ОТЗЫВ

официального оппонента на диссертационную работу Бондарчука Дмитрия Вадимовича на тему «Алгоритмы интеллектуального поиска на основе метода категориальных векторов», представленную на соискание ученой степени кандидата физико-математических наук по специальности 05.13.17 — «Теоретические основы информатики»

Актуальность темы диссертации

В настоящее время происходит интенсивное развитие и широкое внедрение информационных технологий, расширение масштабов хранения и обработки больших объемов данных, развитие Интернета вещей, баз данных, хранилищ и ряда других информационных направлений. Это приводит к необходимости решения задач автоматизированной обработки информации на предмет выявления семантической близости документов, их классификации, кластеризации и обработки. Обычно целями такой обработки является выявление противоречивости в хранимой информации, удаление этой противоречивости и устранение дублирования семантически близкой информации и ряда аналогичных задач.

В современной научной литературе решение задачи разработки эффективных методов семантического анализа и обеспечения высокой релевантности поиска хранимых и обрабатываемых информационных массивов данных на основе учета взаимодействия элементов информации между собой и отношения пользователя к знанию не получила исчерпывающего решения. Все это подтверждает ***актуальность темы*** диссертационной работы Бондарчука Д.В., посвященной улучшению качества интеллектуального анализа текстовых данных, разрешению лексической многозначности и разработке методов, обеспечивающих непустой результат при произвольной обучающей выборке.

Степень обоснованности научных положений, выводов и рекомендаций

Диссертационная работа Бондарчука Д.В., состоящая из введения, четырех глав, заключения и библиографического списка, изложена на 141 страницах машинописного текста, включает семнадцать рисунков и двадцать семь таблиц.

Введение соответствует формальным требованиям. В нем обоснована актуальность темы диссертации, дано краткое описание ее сути, изложены цель и задачи исследования, научная новизна и практическая ценность полученных результатов.

В *первой* главе работы приведен обзор литературы, использованной при проведении диссертационного исследования. Приведено обоснование того, что при разработке алгоритмов классификации перспективным подходом является использование комплекса подходов. Для очистки данных от шумов предлагается использовать стеммер Портера и семантическое ядро, для представления данных выбрана векторная модель, для выделения сематического ядра — модификация подхода ЛСА.

Во *второй* главе предложен эффективный метод, позволяющий из любой непустой выборки сформировать персональные рекомендации. Предложен способ формирования семантического пространства на основе ортогонального разложения матрицы корреспонденций термов, которая подвергается ортогональному разложению. Проведено сравнение ортогонального разложения МКТ и сингулярного разложения ТДМ. Доказано, что при использовании предлагаемого метода термы, содержащиеся в коротких документах, сохраняются в семантическом ядре.

В *третьей* главе дано описание применения семантической близости термов при обучении классификатора путем повторного взвешивания весов термов векторной модели. Эта модель помогает решить проблему лексической неоднозначности терминов, а также находит скрытые семантические связи между документами, сравнивая семантически близкие

термы. Проанализирована возможность применения словарей и тезаурусных баз данных для вычисления семантической взаимосвязи между термами.

В *четвертой* главе описываются результаты экспериментов по исследованию эффективности разработанных в диссертации моделей, методов и алгоритмов. Приведены и прокомментированы табличные данные, убедительно показывающие выигрыш по качеству поиска, обеспечиваемый от использования полученных результатов на основании мер f-measure и purity. Показано что время генерации ответа с помощью предложенного метода в большинстве случаев заметно меньше, чем у других методов.

В *заключении* кратко подводятся итоги диссертационного исследования, представляются отличия диссертационной работы от выполненных работ других авторов, даются рекомендации по использованию полученных результатов и рассматриваются перспективы дальнейшего развития темы.

Сформулированные в диссертационной работе научные положения, выводы и рекомендации являются вполне обоснованными.

Научная новизна, достоверность и практическая значимость

Научная новизна работы определяется предложенным автором способом формирования семантического пространства, основанным на ортогональном разложении матрицы корреспонденций термов (МКТ), методом перехода к категориальным векторам с переопределением исходных весов термов и с помощью учета семантической взаимосвязи между термами.

Достоверность результатов, полученных автором, подтверждается утверждениями, снабженными математически строгими доказательствами, обширными вычислительными экспериментами, в том числе на стандартных наборах данных, а также апробацией результатов работы на научно-технических конференциях и семинарах, схожих по своему направлению с тематикой исследования.

Теоретическая ценность работы состоит в том, что в ней проведен сравнительный анализ свойств сингулярного разложения терм-документной матрицы (ТДМ) и ортогонального разложения МКТ. Автором предложено доказательство того, что термы, содержащиеся только в коротких документах, отбрасываются при использовании сингулярного разложения ТДМ, но учитываются при использовании предлагаемого подхода.

Практическая значимость работы заключается в том, что результаты работы являются основой для разработки поисковых систем, использующих интеллектуальный анализ текстовых данных. Результаты доведены до готовых к непосредственному применению алгоритма построения семантического ядра для текстового классификатора и способа определения семантической близости термов.

Замечания по диссертационной работе

1. Приведенный в первой главе диссертации обзор состояния исследований по тематике, близкой к диссертации, представляется в определенной степени обширным, его можно немного сократить без большого ущерба для работы.

2. В соответствии с описанием алгоритма (гл.2, с. 74) разбиение текстов по категориям для обучающей выборки должно проводиться экспертами вручную, что требует значительных затрат времени.

3. Недостаточно обоснован выбор расстояния Хэмминга для уточнения вхождения терма в текст (на стр. 76-77).

4. В диссертации приведены табличные данные, убедительно показывающие выигрыш, обеспечиваемый от использования полученных результатов на основании мер f-measure и purity. В автореферате данные приведены очень кратко и фактически не прокомментированы.

5. В диссертации присутствуют некоторые стилистические и пунктуационные ошибки, в частности, при написании формул, нумерованных списков и правых причастных оборотов.

Указанные замечания не влияют существенно на общую положительную оценку работы в целом.

Заключение по диссертации в целом

Результаты являются новыми и достаточно опубликованы в публикациях из перечня ВАК, в индексируемых в Scopus, представлены на научно-технических конференциях. Автореферат отражает основные результаты.

Диссертация является завершенной ***научно-квалификационной*** работой, в которой решена актуальная задача разработки эффективного алгоритма интеллектуального анализа текстов, имеющей ***существенное значение*** в области автоматизированной обработки информации.

Диссертация соответствует критериям пункта 9 действующего положения о присуждении ученых степеней, утвержденного постановлением Правительства РФ от 24.09. 2013 № 842 (ред. от 30.07.2014), а Дмитрий Вадимович Бондарчук заслуживает присуждения ученой степени кандидата физико-математических наук по специальности 05.13.17 – «Теоретические основы информатики».

Заведующий кафедрой «Информационные и вычислительные системы» федерального государственного бюджетного образовательного учреждения высшего образования «Петербургский государственный университет путей сообщения Императора Александра I» (ФГБОУ ВО ПГУПС), доктор технических наук, профессор Анатолий Дмитриевич Хомоненко, 190031, Санкт-Петербург, Московский пр., 9, ФГБОУ ВО ПГУПС, телефон; (812)457-80-23, E-mail: khomonenko@pgups.ru, khomon@mail.ru

«20» февраля 2017 г._____ А.Д. Хомоненко

Подпись руки А.Д. Хомоненко
удостоверяю.
Начальник Службы управления персоналом
Университета Г.Е. Егоров
20 февраля 2017 г.