

ОТЗЫВ

официального оппонента

д.т.н., доцента, профессора факультета программной инженерии и компьютерной техники ФГАОУ ВО «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики»

Бессмертного Игоря Александровича

на диссертационную работу Потапенко Анны Александровны «Семантические векторные представления текста на основе вероятностного тематического моделирования», представленную на соискание ученой степени кандидата физико-математических наук по специальности 05.13.17 – Теоретические основы информатики

АКТУАЛЬНОСТЬ ТЕМЫ ИССЛЕДОВАНИЯ

Наблюдаемый в последние десятилетия экспоненциальный рост объемов оцифрованных текстовых документов вступает в противоречие с возможностями информационного поиска. Существующие поисковые машины, построенные на индексах, не учитывают семантику текстов, что приводит к низкой пертинентности результатов поиска поисковому запросу. Относительно новое направление в информационном поиске – векторное представление слов – позволяет представлять семантику слов посредством контекста, но обладает двумя основными недостатками: отсутствием разреженности и интерпретируемости результатов. Это обуславливает актуальность диссертационного исследования, посвященного обеспечению разреженности и интерпретируемости векторного представления текстов.

ДОСТОВЕРНОСТЬ И ОБОСНОВАННОСТЬ НАУЧНЫХ ПОЛОЖЕНИЙ, ВЫВОДОВ И ПРАКТИЧЕСКИХ РЕКОМЕНДАЦИЙ

Научные положения, выводы и практические рекомендации, сформулированные в диссертационной работе, достаточно хорошо обоснованы и не противоречат результатам исследований отечественных и зарубежных авторов. Кроме того, они подтверждены результатами вычислительных экспе-

риментов и внедрения разработанного математического и программного обеспечения.

ХАРАКТЕРИСТИКА НОВИЗНЫ РЕЗУЛЬТАТОВ

Научной новизной обладают следующие научные положения диссертационного исследования:

1. Обобщенный ЕМ-алгоритм для обучения тематических моделей, отличающийся использованием набора универсальных эвристик и обеспечивающий контроль перплексии, разреженности и робастности в процессе обучения модели.

2. Тематическая модель фоновых и предметных тем, отличающаяся набором аддитивных регуляризаторов и обеспечивающая интерпретируемость, различность и высокую разреженность предметных тем.

3. Алгоритм построения тематических векторных представлений слов, отличающийся отказом от гипотезы мешка слов в пользу моделирования совместной встречаемости слов в локальных контекстах и обеспечивающий интерпретируемость компонент векторов слов.

4. Алгоритм построения тематических векторных представлений слов и документов, отличающийся мультимодальным представлением каждой темы и обеспечивающий как улучшение определения семантической близости слов, так и возможность определения семантической близости между токенами различных модальностей.

Уровень новизны основных положений, выносимых на защиту, считаю достаточным.

СОДЕРЖАНИЕ РАБОТЫ

Глава 1 содержит обзор методов дистрибутивной семантики, позволяющих определять семантическую близость слов на основе низкоранговых разложений матриц совместной встречаемости слов в локальных контекстах.

Глава 2 содержит обзор вероятностных тематических моделей, позволяющих определять тематическую структуру коллекций документов, байесовских методов вывода и численных методов обучения таких моделей.

В главе 3 рассматривается ряд полезных эвристик, применимых к базовой тематической модели PLSA, и предлагается обобщенный ЕМ-подобный алгоритм, обобщающий данные эвристики. Рассматриваются взаимосвязи моделей PLSA, LDA, SWB, а также способов вывода в модели LDA. В результате удастся построить набор моделей, комбинирующих свойства разреженности тематических распределений и робастности к шуму в данных при несильном ухудшении основного критерия перплексии модели.

В главе 4 рассматривается проблема неединственности стохастического матричного разложения, и вводятся дополнительные критерии оптимизации – регуляризаторы тематической модели. Предлагается набор из 5 регуляризаторов, на основе которого разрабатывается модель предметных и фоновых тем, позволяющая отделять специфичные термины и фоновую лексику, повышая интерпретируемость, различность и разреженность тем. Дополнительно вводится регуляризатор отбора тем, позволяющий удалять из модели незначимые темы. Демонстрируется возможность совместного использования всех предложенных регуляризаторов.

В главе 5 предлагается формализация дистрибутивной гипотезы в рамках подхода аддитивной регуляризации тематического моделирования. Это позволяет построить тематические представления слов, сопоставимые по качеству на задачах определения семантической близости слов со стандартными нейросетевыми моделями семейства word2vec. При этом в экспериментах демонстрируются преимущества предлагаемого алгоритма: интерпретируемость компонент векторов, высокая разреженность, возможность дополнительной регуляризации.

Далее предлагается мультимодальная тематическая модель векторных представлений, в которой помимо слов в единое векторное пространство погружаются такие сущности как даты, авторы, категории и другие метаданные документов. На мультимодальном новостном корпусе Lenta.ru демонстрируется улучшение качества определения семантической близости слов, а также осмысленные семантические близости между датами и словами. На задаче предсказания семантической близости научных статей arXiv демонстрируется высокое качество предложенного метода по сравнению с другими методами, традиционно используемыми для данной задачи.

