УДК 004.021
DOI: 10.24160/1993-6982-2018-4-121-127

Анализ и мониторинг рубрицирования электронных текстовых документов

В.В. Борисов, М.И. Дли, П.Ю. Козлов

Предложено решение задачи анализа и мониторинга рубрицирования электронных текстовых документов (ЭТД) в системах анализа жалоб, предложений и обращений граждан, поступающих в органы законодательной и исполнительной власти различного уровня с целью повышения качества и оперативности подготовки ответов. Показано, что рубрицирование подобных документов и мониторинг осуществляются в условиях нестационарности тезауруса (состава и важности слов), зависят от актуализации новых нормативных документов, что приводит к необходимости использования процедур динамической классификации при мониторинге рубрик ЭТД. Это определяет целесообразность мониторинга и периодического пересмотра рубричного поля с последующим изменением административных регламентов, закрепляющих выделенные рубрики за профильными департаментами в соответствии с ситуациями, требующими динамического изменения рубричного поля: выделением дополнительных «стыковых» рубрик, формированием новых рубрик, разделением, исключением и объединением рубрик.
Для анализа и мониторинга рубрицирования электронных текстовых документов предложен способ, включающий в себя следующие основные этапы: задание рубрик и совокупности ЭТД, сопоставление ЭТД рубрикам, проверку условий изменения рубричного поля. Рассмотрены ситуации, требующие изменения рубричного поля, определены критериальные показатели, описаны условия и предложены правила для изменения рубрик.
Предложеный способ анализа и мониторинга электронных текстовых документов позволяет обеспечить актуализацию рубрик в зависимости от структуры и показателей текстовых документов в условиях нестационарности состава тезауруса и важности ключевых слов рубрик.

Ключевые слова: автоматизированный анализ и нечеткое рубрицирование текстов, динамичный тезаурус, обработка обращений граждан, формализация электронных текстовых документов.

Для цитирования: Борисов В.В., Дли М.И., Козлов П.Ю. Анализ и мониторинг рубрицирования электронных текстовых документов // Вестник МЭИ. 2018. № 4. С. 121—127. DOI: 10.24160/1993-6982-2018-4-121-127.

Analyzing and Monitoring the Assignment of Rubrics to Electronic Text Documents

V.V. Borisov, M.I. Dli, P.Yu. Kozlov

The article proposes a solution to the problem of analyzing and monitoring the assignment of rubrics to electronic text documents (ETDs) in the systems for analyzing complaints, proposals and appeals of citizens arriving to the bodies of legislative and executive power at different levels, which will help the answers to be prepared with better quality and within a shorter period of time. It is shown that the work on assigning the rubrics to such ETDs and monitoring them is carried out under the conditions of an unsteady thesaurus (i.e., the composition and importance of words). In addition, the way in which this work is carried out depends on the updating of new regulatory documents, which generates the need of using dynamic classification procedures in monitoring the rubrics of ETDs. These circumstances determine the advisability of monitoring and periodically revising the rubric field with subsequently modifying the administrative regulations that define the responsibility of relevant departments for certain rubrics in accordance with the following situations that require dynamic changes to be made in the rubric field: defining additional “interfacing” rubrics; introducing new rubrics; dividing, excluding, and uniting rubrics.
To analyze and monitor the assignment of rubrics to electronic text documents, a method is proposed, which includes the following main steps: assigning rubrics to ETDs and defining their totality, checking the consistency of ETDs to their rubrics, and checking the rubric field alteration conditions.

Situations requiring alteration of the rubric field are considered, criterial indicators are defined, conditions under which the rubrics can be modified are described, and rules of changing them are proposed.
The proposed method for analyzing and monitoring ETDs opens the possibility of updating their rubrics depending on the structure and indicators of text documents under the conditions of an unsteady composition of the thesaurus and importance of key words in the rubrics.

Key words: computer-aided analysis and fuzzy assignment of rubrics to texts, dynamic thesaurus, processing of citizens' appeals, formalization of electronic text documents.

For citation: Borisov V.V., Dli M.I., Kozlov P.Yu. Analyzing and Monitoring the Assignment of Rubrics to Electronic Text Documents. MPEI Vestnik. 2018;4:121—127. (in Russian). DOI: 10.24160/1993-6982-2018-4-121-127.

Одно из направлений использования процедур автоматизированного анализа электронных текстовых документов (ЭТД) — рубрицирование жалоб, предложений и обращений граждан, поступающих в органы законодательной и исполнительной власти различного уровня с целью повышения оперативности подготовки ответа [1 — 3].

