DOI: 10.37791/2687-0649-2020-15-3-36-45

Rubrication of text documents based on fuzzy difference relations

***V. Borisov*[1], *O. Bulygina*[1*], *M. Dli*[1], *P. Kozlov*[1]**
[1]*The Branch of National Research University MPEI in Smolensk, Russia*
** baguzova_ov@mail.ru*

Resume. One of the key areas of informatization of public authorities is to develop and implement the systems of automated processing the electronic appeals (applications, complaints, suggestions) of individuals and legal entities that arrive on official websites and portals of government. The rubrication plays an important role in solving this problem. It consists in the appeals' distribution according to thematic rubrics determining the directions of the activity of departments carrying out processing and preparation of the corresponding response. The results of the analysis of the specific features of such text messages (small size, markup lack, the errors' presence, thesaurus unsteadiness, etc.) confirmed the impossibility of using traditional approaches to rubrication and justified the feasibility of using data mining methods. The article proposes a new approach to the analysis and rubrication of electronic unstructured text documents arrived on official websites and portals of public authorities. It involves the formation of a tree-like structure of the rubric field, based on fuzzy relationships of differences between the syntactic characteristics of documents. The analysis is based on determining the fuzzy correspondence of these documents by their syntactic characteristics with the values of the clusters' centers. It is carried out sequentially from the root to the leaves of the constructed fuzzy decision tree. The proposed rubrication method is programmatically implemented and tested in the automated processing and analysis of appeals (applications, complaints and suggestions) of citizens entering the Administration of Smolensk Region. This made it possible to ensure prompt and high-quality updating of rubrics and document analysis under conditions of non-stationary composition of the thesaurus and the importance of rubric words.

Keywords: rubrication, electronic unstructured document, syntactic characteristic, fuzzy difference, hierarchical clustering, fuzzy correspondence

For citation: Borisov V., Bulygina O., Dli M., Kozlov P. Rubrication of text documents based on fuzzy difference relations. *Prikladnaya informatika* = Journal of Applied Informatics, 2020, vol.15, no.3, pp. 36-45 (in Russian) DOI: 10.37791/2687-0649-2020-15-3-36-45.

Рубрицирование текстовых документов на основе нечетких отношений различия

В. В. Борисов[1]***, О. В. Булыгина***[1*]***, М. И. Дли***[1]***, П. Ю. Козлов***[1]
[1]*Филиал национального исследовательского университета «МЭИ» в г. Смоленске, Россия*
** baguzova_ov@mail.ru*

Аннотация. Одним из ключевых направлений информатизации деятельности органов государственной власти является разработка и внедрение систем автоматизированной обработки электронных обращений (заявлений, жалоб, предложений) физических и юридических лиц, поступающих на официальные веб-сайты и порталы органов власти федеральных округов, администраций областей и других территориальных образований. Важную роль при решении данной задачи играет рубрицирование, которое заключается в распределении обращений по тематическим рубрикам, определяющих направления деятельности департаментов, осуществляющих их обработку и подготовку соответствующего ответа. Результаты анализа специфических особенности таких текстовых сообщений (небольшой размер, отсутствие разметки, наличие ошибок, нестационарность тезауруса и т. п.) подтвердили невозможность применения традиционных подходов к рубрицированию и обосновали целесообразность применения методов интеллектуального анализа данных. В статье предложен новый подход к анализу и рубрицированию электронных неструктурированных текстовых документов, поступающих на официальные веб-сайты и порталы органов государственной власти. Он предполагает формирование древовидной структуры рубричного поля, основанной на нечетких отношениях различия между синтаксическими характеристиками документов. Анализ основывается на определении нечеткого соответствия этих документов по синтаксическим характеристикам со значениями центров кластеров, проводимого последовательно от корня к листьям построенного нечеткого дерева решений. Предлагаемый метод рубрицирования программно реализован и апробирован при автоматизированной обработке и анализе обращений (заявлений, жалоб и предложений) граждан, поступающих в Администрацию Смоленской области. Это позволило обеспечить оперативную и качественную актуализацию рубрик и анализ документов в условиях нестационарности состава тезауруса и значимости слов рубрик.

Ключевые слова: рубрицирование, электронный неструктурированный документ, синтаксическая характеристика, нечеткое различие, иерархическая кластеризация, нечеткое соответствие

Для цитирования: *Борисов В. В., Булыгина О. В., Дли М. И., Козлов П. Ю.* Рубрицирование текстовых документов на основе нечетких отношений различия // Прикладная информатика. 2020. Т. 15. № 3. С. 36–45. DOI: 10.37791/2687-0649-2020-15-3-36-45.