deals with the offered method of fuzzy analysis of texts and their rubrics actualization. The method is oriented to analyze electronic nonstructural texts of not big size in the following conditions: first, nonstationary composition and the importance of the keywords of the rubric field, second, in the absence or weak stucturization of these texts, third, if there are grammar or syntaxes inaccuracy and errors. The offered method is based on the original approach to the identification of the degree of the texts words fuzzy correspondence according to the well-founded set of syntactical characteristics with subsequent finding the degrees of text documents fuzzy correspondence to all rubrics. The method also allows to carry out monitoring of changes and actualization of rubrics according to the results of checking the formulated conditions of rubric field changes for the following typical situations: formation of the additional rubrics on the "boundary" of the existing rubrics; rubrics division, creating new rubrics, rubrics exclusion, rubrics combining. The offered method allows to raise the accuracy of analysis and the quality of texts classification at the expense of using the fuzzy approach for the accounting of analysis conditions uncertainty and nonstationarity of thesaurus of these texts as well as at the expense of operational actualization of rubrics depending on the composition and importance of the rubrics key words.

1. Introduction

Nowadays in the conditions of permanent perfection of internet technologies the tasks of automatic analysis of electronic nonstructural texts are actual, they possess the following features:

- relatively small size of such texts;
- such texts weak structuredness or no structuredness at all (no marking and fields for computer processing);
- presence of grammar and syntaxes inaccuracy and errors;
- analysis conditions uncertainty and nonstationarity of composition and importance of rubric field key words;
- high degree of rubrics interdependency.

These features put considerable limitations on the usage of traditional models and methods of morphological, syntaxes and semantic analysis of the texts. However, famous models and methods of knowledge acquisition from the text information take the requirements of operational rubric changes into account not sufficiently, this leads to the growth of the number of errors because of the wrong classification of the processing texts [1–7].

Therefore, the actual problem is to make a method of fuzzy analysis of electronic nonstructural texts and actualization of rubrics taking into account the detection of the following situations requiring operational changes of the rubric field: the additional rubrics formation on the "boundary" of the already exiting rubrics, rubrics division, creating new rubrics, rubrics exclusion, combining rubrics.

The offered method of texts fuzzy analysis and rubrics actualization includes the following main stages:

Stage 1. Rubric tasks and texts presentation on the basis of the detected syntaxes characteristics.

Stage 2. Texts analysis on the basis of the degree defining of their fuzzy correspondence to the rubrics.

Stage 3. Checking the rubric field changes conditions and rubrics of field actualization according to the results of this checking.

Let us consider the problems solving on the stages of the offered method in more details.

2. Rubric tasks and text presentation on the basis of the detected syntaxes characteristics

On the basis of the preliminary texts analysis the initial rubric multitude is given:

$$R = \{R_j \mid j \in 1..J\},$$

where for all $j \in 1..J$ $R_j = \left\{ \left\langle w_{jm}, r_{jm} \right\rangle \mid m \in 1..M_j \right\}$, w_{jm} – m- word in the rubric R_j, $r_{jm} \in [0,1]$ – the degree of correspondence of the word w_{jm} to rubric R_j.

For such texts presentation the «unification» of the set of the following syntaxes characteristics, detected, for example, by analyzer LinkGrammar is done ([8]):

$$S = \{s_n \mid n \in 1..N\}, \text{ then } N = 5,$$

where s_1 – the root word or predicate; s_2 – the subject; s_3 – the adverbial modifier; s_4 – the subject under action; s_5 –the predicate [9].

The texts multitude is presented in the form of:

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

where $SD_k = \{ SD_n^{(k)} \mid n \in 1..N\}$, $SD_n^{(k)}$ – word multitude of k- text, corresponding the syntaxes parameter s_n.

3. Texts analysis on the basis of the identification of the degree of their fuzzy correspondence to the rubrics

First, the degrees of fuzzy correspondence $\mu_{jn}(SD_n^{(k)}) \in [0,1]$ relative to syntaxes characteristics $SD_n^{(k)}$ to all rubrics are determined:

$$\forall j \in J, \quad \mu_{jn}(SD_n^{(k)}) = \frac{1}{L_n^{(k)}} \sum_{p=1}^{L_n^{(k)}} u_{jp}^{(k)}, n \in 1..N.$$

where $u_{jp}^{(k)}$ – the degree of correspondence of p-word from $SD_n^{(k)}$, primarily given for this word from rubric R_j.

