УДК 681.3

ИССЛЕДОВАНИЕ ХАРАКТЕРИСТИК РУБРИКАЦИИ ТЕКСТОВ

© 2022 Я. Е. Львович, Ю. П. Преображенский, Е. Ружицкий

Воронежский государственный технический университет (Воронеж, Россия)
Воронежский институт высоких технологий (Воронеж, Россия)
Паньевропейский университет (Братислава, Словакия)

В статье исследуются характеристики рубрикации текстов. Показаны основные применяемые методы, ключевые этапы, связанные с тематическим представлением текстов.

Ключевые слова: анализ текстов, информация, принятие решений.

Процессы обмена информацией и организации знаний во многих случаях рассматриваются в задачах, связанных с классификацией различных видов информации [1, 2]. К настоящему времени разработано большое число подходов, позволяющих проводить решение указанных задач автоматическим образом.

Виды применяемых подходов указаны на рисунке 1 – это базирующихся на знаниях методы, а также методы машинного обучения.

В первом методе анализ рубрикаторов осуществляется со стороны экспертов. Во втором методе предварительным образом осуществляют публикацию коллекции документов [3, 4].

На ее основе формируется процедура, позволяющая проводить классификацию документов, в которых применяется машинное обучение. Во многих случаях исследователи прибегают к векторным моделям, например, SVM. С точки зрения практики, структура рубрикаторов и тематическая классификация при этом не учитываются [5, 6].

Это происходит вследствие того, что применяется абстрактная векторная модель документов. Каким образом можно сделать преобразование к векторам пространство признаков текстовых данных?

Львович Яков Евсеевич – Воронежский государственный технический университет, доктор техн. наук, профессор, e-mail: office@vivt.ru.
Преображенский Юрий Петрович – Воронежский институт высоких технологий, канд. техн. наук, профессор, e-mail: petrovich@vivt.ru.
Ружицкий Евгений – Панъевропейский университет, канд. техн. наук, доцент, rush_evg_br53@yandex.ru.

В ряде случаев осуществляется процесс нормализации в рамках морфологического подхода. В ходе машинного обучения необходимо правильным образом осуществить процесс выбора по коллекции документов.

На рисунке 2 указаны свойства, которыми она должна обладать. Тогда результаты сравнения достигаются достоверным образом [7, 8].

Простота математических моделей, а также высокая скорость при функционировании характерна для метода Байеса. Для этапов рубрикации необходимо использовать весьма много вычислительных затрат, чтобы в методе ближайших соседей были обеспечены требуемые показатели эффективности [9].

Когда добавляются новые отрубрицированные примеры, тогда в классификаторе Роше можно осуществить быстрый пересчёт по взвешенным центроидам.

В задачах классификации широкая область применения связана с нейронными сетями. Исходя из того, какие значения переменных в пространстве признаков, данные будут разбиваться по группам [10, 11] в подходе, связанном с использованием деревьев решений.

Правила классификации в методе булевых функций приведены на рисунке 3. Вследствие того, что для предметной области строится абстрактная векторная модель, в подходе SVM, он может быть использован для того, чтобы решать широкий круг задач в машинном обучении.

На рисунке 4 показаны ключевые этапы, связанные с тематическими представлениями текстов. На рисунке 5 приведено сравнение между методами.