Отличительная особенность условий решения указанной задачи — нестационарность тезауруса (состава и важности слов), который зависит от изменений нормативной правовой базы, выступлений должностных лиц и политических деятелей. Это увеличивает количество ошибок системы документооборота из-за неправильного рубрицирования обрабатываемых текстов и ведет к необходимости мониторинга и периодического пересмотра рубричного поля с последующим изменением административных регламентов, закрепляющих выделенные рубрики за профильными департаментами. Такми образом, актуальной задачей является мониторинг рубрик ЭТД для анализа динамики состава и характеристик рубрик и выявления ситуаций, требующих динамического изменения рубричного поля, таких как выделение дополнительных рубрик на стыках уже существующих; разделение, исключение, объединение и формирование новых рубрик [4—10].

Для анализа ЭТД и мониторинга рубричного поля предлагается следующий способ.

Этап 1. Задание рубрик электронных текстовых документов.

Исходя из предварительного анализа ЭТД, поступающих в органы законодательной и исполнительной власти, задается первоначальное множество рубрик:

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

где для всех $j \in 1...J$ $R_j = \left\{\left\langle w_{jm}, r_{jm}\right\rangle \mid m \in 1...M_j\right\}$; w_{jm} — m-е слово в рубрике R_j; $r_{jm} \in [0, 1]$ — степень соответствия слова w_{jm} рубрике R_j.

Этап 2. Задание совокупности электронных текстовых документов.

Для данного представления ЭТД выполняется «унификация» набора следующих синтаксических характеристик, выделяемых анализатором LinkGrammar [11]:

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

где s_1 — корневое слово или сказуемое; s_2 — подлежащее; s_3 — обстоятельство; s_4 — предмет, над которым совершается действие; s_5 — сказуемое [12].

Множество ЭТД представлено в виде

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

где $SD_k = \{ SD_n^{(k)} \mid n \in 1...N\}$, $SD_n^{(k)}$ — множество слов k-го ЭТД, соответствующих синтаксическому параметру s_n.

Этап 3. Определение степени нечеткого соответствия $\mu_{jn}(SD_n^{(k)}) \in [0, 1]$ относительно синтаксических характеристик $SD_n^{(k)}$ ко всем рубрикам.

Для всех $j \in J$

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

где $u_{jp}^{(k)}$ — степень соответствия p-го слова из $SD_n^{(k)}$, изначально заданное для этого слова из рубрики R_j.

Этап 4. Определение степеней нечеткого соответствия документа рубрикам.

Введем показатель $\rho(SD_k, R_j)$, характеризующий степень нечеткого соответствия ЭТД SD_k рубрике R_j:

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

где $\tilde{R}_j = \left\{\left(\mu_{R_j}(R_{jn}) / s_n\right)\right\}$ — нечеткое множество, характеризующее «четкие координаты» рубрики R_j [13].

Для рассматриваемого случая

$$\tilde{R}_j = \left\{(1/s_1), (1/s_2), (1/s_3), (1/s_4), (1/s_5)\right\},$$

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

Документ SD_k в наибольшей степени относится к той рубрике R_l^*, степень принадлежности к которой является максимальной

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

Введем дополнительно следующие показатели.

Для всех $j \in J$

$$\rho_{\widetilde{0,5}}(SD_k, R_j) = 1 - \frac{1}{\sqrt{N}} \sqrt{\sum_{n=1}^{N} \left(0,5 - \mu_{R_j n}(SD_n^{(k)})\right)^2};$$

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

где $\rho_{\widetilde{0,5}}(SD_k, R_j)$ характеризует степень неопределенности отнесения ЭТД SD_k к рубрике R_j, а $\rho_{\tilde{0}}(SD_k, R_j)$ — степень несоответствия ЭТД SD_k рубрике R_j.

Этап 5. Проверка условий изменения рубричного поля.

Реализация данного этапа предлагает вычисление показателей $\rho_{\tilde{0}}(SD_k, R_j)$, $\rho_{\widetilde{0,5}}(SD_k, R_j)$, $\rho_{\tilde{0}}(SD_k, R_j)$ для всех ЭТД и их анализ, по результатам которого на основе определеных условий пересматриваются состав и структура рубричного поля.

Рассмотрим условия и правила разрешения типовых ситуаций изменения рубричного поля.

На рис. 1 приведено отнесение одного ЭТД к первой, а другого ЭТД к третьей рубрикам (при наличии трех рубрик и двух синтаксических параметров).