To determine the degree of the text fuzzy correspondence to the rubrics let us introduce the parameter $\rho(SD_k, R_j)$ characterizing *the degree of text* SD_k *fuzzy correspondence to* rubric R_j:

$$\rho(SD_k, R_j) = 1 - \frac{1}{\sqrt{N}} \sqrt{\sum_{n=1}^{N} \left(\mu_{R_j}(R_{jn}) - \mu_{R_j n}(SD_n^{(k)}) \right)^2},$$

where $\tilde{R}_j = \left\{\left(\mu_{R_j}(R_{jn}) / s_n\right)\right\}$ – fuzzy multitude characterizing the "accurate coordinates" of rubric R_j [10]. For the case under consideration $\tilde{R}_j = \left\{(1/s_1), (1/s_2), (1/s_3), (1/s_4), (1/s_5)\right\}$, i.e.

$$\forall j \in J, \quad \rho_{1\alpha}(SD_k, R_j) = 1 - \frac{1}{\sqrt{N}} \sqrt{\sum_{n=1}^{N} \left(1 - \mu_{R_j n}(SD_n^{(k)})\right)^2}.$$

Text SD_k refers, in the greatest degree, to that rubric R_l^* for which the degree of correspondence is maximum:

$$R_l^* : \max_{j \in 1..J} \rho_{1\alpha}(SD_k, R_j).$$

4. Checking the conditions of rubric field changes and rubrics of field actualization in accordance with the results of this checking

To check the conditions of changing of rubric field let us introduce additionally the following parameters:

$$\forall j \in J, \quad \rho_{0,5}(SD_k, R_j) = 1 - \frac{1}{\sqrt{N}} \sqrt{\sum_{n=1}^{N} \left(0{,}5 - \mu_{R_j n}(SD_n^{(k)})\right)^2},$$

$$\forall j \in J \quad \forall j \in J, \quad \rho_{0\alpha}(SD_k, R_j) = 1 - \rho_{1\alpha}(SD_k, R_j),$$

where parameter $\rho_{0,5}(SD_k, R_j)$ characterizes *the degree of uncertainty of text* SD_k *referring to rubric* R_j and parameter $\rho_{0\alpha}(SD_k, R_j)$ characterizes *the degree of text* SD_k *discrepancy to rubric* R_j.

This stage realization considers the calculation of parameters $\rho_{1\alpha}(SD_k, R_j)$, $\rho_{0,5}(SD_k, R_j)$, $\rho_{0\alpha}(SD_k, R_j)$ for all texts and their analysis, according to the results of the analysis on the basis of the conditions given bellow, the revision of composition and rubric field structure is performed.

Let us consider the formulated conditions of detection and revision rules of composition and rubric field structure for the following basic situations: additional rubric formation, rubric division, new rubrics creation, rubric exclusion, rubric combining.

4.1. Additional rubric formation

The basis for *the additional rubric formation on the «boundary» of the already existed rubrics* R_i and R_j is the revealing of the considerable amount of texts (equal to the rubrics or more than the number of rubrics), for every of which the following condition is valid:

$$\alpha < \rho_{1\alpha}(SD_k, R_i) < \beta \ \wedge \alpha < \rho_{1\alpha}(SD_k, R_j) < \beta \ \wedge$$
$$\rho_{0,5}(SD_k, R_i) < \alpha \ \wedge \ \rho_{0,5}(SD_k, R_j) < \alpha \ \wedge$$
$$\alpha < \rho_{0\alpha}(SD_k, R_i) < \beta \ \wedge \ \alpha < \rho_{0\alpha}(SD_k, R_j) < \beta \ \wedge$$
$$\forall R_l \in R, l \neq i \neq j : \left(\rho_{1\alpha}(SD_k, R_l) > \beta \ \wedge \ \rho_{0,5}(SD_k, R_l) > \alpha \ \wedge \ \rho_{0\alpha}(SD_k, R_l) < \alpha\right),$$

where α and β – the upper and the lower boundary values (usually, α = 0.4 and β = 0.7 [11]), defining the reasonability of rubric field revision.

When revealing the number of texts equal to the rubrics or more than the number of rubrics, for which the above mentioned condition is performed, the conclusion about the reasonability of additional «boundary» rubric is made.

4.2. Rubric division

The basis for *the rubric division* R_j is the revealing of the considerable number of texts for each of which the following condition is fulfilled:

$$\alpha < \rho_{1\alpha}(SD_k, R_j) < \beta \ \wedge \ \rho_{0,5}(SD_k, R_j) < \alpha \ \wedge \ \alpha < \rho_{0\alpha}(SD_k, R_j) < \beta \ \wedge$$
$$\forall R_l \in R, l \neq j : \left(\rho_{1\alpha}(SD_k, R_l) > \beta \ \wedge \ \rho_{0,5}(SD_k, R_l) > \alpha \ \wedge \ \rho_{0\alpha}(SD_k, R_l) < \alpha\right),$$

where *j* – the number of the divided rubric.

