УТВЕРЖДАЮ
Первый проректор федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет «Высшая школа экономики»
д.э.н., профессор Вадим Валерьевич Радаев

«16» февраля 2017 г.

ОТЗЫВ

ведущей организации на диссертацию Бондарчука Дмитрия Вадимовича «Алгоритмы интеллектуального поиска на основе метода категориальных векторов», представленную на соискание учёной степени кандидата физико-математических наук по специальности 05.13.17 – теоретические основы информатики

Актуальность темы диссертации

Интеллектуальный анализ текстовых данных в последние десятилетия получил широкое распространение в связи с колоссальным увеличением количества документов, хранящихся в электронном виде, необходимостью их систематизации. Исследованию интеллектуального анализа текстовых данных и развитию методов автоматической классификации и кластеризации посвящены десятки работ российских и зарубежных авторов. Вместе с тем, существующие алгоритмы не всегда корректно решают задачу интеллектуального поиска, особенно при неравномерном распределении данных по категориям. Таким образом, актуальной является задача улучшения качества интеллектуального анализа текстовых данных за счет учета семантической и лексикографической взаимосвязи термов, и решения проблемы лексической многозначности и разработки методов, обеспечивающих непустой результат для любой обучающей выборки.

Структура и содержание диссертации. Диссертация состоит из введения, четырех глав, заключения, списка литературы и приложений. Объем диссертации составляет 141 страницу, список литературы содержит 124 наименования.

Во введении обоснована актуальность темы диссертации, изложены цель и задачи исследования, научная новизна и практическая ценность полученных результатов. В

первой главе дан обзор современных методов и подходов интеллектуального анализа текстовых данных.

Во второй главе предлагается новый алгоритм подбора персональных рекомендаций, который на любой запрос пользователя дает непустую выдачу, отсортированную по степени релевантности запросу пользователя.

Предлагаемый алгоритм обучения состоит из следующих этапов: обработка текстов с использованием стеммера и определение частот вхождений термов в документы; уменьшение количества термов с использованием сингулярного разложения матрицы корреспонденций термов (МКТ); вычисление категориальных векторов документов базы данных вакансий и пользовательского запроса; вычисление коэффициентов близости между пользовательским запросом и данными базы, сортировка по убыванию и выбор нескольких первых элементов.

В диссертации проведен сравнительный анализ свойств сингулярного разложения терм-документной матрицы (ТДМ), который используется в стандартном методе латентно-семантического анализа, и ортогонального разложения МКТ.

В третьей главе для решения проблем синонимии и полисемии предложена векторная модель представления знаний, использующая семантическую близость термов. Для вычисления семантической близости термов используется авторская адаптация расширенного алгоритма Леска. Так же предлагается способ вычисления семантической близости, основанный построении контекстного множества, т.е. множества слов, связанных с заданным термином. Для построения контекстного множества используется матрица корреспонденций термов.

В четвертой главе систематизированы результаты экспериментов по исследованию эффективности разработанных в диссертации моделей, методов и алгоритмов.

Основные результаты диссертационного исследования

являются новыми и состоят в следующем:

1. Разработан метод отображения текста в семантическое пространство, обеспечивающие компактное представление текстового документа в оперативной памяти на основе матрицы корреспонденций термов, которая подвергается ортогональному разложению.
2. Доказано, что термы, содержащиеся только в коротких документах, отбрасываются при использовании сингулярного разложения ТДМ, но учитываются при использовании предлагаемого подхода.
3. Разработан алгоритм интеллектуального анализа текстов, гарантирующий непустой результат независимо от распределения обучающей выборки по категориям на

основе использования вычисления категориальных векторов.

4. Для учета особенностей языка предложены: метод перевзвешивания термов векторной модели с помощью вычисления их семантической взаимосвязи друг с другом на основе авторской версии алгоритма Леска и статистический метод вычисления семантической близости термов, основанный на сборе контекстных множеств термов.

Апробация работы и публикации

Результаты диссертационных исследований докладывались на научных семинарах, международных и всероссийских конференциях, материалы диссертации достаточно полно представлены в 10 статьях, опубликованных соискателем, в том числе в 5 статьях в журналах, входящих в список изданий, рекомендованных ВАК, и 2 статьях в изданиях, индексируемых в международных базах данных WoS и Scopus.

