

М. И. Дли, докт. техн. наук, зам. директора филиала Национального исследовательского университета «МЭИ» в г. Смоленске, midli@mail.ru

О. В. Булыгина, канд. экон. наук, доцент филиала Национального исследовательского университета «МЭИ» в г. Смоленске, baguzova_ov@mail.ru

П. Ю. Козлов, канд. техн. наук, ассистент филиала Национального исследовательского университета «МЭИ» в г. Смоленске, originaldod@gmail.com

Формирование структуры интеллектуальной системы анализа и рубрицирования неструктурированной текстовой информации в различных ситуациях¹

Анализ электронных текстовых документов, написанных на естественном языке, является одной из важнейших задач, реализуемых в системах автоматизированного анализа лингвистической информации. Известно, что такие документы могут характеризоваться различными параметрами: размер, наличие структуры, частота упоминания ключевых слов и т. п. Проведенный анализ показал невозможность построения единой модели для рубрицирования неструктурированных текстовых документов в различных ситуациях. Для решения указанной проблемы предложен мультимодельный подход к рубрицированию, отличающийся комбинированным использованием интеллектуальных и вероятностно-статистических методов анализа текстовых документов. Конкретная модель выбирается с использованием нечетко-логических алгоритмов на основе предложенных характеристик. Его применение позволит повысить точность отнесения электронных текстовых документов к конкретным рубрикам с учетом их специфики и различных целей практического применения в организации.

Ключевые слова: электронный неструктурированный текстовый документ, рубрицирование, мультимодельный подход, растущие пирамидальные сети, нечетко-логические алгоритмы.

Введение

В настоящее время отмечается активный рост объема текстовой информации в электронной форме как в сети Интернет, так и в глобальных и локальных хранилищах данных и знаний [2, 3].

Проблема обработки электронной текстовой информации относится к числу приоритетных задач, решить которую необходимо для достижения цели перехода к цифровой экономике, сформулированной в Стратегии развития информационного общества в Российской Федерации на 2017–2030 годы. В программе «Цифровая экономика», утвержденной Правительством в июле 2017 г., рассматривается в качестве одной из важнейших

¹ Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 18-01-00558.