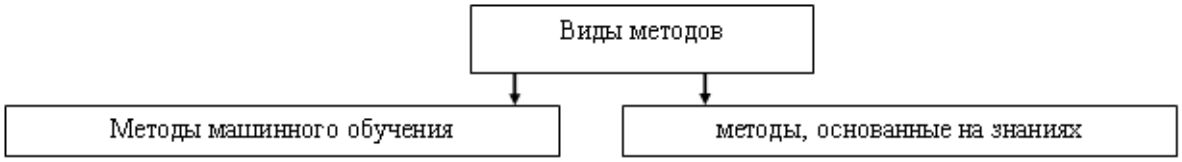

Рисунок 1. Виды методов

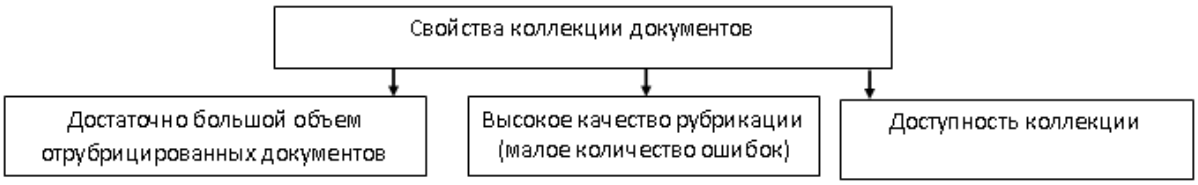

Рисунок 2. Свойства коллекции документов

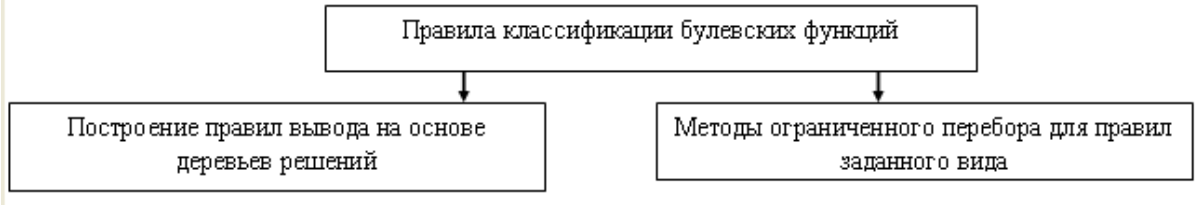

Рисунок 3. Правила классификации булевских функций

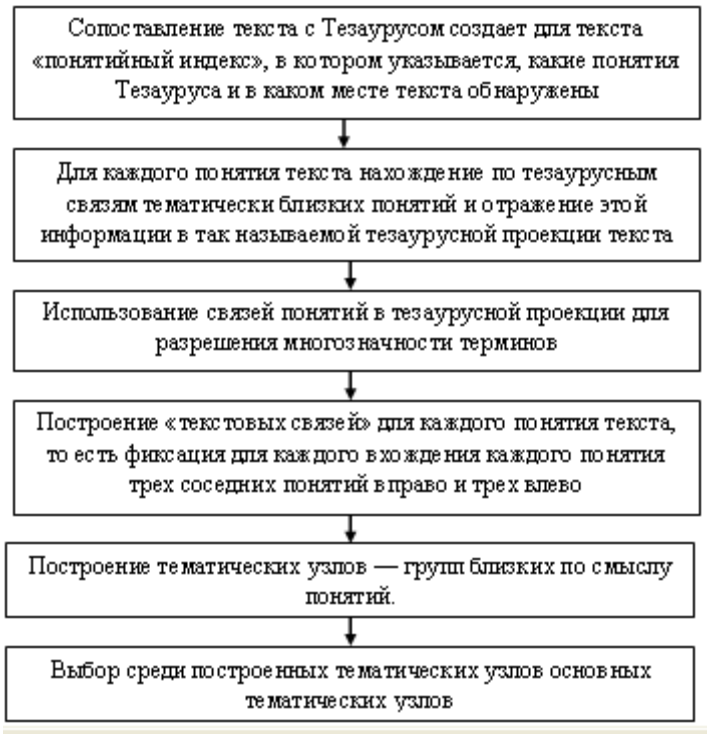

Рисунок 4. Ключевые этапы, связанные с тематическим представлением текстов

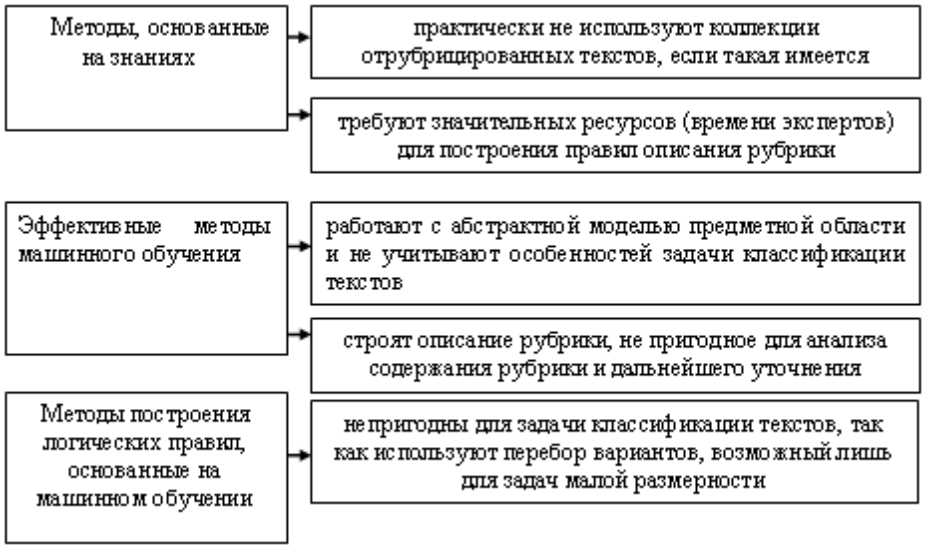

Рисунок 5. Сравнительный анализ методов

СПИСОК ЛИТЕРАТУРЫ

1.Преображенский А. П. Тематический анализ текстовой информации на основе частотных характеристик / А. П. Преображенский, Д. В. Меняйлов, Е. И. Чопорова // Моделирование, оптимизация и информационные технологии. – 2021. – Т. 9. – № 1 (32).