4.3. New rubric creation

The basis for *the new rubric creation* is the revealing of the considerable number of texts for each of which the following condition is fulfilled:

$$\forall R_l \in R : \left(\rho_{1\alpha}(SD_k, R_l) > \beta \ \wedge \ \rho_{0,5}(SD_k, R_l) > \alpha \ \wedge \ \rho_{0\alpha}(SD_k, R_l) < \alpha\right).$$

4.4. Rubric exclusion

The basis for *the rubric exclusion* is the revealing of the considerable number of texts for each of which the following condition is fulfilled:

$$\rho_{1\alpha}(SD_k, R_j) > \beta \ \wedge \ \rho_{0,5}(SD_k, R_j) > \alpha \ \wedge \ \rho_{0\alpha}(SD_k, R_j) < \alpha.$$

4.5. Rubrics combining

The basis for *the rubric* R_i and R_j *combining* is the revealing of the considerable number of texts for which the following condition is fulfilled:

$$\rho_{1\alpha}(SD_k, R_i) < \alpha \ \wedge \ \rho_{1\alpha}(SD_k, R_j) < \alpha \ \wedge$$
$$\rho_{0,5}(SD_k, R_i) > \alpha \ \wedge \ \rho_{0,5}(SD_k, R_j) > \alpha \ \wedge$$
$$\rho_{0\alpha}(SD_k, R_i) > \beta \ \wedge \ \rho_{0\alpha}(SD_k, R_j) > \beta \ \wedge$$
$$\forall R_l \in R, l \neq i \neq j : \left(\rho_{1\alpha}(SD_k, R_l) > \beta \ \wedge \ \rho_{0,5}(SD_k, R_l) > \alpha \ \wedge \ \rho_{0\alpha}(SD_k, R_l) < \alpha\right),$$

where R_i and R_j – combining rubrics.

5. Experimental results

The offered method was used in Administration of Smolensk region when automated analysis of electronic nonstructural texts documents was performed, and it allowed to provide the operational actualization of rubrics depending on the structure and parameters of the text documents in the conditions of nonstationary composition of thesaurus and changes of the rubrics keywords importance. Automated rubrication of 5062 massages received in 2016–2017 was performed through the internet portal and by electronic mail. The analysis showed the presence of 17 different interconnected rubrics, among them there are rubrics such as general issues of society and politics, separation of powers and duties in Administration, social sphere, education, family, culture, housing and communal service etc. The results of rubrication showed that rubrics dynamic accounting, when using the probabalistic classification algorithm of text information as a basic tool of analysis [4, 6], allowed to reduce the number of erroneously rubricated texts up to 13,3 % in general.

6. Conclusion

The offered method was used in Administration of Smolensk region when automated analysis of electronic nonstructural texts documents was performed, and it allowed to provide the operational actualization of rubrics depending on the structure and parameters of the text documents in the conditions of nonstationary composition of thesaurus and changes of the rubrics keywords importance.

Eventually, the number of erroneously rubricated texts was managed to be reduced to 13.3 % on average.

7. References

[1] Ageev M S, Dobrov B V and Lukashevich N V 2008 *Automatic Text Rubrication: Methods and Problems*, *Scientific notes of Kazan State University*. Vol 150. No. 4. pp 25–40. (in Russian)

[2] Dumais S, Platt J, Heckerman D and Sahami M 1998 *Inductive Learning Algorithms and Representations for Text Categorization*, Proc. Int. Conf. on Inform. and Knowledge Manage. pp 148–155.

[3] Yang Y and Liu X 1999 *A Re-examination of Text Categorization Methods*, Proc. of Int ACM Conf. on Research and Development in Information Retrieval (SIGIR-99). pp 42–49.

[4] Zaboleeva-Zotova A V, Petrovsky A B, Orlova Yu A and Shitova T A 2016 *Automated Analysis of News Texts Themes* , *Int. J. Information Content and Processing.* Vol. 3. No. 3. Pp 288–299. (in Russian)

[5] Kozlov P Yu 2017 *Methods of Automated Analysis of Short Nonstructural Text Documents*, *Software products and systems* No. 1. pp 100–106. (in Russian)

[6] Borisov V V, Dli M I and Kozlov P Yu 2017 *Intellectual Methods of Nonstructural Texts* (Smolensk: Universum). p 156 ISBN 978-5-91412-364-9 (in Russian)

[7] Uchitelev N V 2013 *Classification of Text Information with the Help of SVM, Information technologies and systems.* No. 1. pp 335–340. (in Russian).

[8] Sajadi A and Borujerdi M 2013 *Machine Translation Based on Unification Link Grammar*, *Journal of Artificial Intelligence Review* pp 109–132. DOI: 10.1007/s10462-011-9261-7.

[9] Protasov S *Link Grammar* (Electronic materials). – http://sz.ru/parser/doc/ (Accessed July, 2018).

[10] Borisov V V, Fedulov A S and Zernov M M 2014 *The Base of the Fuzzy Sets Theory, The Base of Fuzzy Mathematics Series* book 1 (Moscow: Hot line–Telecom). p 88 (in Russian)

[11] Gimarov V A 2004 *Methods and Automated Systems of Dynamic Classification of Complex Technogenic Objects, Synopsis of a thesis paper of Dr.Tech.Sc*. (Moscow) (in Russian)

Acknowledgments

The work is conducted under support of the Russian Foundation for Basic Research (project 18-01-00558).