Изучим предлагаемые условия выявления основных ситуаций и правила пересмотра состава и структуры рубричного поля.

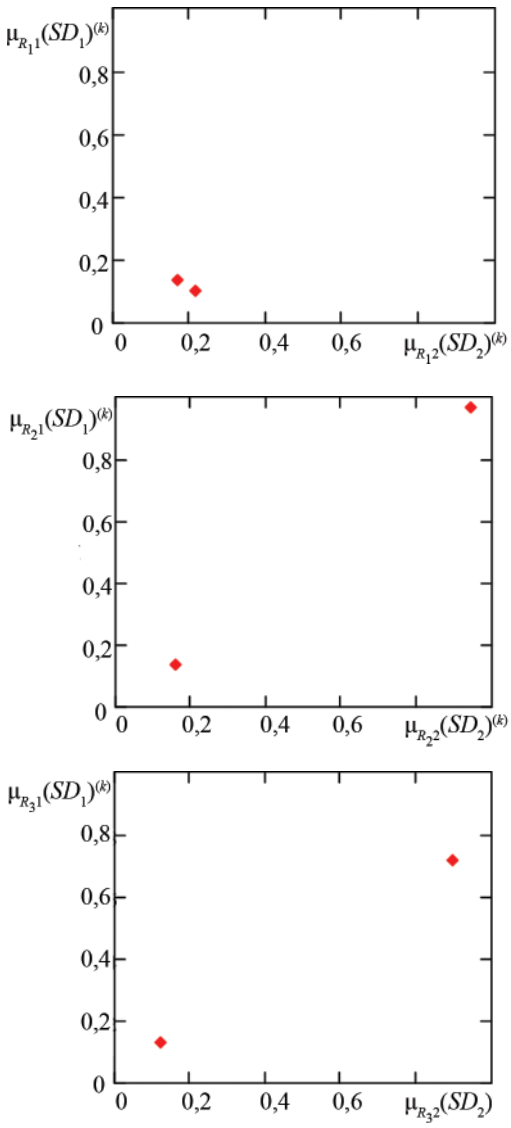

Рис. 1. Пример рубрицирования двух ЭТД по двум синтаксическим параметрам

Основанием для выделения дополнительной рубрики на стыке уже существующих R_i и R_j является поступление в систему автоматизированного рубрицирования количества ЭТД (равного или большего числу рубрик), для которых выполняются следующие условия:

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

где α, β — нижнее и верхнее граничные значения, определяющие целесообразность пересмотра рубричного поля, α = 0,4; β = 0,7 [14].

В процессе анализа ЭТД решение о введении дополнительной рубрики принимается по правилу

$$K_1 C_1 + K_2 C_2 > (K_1 + K_2) C_3 \ \text{ при } \ t_{\text{отв}} < t_{\text{доп}}, \tag{2}$$

где K_1 — число неправильно рубрицированных ЭТД на момент мониторинга рубричного поля; K_1 — число неправильно рубрицированных ЭТД; C_1 — затраты, вызванные неправильным рубрицированием одного ЭТД; C_2 — затраты, вызванные трудностью подготовки ответа на ЭТД; C_3 — затраты, связанные с обработкой ЭТД при использовании дополнительной сформированной рубрики; $t_{\text{отв}}$, $t_{\text{доп}}$ — отведенное и максимально допустимое время ответа на ЭТД.

При выполнении данного правила (с учетом условия (1)) делается вывод о целесообразности формирования дополнительной стыковой рубрики, к тому же оно применимо и для других ситуаций, требующих изменения рубричного поля.

На рис. 2 показана модель ситуации целесообразности формирования дополнительной стыковой рубрики для двух синтаксических параметров.

Основанием для разделения рубрики R_j служит поступление в систему автоматизированного рубрицирования значимого (с точки зрения выполнения правила (2)) количества ЭТД, для которых выполняется следующее условие:

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

где j — номер разделяемой рубрики.

На рис. 3 представлена ситуация целесообразности разделения рубрики для двух синтаксических параметров, а на рис. 4 — результаты рубрицирования ЭТД после разделения рубрики.

Итоги рубрицирования ЭТД (рис. 4) являются целевыми и для других ситуаций, требующих динамического изменения рубричного поля.