Обоснованность научных положений и выводов, сформулированных в диссертации

Теоретические утверждения математически корректно обоснованы с использованием аппарата современной линейной алгебры; эффективность предложенных алгоритмов подтверждается значительным объемом проведенных численных экспериментов, в том числе на известных наборах тестовых данных, а также опытом практического внедрения результатов на нескольких предприятиях.

Соответствие содержания диссертации автореферату и указанной специальности

Диссертационное исследование соответствует специальности 05.13.17 – теоретические основы информатики. Автореферат диссертации правильно отражает ее основное содержание, научную новизну, выводы и другие ключевые моменты.

Теоретическая и практическая значимость. В работе разработан комплекс теоретически обоснованных и эвристических методов для решения проблемы организации эффективного поиска текстовых данных. В том числе, дано математическое обоснование применения нормированной терм-документной матрицы через исследование сравнительных свойств сингулярных разложений нормированной и ненормированной терм-документных матриц.

Практическая значимость состоит в том, что предложенные методы и подходы позволяют повысить скорость и качество интеллектуального поиска в системах автоматической рубрикации, системах формирования персональных рекомендаций и системах интеллектуального поиска.

Рекомендации по использованию результатов и выводов, приведенных в диссертации. Разработанные алгоритмы позволяют производить поиск, классификацию и формировать персональные рекомендации пользователю, гарантируя получение непустой выдачи на любой запрос пользователя, в том числе и в случае неравномерного распределения данных по категориям. Представленные в диссертационной работе подходы формирования персональных рекомендаций могут быть использованы для построения коммерческих и свободных поисковых систем, тезаурусов, систем родительского контроля, систем фильтрации спама и пр. Кроме того, результаты диссертационного исследования могут быть использованы при разработке специального учебного курса по направлению подготовки 09.03.01 «Информатика и вычислительная техника».

Замечания по диссертационной работе

1. В работе для получения основы слова используется алгоритм стемминга Портера, который хорошо работает на английском языке, но не является оптимальным для русского языка, для которого больше подходит стеммер Snowball.

2. В выводах по второй главе (стр. 83) упоминается, что разработанный метод оптимизирован для реляционных хранилищ, однако, подтверждений этому факту автором не приводится.

3. В вычислительных экспериментах (стр. 115) автором приведена сравнительная таблица потребления памяти разными алгоритмами. Как хранятся результаты, каким образом они были получены и на каких реализациях алгоритмов использованы, подробно не объяснено.

Перечисленные замечания не влияют на общую положительную оценку работы.

Вывод

Диссертационная работа Бондарчука Дмитрия Вадимовича «Алгоритмы интеллектуального поиска на основе метода категориальных векторов» является законченной научно-квалификационной работой, в которой содержится решение актуальной задачи совершенствования методов интеллектуального поиска текстовых данных, имеющей существенное значение для развития теоретических основ информатики. Результаты диссертации являются новыми и получены лично автором.

Диссертационная работа Бондарчука Дмитрия Вадимовича «Алгоритмы интеллектуального поиска на основе метода категориальных векторов», соответствует

требованиям пунктов 9-10 Положения о присуждении ученых степеней, утвержденного постановлением Правительства РФ от 24 сентября 2013 года № 842, предъявляемым к кандидатским диссертациям, а её автор заслуживает присуждения учёной степени кандидата физико-математических наук по специальности 05.13.17 – теоретические основы информатики.

Отзыв подготовлен доктором технических наук, профессором департамента прикладной математики федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет «Высшая школа экономики» Афанасьевым Валерием Николаевичем.

Отзыв рассмотрен и одобрен на заседании департамента прикладной математики федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет «Высшая школа экономики», протокол № 11 от «02» февраля 2017 года.

Сведения о ведущей организации: Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет «Высшая школа экономики» (НИУ ВШЭ)

Адрес: 101000 г. Москва, ул. Мясницкая, 20.
Тел.: (495) 771-32-32
Электронная почта: hse@hse.ru
Сайт: http://www.hse.ru

Руководитель
департамента «Прикладная математика»
кандидат техн. наук, доцент
Белов Александр Владимирович

Профессор
департамента прикладной математики
доктор техн. наук, профессор
Афанасьев Валерий Николаевич

Подпись заверяю

Специалист по кадрам

Исхакова Л.К.

16. 02. 2017