2. Меняйлов Д. В. Сравнительный анализ результатов, полученных при решении задачи анализа тональности текста с помощью сверточной и рекуррентной нейронных сетей / Д. В. Меняйлов, А. П. Преображенский // Моделирование, оптимизация и информационные технологии. – 2021. – Т. 9. – № 4 (35).

3. Диденко С. С. Применение мультиагентных технологий в контекстно-ориентированной среде компонента умного дома / С. С. Диденко // Моделирование, оптимизация и информационные технологии. – 2021. – Т. 9. – № 2 (33). – С. 18-19.

4. Lvovich I. Ya. Modeling of information processing in the internet of things at agricultural enterprises / I. Ya. Lvovich, Ya. E. Lvovich, A. P. Preobrazhenskiy, Yu. P. Preobrazhenskiy, O. N. Choporov // IOP Conference Series: Earth and Environmental Science. Krasnoyarsk Science and Technology City Hall of the Russian Union of Scientific and Engineering Associations. – 2019. – С. 32029.

5. Машков В. Г. Предварительная оценка вероятности принятия правильного решения в автоматизированных системах управления / В. Г. Машков, В. А. Малышев, Ю. В. Никитенко // Моделирование, оптимизация и информационные технологии. – 2021. – Т. 9. – № 3 (34). – С. 12-13.

6. Lvovich I. Optimization of the subsystem for the movement of electronic documents in educational organization / I. Lvovich, A. Preobrazhenskiy, Y. Preobrazhenskiy, Y. Lvovich, O. Choporov // Proceedings – 2021 1st International Conference on Technology Enhanced Learning in Higher Education, TELE 2021. – 1. – 2021. – С. 328-332.

7. Борзова А. С. Особенности построения системы принятия решений при многовариантной оптимизации структуры цифрового управления логистическим процессом в организационной системе на основе имитационного моделирования / А. С. Борзова, В. В. Муха // Моделирование, оптимизация и информационные технологии. – 2021. – Т. 9. – № 3 (34). – С. 15-16.

8 Печенкин В. В. Моделирование динамики серверной нагрузки стохастическими сетями Петри с приоритетами (на примере системы видеоконференцсвязи) / В. В. Печенкин, А. Т. Х. Аль-Хазраджи, С. С. Гельбух // Моделирование, оптимизация и информационные технологии. – 2021. – Т. 9. – № 1 (32). – С. 10-11.

9. Lvovich I. Managing developing internet of things systems based on models and algorithms of multi-alternative aggregation / I. Lvovich, A. Preobrazhenskiy, Y. Preobrazhen-

skiy, Y. Lvovich, O. Choporov // 2019 International Seminar on Electron Devices Design and Production, SED 2019 – Proceedings. – 2019. – С. 8798413.

10. Новосадов К. С. Анализ спектрально эффективных схем модуляции, применяемых в высокоскоростных системах радиосвязи / К. С. Новосадов // Моделирование, оптимизация и информационные технологии. – 2021. – Т. 9. – № 1 (32). – С. 20-21.

11 Lvovich I. Ya. Modelling of information systems with increased efficiency with application of optimization-expert evaluation / I. Ya. Lvovich, Ya. E. Lvovich, A. P. Preobrazhenskiy, Yu. P. Preobrazhenskiy, O. N. Choporov // Journal of Physics: Conference Series. International Scientific Conference «Conference on Applied Physics, Information Technologies and Engineering – APITECH-2019». Krasnoyarsk Science and Technology City Hall of the Russian Union of Scientific and Engineering Associations; Polytechnical Institute of Siberian Federal University. – 2019. – С. 33079.

CHARACTERIZATION OF STUDY TEXT RUBRICATIONS

© 2022 Ya. E. Lvovich, Yu. P. Preobrazhensky, E. Ruzhitsky

Voronezh State Technical University (Voronezh, Russia)
Voronezh Institute of High Technologies (Voronezh, Russia)
Pan-European University (Bratislava, Slovakia)

The paper examines the characteristics of the rubric of texts. The main methods used, the key stages related to the thematic presentation of texts are shown.

Keywords: text analysis, information, decision-making.