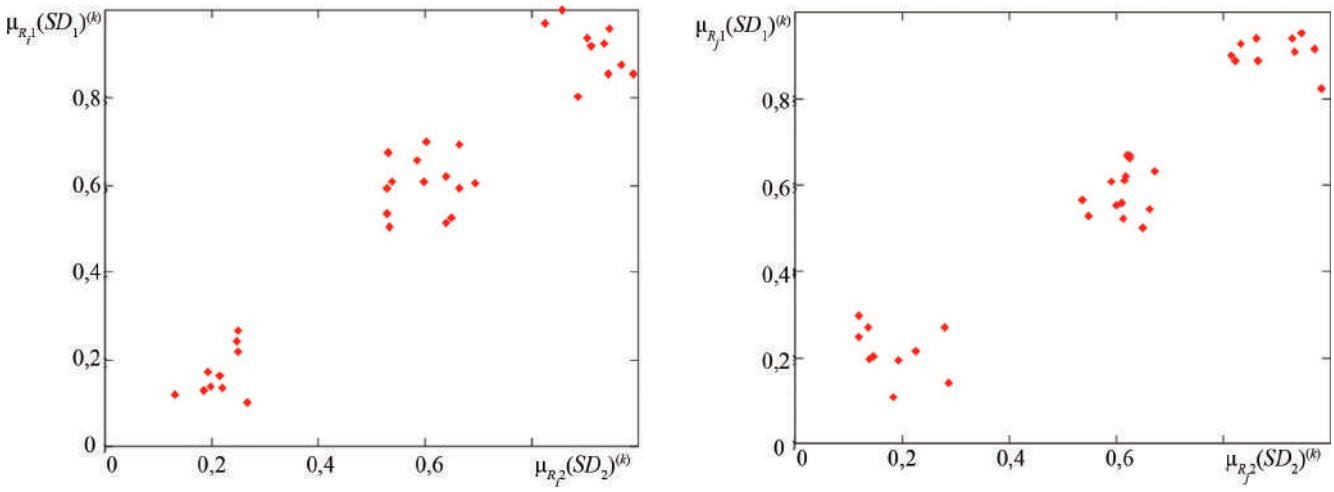

Рис. 2. Ситуации целесообразности формирования дополнительной стыковой рубрики

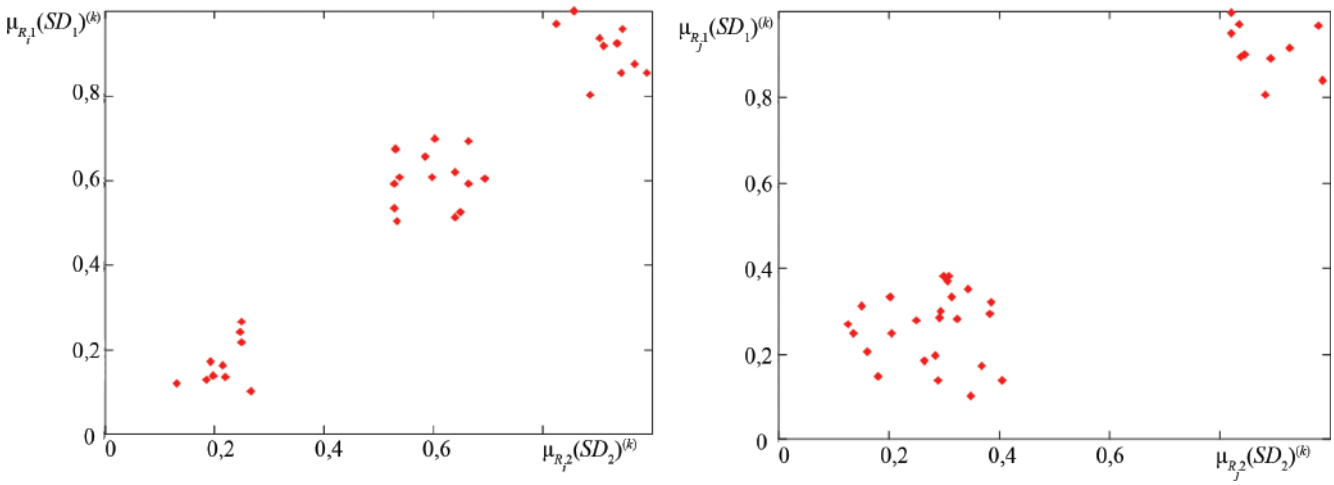

Рис. 3. Ситуации целесообразности разделения рубрики

Основанием для формирования новой рубрики служит ситуация (рис. 5), возникающая при появлении достаточного количества ЭТД, для которых выполняется следующее условие:

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

Основанием для исключения рубрики является ситуация, возникающая при появлении подавляющего числа ЭТД, для которых выполняется следующее условие:

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

Если для рубрики практически все ЭТД позиционируются так, как это показано на рис. 6, то эту рубрику целесообразно исключить.

