Отзыв

На автореферат диссертации Карповича Сергея Николаевича «Математическое и программное обеспечение вероятностного тематического моделирования потока текстовых документов», представленной к защите на соискание ученой степени кандидата технических наук по специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Актуальность диссертационного исследования определяется его направленностью на развитие методов, алгоритмов и программных средств для решения практических задач по анализу и обработке текстов на естественном языке. Актуальность научных задач разработки методов и инструментов, улучшающих существующие подходы к автоматической обработке естественного языка, обусловлена постоянным ростом информационного обмена с использованием текстовых данных.

Цель диссертационного исследования состоит в разработке открытых программно-алгоритмических инструментов, позволяющих анализировать коллекции и потоки текстовых документов. Цель достигнута за счет разработанного метода построения вероятностной тематической модели через обучение с учителем, созданного алгоритма многозначной классификации, представленного метода определения тем для новых слов в динамической вероятностной тематической модели.

Научная новизна результатов заключается в следующем:

1. Создан русскоязычный корпус текстов SCTM-ru, позволяющий исследовать алгоритмы вероятностного тематического моделирования и отличающийся от других корпусов наличием оригинального текста документа и метатекстовой разметки: автор, время описанных событий, тема.
2. Разработан метод расчета матриц BTM на основе обучения с учителем (авторами документов), учитывающий заданные связи между документами и темами.
3. Предложен алгоритм многозначной классификации текстовых документов ml-PLSI, основанный на вероятностном тематическом моделировании, заключающийся в использовании матрицы «слово-тема» BTM для классификации документов.
4. Предложен метод определения тем «нового слова», основанный на использовании произведения Адамара тематических векторов документов, содержащих это слово, позволяющий определять вектора тем для «новых слов» в потоке текстовых документов при построении динамической тематической модели.

Достоверность полученных автором результатов подтверждается анализом публикаций по тематике вероятностного тематического моделирования, апробацией результатов на научных мероприятиях различного уровня оценкой и экспериментальным сравнением результатов с существующими аналогами.

Представленные в работе результаты представляют теоретическую и практическую значимость для развития методов и инструментальных средств автоматизированной классификации текстовых документов. С точки зрения теоретической значимости, интерес представляют алгоритм многозначной классификации текстовых документов, позволяющий проводить анализ потока текстовых документов и метод определения тем новых слов использующий произведение Адамара для динамического перестроения вероятностной тематической модели. С точки зрения практической значимости интерес представляют разработанный корпус текстовых документов и использование микросервисной

архитектуры при построении комплекса программных средств для вероятностного тематического моделирования.

Замечания по содержанию автореферата:

– В тексте автореферата отсутствуют пояснения к указанным на рисунках 2 и 6 числовым обозначениям, определяющим, вероятно, порядок выполнения операций или информационного взаимодействия модулей разработанного программного комплекса.

– Характеризуя научную новизну полученных результатов диссертационного исследования (п.5, стр. 5), автор отмечает обеспечение в рамках созданного прототипа комплекса программных средств возможности варьирования способов решения конкретных практических задач анализа потока текстовых документов. На стр.13 автореферата (абз.3) также указано, что любой микросервис, входящий в состав комплекса, может быть изменен без нарушения целостности системы. Однако в тексте автореферата не представлено, каким образом выбирается «подходящий» способ решения задачи, кто и как может изменять микросервисы.

– В тексте автореферата неоднократно (стр. 5, 13) упоминается эффективность алгоритмов, но не приводятся ее количественные характеристики.

Указанные недостатки не являются принципиальными и не снижают значимости проведенного исследования. Автореферат дает достаточно полное представление о содержании выполненной диссертационной работы и основных полученных результатах. В целом диссертационная работа С.Н. Карповича «Математическое и программное обеспечение вероятностного тематического моделирования потока текстовых документов» является законченной научно-исследовательской работой и соответствует требованиям п. 9 Положения о присуждении ученых степеней, утвержденного постановлением Правительства Российской Федерации от 24 сентября 2013 № 842, предъявляемым к кандидатским диссертациям, а ее автор заслуживает присуждения ученой степени кандидата технических наук по специальности 05.13.11 – «Математическое и программное обеспечение вычислительных машин, комплексов и компьютерных сетей».

Отзыв составил:

Врио директора ФГБУН Институт
информатики и ма
моделирования те
процессов Кольск

д.т.н. Олейник Андрей Григорьевич

09.10.2017

Почтовый адр иты, ул. Ферсмана 24А
Тел.: +7 81555
Эл. почта: oley