ПРАКТИЧЕСКАЯ ЗНАЧИМОСТЬ РЕЗУЛЬТАТОВ РАБОТЫ

Несмотря на то, что диссертация представлена на соискание ученой степени в области физико-математических наук, результаты, полученные соискателем, обладают несомненной практической значимостью, поскольку доведены до уровня алгоритмов, что позволяет создавать программы, реализующие предложенные методы. Кроме того, диссертация написана в хорошем стиле и снабжена достаточным количеством практических примеров, что позволяет мне рекомендовать издание данной диссертации в виде научной монографии, которая безусловно будет полезна научным работникам и аспирантам.

РЕЗУЛЬТАТЫ РАБОТЫ

Результаты работы опубликованы в 12 изданиях, 7 из которых являются изданиями, рекомендованными ВАК Минобрнауки РФ, 5 индексируются в базе SCOPUS. Состав публикаций отражает основные положения, выносимые на защиту.

ЗАМЕЧАНИЯ ПО ДИССЕРТАЦИОННОЙ РАБОТЕ

1. Из текста диссертации неясно, возможно ли дообучение модели на расширение обучающей коллекции вместо полного повторного обучения. Если дообучение невозможно, то возникает вопрос вычислительной сложности алгоритма при его практической реализации.

2. Непонятно, насколько автоматизирован процесс обучения модели. Судя по рекомендациям типа «Коэффициенты регуляризации для разреживания предметных тем рекомендуется включить только после того, как итерационный процесс начал сходиться и определились близкие к нулю элементы матриц Φ и Θ . Более раннее или более резкое разреживание может ухудшать перплексию» (стр.94-95), здесь требуется постоянное наблюдение за процессом обучения.

3. На стр.65 автор относит к шумовой компоненте текста редкие термины. С точки зрения математики это оправдано, но на практике редкие термины могут быть специфичными не только для отдельного документа, но и для темы.

4. На стр. 68-69 диссертации говорится о том, что три эвристики (S/P, D и R) могут рассматриваться независимо и образовывать 8 различных алгоритмов. В таком случае, по правилам комбинаторики, один из алгоритмов не должен использовать ни одной из эвристик, однако в диссертации такого варианта нет. Здесь же автор утверждает, что возможны и другие гибридные модели. Непонятно, имеются в виду комбинации тех же тех эвристик или каких-то других, а также роль человеческого фактора в синтезе результирующих моделей.

5. Утверждение «Упрощённая робастная модель не требует дополнительных затрат памяти или времени» кажется тривиальным. Следовало бы говорить об экономии памяти и времени, обеспечиваемой с помощью робастной модели.

6. Формула после (4.9) не отображается в PDF файле. Кроме того, следом в тексте приведены формулы, не имеющие нумерации, что затрудняет работу с текстом.

7. Фраза «По сравнению с байесовским подходом, ARTM радикально упрощает вывод алгоритма и позволяет комбинировать регуляризаторы в

произвольных сочетаниях» на стр. 83 вступает в противоречие с утверждением, упомянутом в замечании 4.

8. В тексте имеется незначительное число опечаток («деккоррелирования», стр. 100).

9. В автореферате присутствуют ссылки на источники.

ЗАКЛЮЧЕНИЕ

Представленная на рассмотрение диссертационная работа является завершенным научно-исследовательским трудом, выполненным автором на достаточно высоком уровне. Полученные автором результаты достоверны, выводы и заключения обоснованы и подтверждены вычислительными экспериментами.

Диссертационная работа относится к следующим областям исследования паспорта научной специальности 05.13.17 – Теоретические основы информатики:

П.4. Исследование и разработка средств представления знаний. Принципы создания языков представления знаний, в том числе для плохо структурированных предметных областей и слабоструктурированных задач; разработка интегрированных средств представления знаний, средств представления знаний, отражающих динамику процессов, концептуальных и семиотических моделей предметных областей.

П.5. Разработка и исследование моделей и алгоритмов анализа данных, обнаружения закономерностей в данных и их извлечения; разработка и исследование методов и алгоритмов анализа текста, устной речи и изображений.

П.6. Разработка методов, языков и моделей человекомашинного общения; разработка методов и моделей распознавания, понимания и синтеза речи, принципов и методов извлечения данных из текстов на естественном языке.

П.9. Разработка новых интернет- технологий, включая средства поиска, анализа и фильтрации информации, средства приобретения знаний и создания онтологии, средства интеллектуализации бизнес-процессов.

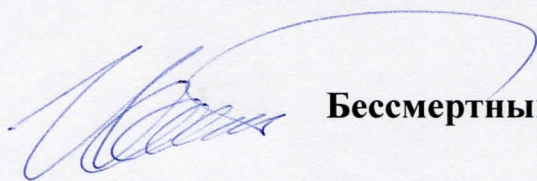
Диссертационная работа соответствует пп.9-14 Постановления Правительства РФ от 24 сентября 2013 г. № 842 "О порядке присуждения ученых степеней".

Автореферат отражает основные положения и полученные результаты, подробнее описанные в самой работе.

С учетом вышеизложенного, диссертация по всем критериям отвечает требованиям ВАК Минобрнауки РФ, предъявляемым к кандидатским диссертациям, а её автор, Потапенко Анна Александровна, заслуживает присуждения ученой степени кандидата физико-математических наук по специальности 05.13.17 – Теоретические основы информатики.

Официальный оппонент:

доктор технических наук (05.13.11), доцент, профессор факультета программной инженерии и компьютерной техники ФГАОУ ВО «Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики»



Бессмертный Игорь Александрович

«25» января 2019 г.

«Подпись Бессмертного И.А. заверяю»:

Ученый секретарь Ученого Совета

д.т.н., профессор



Марусина М.Я.