Основанием для объединения рубрик R_i и R_j служит ситуация, возникающая при появлении достаточного количества ЭТД, для которых выполняется следующее условие:

$$\rho_{\tilde{1}}(SD_k, R_i) < \alpha \wedge \ \rho_{\tilde{1}}(SD_k, R_j) < \alpha \wedge$$
$$\rho_{\widetilde{0,5}}(SD_k, R_i) > \alpha \ \wedge \ \rho_{\widetilde{0,5}}(SD_k, R_j) > \alpha \wedge$$
$$\rho_{\tilde{0}}(SD_k, R_i) > \beta \wedge \ \rho_{\tilde{0}}(SD_k, R_j) > \beta \wedge \qquad (3)$$
$$\wedge \forall R_l \in R, l \neq i \neq j:$$
$$:\left(\rho_{\tilde{1}}(SD_k, R_l) > \beta \ \wedge \ \rho_{\widetilde{0,5}}(SD_k, R_l) > \alpha \ \wedge \ \rho_{\tilde{0}}(SD_k, R_l) < \alpha\right),$$

где R_i и R_j — объединяемые рубрики.

Выполнение условия (3) для объединения рубрик R_i и R_j изображено на рис. 7.

Предложенный способ анализа и мониторинга электронных текстовых документов использован при автоматизированном анализе электронных неструктурированных текстовых документов в Администрации Смоленской области и позволил обеспечить оперативную актуализацию рубрик в зависимости от структуры и показателей текстовых документов в условиях нестационарности состава тезауруса и важности ключевых слов.

Работа выполнена при поддержке РФФИ (проект № 18-01-00558_а).

