УДК 519.6+004.942

Задача обработки текстовых данных на примере коллекций выпускных квалификационных работ студентами ФМиИТ АлтГУ

Бабкина Н.С., Оскорбин Н.М., Половикова О.Н., Смолякова Л.Л.
Алтайский государственный университет, г. Барнаул
inf.asu@gmail.com, osk46@mail.ru, ponolgap@gmail.com, knaus.larisa@gmail.com

Аннотация

В работе рассматривается задача обработки текстовых данных на примере коллекций выпускных квалификационных работ кафедры информатики, которые подготовлены и защищены в последние годы студентами ФМиИТ АлтГУ.

Ключевые слова: анализ текстовых данных, тематический анализ коллекций, тематическое моделирование, тестирование программных средств.

Неотъемлемым звеном современного развития информационного общества являются исследования по созданию систем автоматизированной обработки текстов на естественном языке. Непрерывный рост объема информации, который нужно просмотреть, отобрать, проанализировать субъекту (человеку) заставляет использовать системы автоматической (или полуавтоматической) классификации, фильтрации, аннотирования, экспертной оценки и анализа обработки текстов. Несмотря на существенное множество зарекомендовавших себя на практике подходов и методов по обработки текстовых документов, данное направление исследования интенсивно развивается и осваивает все новые актуальные для теории и практики области приложений [1, 2].

В данной работе представлены результаты исследования методов тематического моделирования коллекции текстовых документов. В развитие подхода авторов [3] разработан алгоритм поиска документов коллекции по заданным темам и информационные технологии его поддержки. На коллекции выпускных квалификационных работ магистрантов и бакалавров кафедры информатики ФМиИТ АлтГУ проведено тестирование разработки и обоснованы рекомендации по ее использованию на практике и в учебном процессе.

На основании проведённого анализа можно сделать следующие выводы:

1. Литературный анализ методов тематической классификации коллекций текстовых документов и информационные технологии их поддержки развиты в настоящее время, однако проблемно ориентированные информационные технологии для решения практических задач классификации текстов требуют самостоятельных разработок.

2. Задачу автоматической классификации коллекции документов следует проводить в три этапа:

- для выбранной коллекции текстовых документов выделить базовые темы и для каждой из них сформировать список ключевых слов;
- провести предварительную обработку документов коллекции.
- произвести оценку значения классифицирующей функции принадлежности или не принадлежности каждого из исследуемых документов коллекции к выделенным темам.

Для реализации метода проведена разработка алгоритма классификации коллекции текстовых документов для применяемых информационных технологий. Предложен алгоритм поиска документов коллекции по заданным темам, который включает 5 шагов:

1. Формируем коллекцию текстовых документов, из которой необходимо найти документы по заданным темам.

2. Для заданных базовых тем, с помощью экспертов составляем список ключевых слов в нормальной форме.

3. Проводим операцию, обратную лемматизации в автоматизированном режиме, т.е. расширяем список ключевых слов для характеристики заданных базовых тем.

4. Выполняем в диалоговом режиме предварительную обработку документов коллекции, включая их очистку, уменьшение размерности, выделение информативных разделов. Формируем сравниваемый документ в текстовом формате.

5. Посчитываем значения классифицирующей функции принадлежности или не принадлежности каждого из исследуемых документов коллекции к выделенным темам и выделяем искомые документы коллекции, которые относятся к заданным темам.

Обосновано, что анализ текстовых документов можно проводить последовательно, или в процессе классификации уточнять решения предыдущих этапов. Это касается как составов коллекции документов, ключевых слов, так и пороговых значений классифицирующей функции.

Для оценки возможности использования разработки на практике проведена классификация ВКР кафедры информатики по трем заданным темам и выделены документы коллекции, которые к заданным темам не относятся. Методика тестирования состояла в сравнении автоматизированной классификации с классификацией документов коллекции, выполненной экспертами.

В результате проведенных исследований показано, что точность классификации по относительной достоверности составляет более 70%. Полученная оценка достоверности классификации позволяет утверждать о возможности использования разработки не только на практике решения подобных задач, но и в учебном процессе. Результатом работы по данной тематике являются следующие положения:

1. Предложены информационные технологии классификации коллекции текстовых документов, апробированные на реальных данных, которые применимы к текстовым данным с малым количеством информативных переменных.

2. Проведено тестирование информационных технологий тематической классификации коллекции текстовых документов на примере ВКР кафедры информатики, защищённые в период 2016-2018 г.

3. Подготовлены учебные материалы для лабораторной работы по анализу текстовых документов в составе курса “Программирование” для бакалавров ФМиИТ АлтГУ.

Список литературы

1. Ерланова Р.Е., Нугуманова А.Б., Жантасова Ж.З., Байбурин Е.М. Тематическое моделирование текстовых учебных материалов по информатике средствами языка R // Известия Алтайского государственного университета. — 2018. — № 2.

2. Федотов А.М., Прозоров О.В., Федотова О.А., Бапанов А.А. О подходе к тематической классификации документов // Вестн. НГУ. Серия: Информационные технологии. — 2017. — Т. 15, № 1. — С. 79–88.

3. Половикова О.Н., Бабкина Н.С., Смолякова Л.Л. Прикладное направление тематического моделирования в учебном процессе // МАК: Математики – Алтайскому краю

сборник трудов всероссийской конференции по математике с международным участием. / Главный редактор профессор Н.М. Оскорбин. — Барнаул : Изд-во Алт. ун-та, 2018. — С. 139–142.