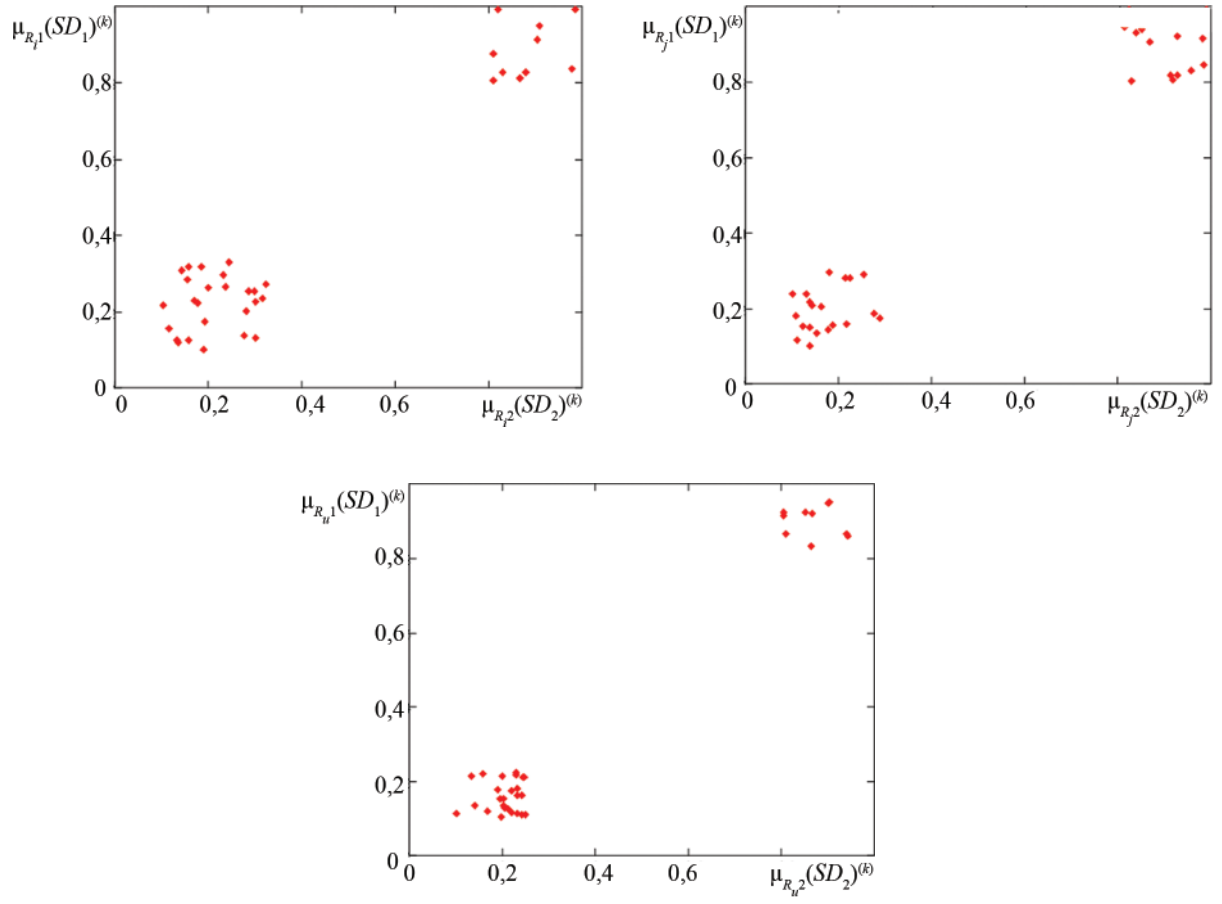

Рис. 4. Результаты рубрицирования ЭТД после разделения рубрики

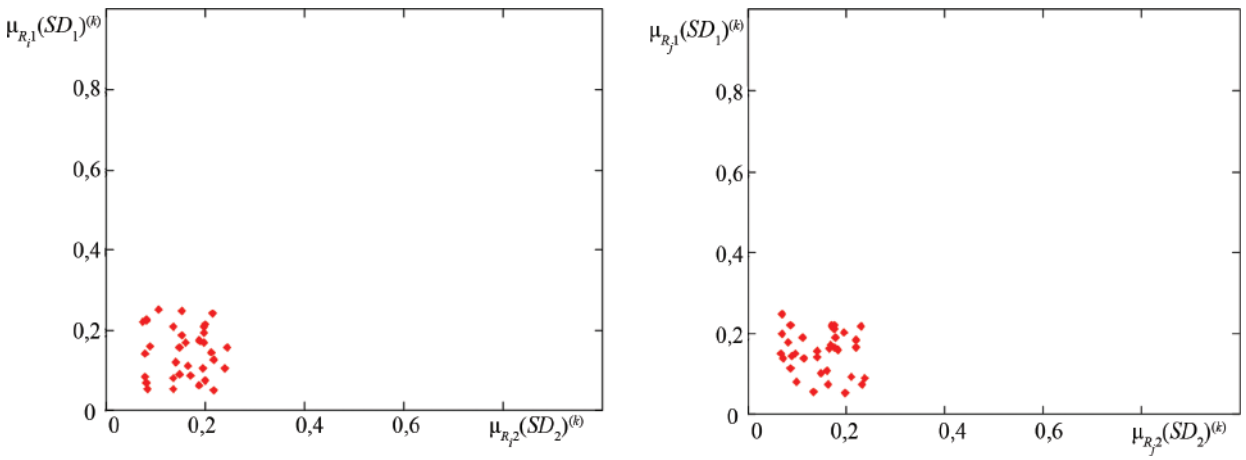

Рис. 5. Ситуации целесообразности формирования новой рубрики

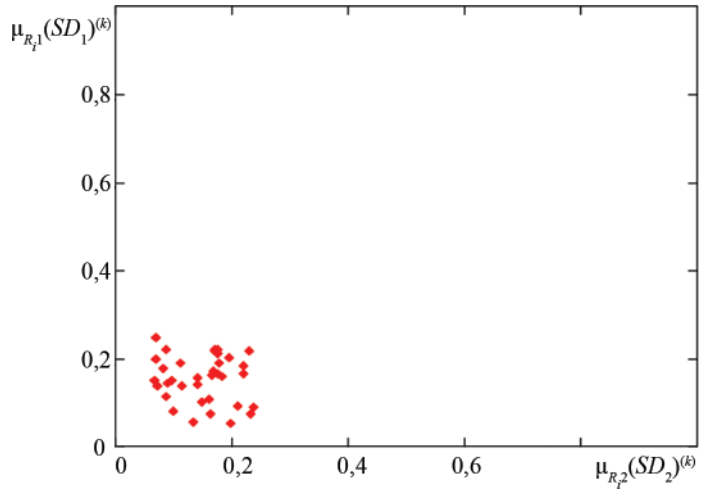

Рис. 6. Ситуация целесообразности исключения рубрики

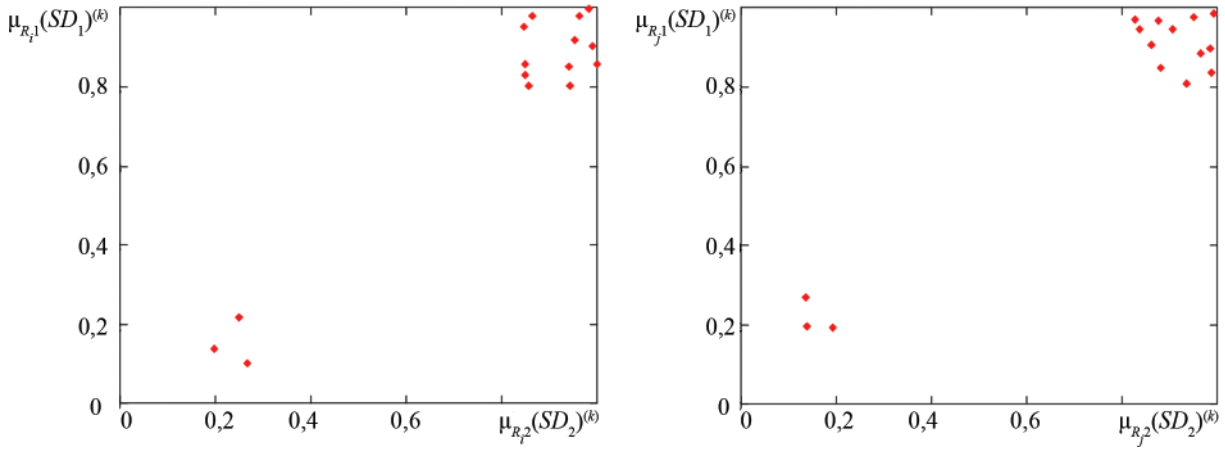

Рис. 7. Ситуации целесообразности объединения рубрик

Литература

1. **Козлов П.Ю.** Методы автоматизированного анализа коротких неструктурированных текстовых документов // Программные продукты и системы. 2017. № 1. С. 100—106.

2. **Аналитическая** справка о работе Аппарата Администрации Смоленской области с обращениями граждан [Офиц. сайт] https://www.admin smolensk.ru/obrascheniya_grazhdan/obzori_obrascheniy/news_16096.html (дата обращения 17.06.2017).

3. **Обзор** обращений граждан Администрации города Санкт-Петербурга [Офиц. сайт] http://gov.spb.ru/gov/obrasheniya–grazhdan/otchet–obrasheniya/?page=1 (дата обращения 25.06.2017).

4. **Гимаров В.А., Дли М.И., Круглов В.В.** Задачи распознания нестационарных образов // Известия РАН. Серия «Теория и системы управления». 2004. № 3. С. 92—96.

5. **Гимаров В.А., Дли М.И.** Нейросетевой алгоритм классификации сложных объектов // Программные продукты и системы. 2004. № 4. С. 51—56.

6. **Singh. S.** Dynamic Pattern Recognition for Temporal Data // Proc. 5th European Congress on Intelligent Techniques and Soft Computing. Aachen, 1997. V 3. Pp. 1993—1997.

7. **Учителев Н.В.** Классификация текстовой информации с помощью SVM // Информационные технологии и системы. 2013. № 1. С. 335—340.

8. **Заболеева-Зотова А.В., Петровский А.Б., Орлова Ю.А., Шитова Т.А.** Автоматизированный анализ тематики текстов новостей // Intern. J. Information Content and Proc. 2016. V. 3. No. 3. Pp. 288—299.

9. **Шграев А.Г., Фальк В.Н.** Линейные классификаторы в задаче классификации текстов // Вестник МЭИ. 2013. № 4. С. 204—209.

10. **Фальк В.Н., Бочаров И.А., Шграев А.Г.** Трансдуктивное обучение логистической регрессии в задаче классификации текстов // Программные продукты и системы. 2014. № 2. С. 114—117.

11. **Козлов П.Ю.** Сравнение частотного и весового алгоритмов автоматического анализа документов // Научное обозрение. 2015. № 14. С. 245—250.

12. **Протасов С.** Грамматика связей Link Grammar. [Электрон. ресурс] http://sz.ru/parser/doc/ (дата обращения 10.07.2017).

13. **Борисов В.В., Федулов А.С., Зернов М.М.** Основы теории нечетких множеств. М.: Горячая линия–Телеком, 2014.

14. **Гимаров В.А.** Методы и автоматизированные системы динамической классификации сложных техногенных объектов: автореф. дисс. ... доктора техн. наук. М., 2004.

References

1. **Kozlov P.Yu.** Metody Avtomatizirovannogo Analiza Korotkih Nestrukturirovannyh Tekstovyh Dokumentov. Programmnye Produkty i Sistemy. 2017;1:100—106. (in Russian).

2. **Analiticheskaya** Spravka o Rabote Apparata Administratsii Smolenskoy Oblasti s Obrashcheniyami Grazhdan [Ofits. Sayt] https://www.admin smolensk.ru/obrascheniya_grazhdan/obzori_obrascheniy/news_16096.html (Data Obrashcheniya 17.06.2017). (in Russian).

3. **Obzor** Obrashcheniy Grazhdan Administratsii Goroda Sankt-Peterburga [Ofits. Sayt] http://gov.spb.ru/gov/obrasheniya–grazhdan/otchet–obrasheniya/?page=1 (Data Obrashcheniya 25.06.2017). (in Russian).

4. **Gimarov V.A., Dli M.I., Kruglov V.V.** Zadachi Raspoznaniya Nestatsionarnyh Obrazov. Izvestiya RAN. Seriya «Teoriya i Sistemy Upravleniya». 2004;3:92—96. (in Russian).

5. **Gimarov V.A., Dli M.I.** Neyrosetevoy Algoritm Klassifikatsii Slozhnyh Ob′ektov. Programmnye Produkty i Sistemy. 2004;4:51—56. (in Russian).

6. **Singh S.** Dynamic Pattern Recognition for Temporal Data. Proc. 5th European Congress on Intelligent Techniques and Soft Computing. Aachen, 1997;3:1993—1997.

7. **Uchitelev N.V.** Klassifikatsiya Tekstovoy Informatsii s Pomoshch'yu SVM. Informatsionnye Tekhnologii i Sistemy. 2013;1:335—340. (in Russian).

8. **Zaboleeva-Zotova A.V., Petrovskiy A.B., Orlova Yu.A., Shitova T.A.** Avtomatizirovannyy Analiz Tematiki Tekstov Novostey. Intern. J. Information Content and Proc. 2016;3;3:288—299. (in Russian).

9. **Shagraev A.G., Fal'k V.N.** Lineynye Klassifikatory v Zadache Klassifikatsii Tekstov. Vestnik MPEI. 2013;4:204—209. (in Russian).

10. **Fal'k V.N., Bocharov I.A., Shagraev A.G.** Transduktivnoe Obuchenie Logisticheskoy Regressii v Zadache Klassifikatsii Tekstov. Programmnye Produkty i Sistemy. 2014;2:114—117. (in Russian).

11. **Kozlov P.Yu.** Sravnenie Chastotnogo i Vesovogo Algoritmov Avtomaticheskogo Analiza Dokumentov. Nauchnoe Obozrenie. 2015;14:245—250. (in Russian).

12. **Protasov S.** Grammatika Svyazey Link Grammar. [Elektron. Resurs] http://sz.ru/parser/doc/ (Data Obrashcheniya 10.07.2017). (in Russian).

13. **Borisov V.V., Fedulov A.S., Zernov M.M.** Osnovy Teorii Nechetkih Mnozhestv. M.: Goryachaya Liniya–Telekom, 2014. (in Russian).

14. **Gimarov V.A.** Metody i Avtomatizirovannye Sistemy Dinamicheskoy Klassifikatsii Slozhnyh Tekhnogennyh Ob′ektov: Avtoref. Diss. ... Doktora Tekhn. Nauk. M., 2004.

Сведения об авторах

Борисов Вадим Владимирович — доктор технических наук, профессор кафедры вычислительной техники Смоленского филиала НИУ «МЭИ», e-mail: vbor67@mail.ru

Дли Максим Иосифович — доктор технических наук, заведующий кафедрой менеджмента и информационных технологий в экономике, зам. директора по научной работе Смоленского филиала НИУ «МЭИ», e-mail: midli@mail.ru

Козлов Павел Юрьевич — аспирант кафедры прикладной математики НИУ «МЭИ», e-mail: originaldod@gmail.com

Information about authors

Borisov Vadim V. — Dr.Sci. (Techn.), Professor of Computer Engineering Dept., Branch of NRU MPEI in Smolensk, e-mail: vbor67@mail.ru

Dli Maksim I. — Dr.Sci. (Techn.), Head of Management and Information Technology in Economy Dept., Deputy Director for Research, NRU MPEI in Smolensk, e-mail: midli@mail.ru

Kozlov Pavel Yu. — Ph.D.-student of Applied Mathematics Dept., NRU MPEI, e-mail: originaldod@gmail.com

Статья поступила в редакцию 03.08.2017