

«Утверждаю»
Директор Федерального
государственного бюджетного
учреждения науки
Института системного
программирования
им. В.П. Иванникова РАН
чл.-корр. РАН, д.ф.-м.н.



А.И. Аветисян

4 февраля 2019

ОТЗЫВ ВЫДУЩЕЙ ОРГАНИЗАЦИИ
Федерального государственного бюджетного учреждения науки «Институт системного
программирования им. В.П. Иванникова»
Российской академии наук
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

Построение векторных представлений для слов и документов – это первый этап при решении многих прикладных задач анализа текстов методами машинного обучения. Ключевым требованием, позволяющим использовать такие представления на практике, является близость векторов семантически близких слов. Другим требованием, все чаще предъявляемым к любым моделям в области автоматической обработки текстов, является интерпретируемость. Интерпретируемость моделей позволяет проводить аналогии с представлением знаний в человеческом мозге, легче встраивать в модели лингвистические предположения, лучше понимать причины возникновения ошибок.

Вероятностные тематические модели – это инструмент в области анализа текстов, позволяющий выделять интерпретируемые кластеры (темы), которые оказываются полезны в задачах семантического и информационного поиска, категоризации и реферирования документов, визуализации данных. Тем не менее, данные методы обладают рядом проблем, в частности, предположение о независимости слов в документах, неединственность и неустойчивость решения, недостаточное качество интерпретируемости тем, слабая разреженность построенных представлений. В данной работе разрабатываются модели в рамках подхода аддитивной регуляризации, нацеленные на преодоление данных ограничений.

Содержание

Диссертационная работа состоит из введения, пяти глав и заключения.

Во введении обоснована актуальность работы, установлены цели и задачи исследования, сформулированы основные выносимые на защиту положения, установлена научная новизна, обоснованы теоретическая и практическая значимость, достоверность результатов, перечислены основные публикации по теме диссертации.

В первой главе приведен обзор классических и нейросетевых методов семантического векторного представления текстов. Все подходы изложены в единой терминологии, многие из них удастся рассматривать как задачу низкорангового матричного разложения, при этом оптимизируемые функционалы и методы оптимизации различаются.

Во второй главе поставлена задача вероятностного тематического моделирования, и рассмотрены основные алгоритмы построения таких моделей. Для модели латентного размещения Дирихле приведены несколько альтернативных способов оценивания параметров, и показана взаимосвязь этих методов. Рассмотрен ЕМ-алгоритм в общем виде, приведены известные свойства о его сходимости.

В третьей главе предложено обобщенное семейство ЕМ-подобных алгоритмов, исследованы свойства разреженности, робастности к шумовым и фоновым терминам. Разработаны упрощенные робастные модели, позволяющие строить сильно разреженные распределения, не теряя в качестве по основному критерию правдоподобия моделей. Сделаны выводы об эффекте переобучения в тематических моделях.

В четвертой главе рассматривается и развивается подход аддитивной регуляризации тематических моделей. Предлагается новая формализация понятия интерпретируемости, на основе которой разрабатывается модель предметных и фоновых тем. Предлагается регуляризатор отбора тем. По результатам многочисленных экспериментов на двух коллекциях научных статей вырабатываются рекомендации о выборе коэффициентов регуляризации.

В пятой главе в рамках подхода аддитивной регуляризации формализуется гипотеза дистрибутивной семантики, что позволяет ослабить предположение о независимости слов в документах, являющееся одним из самых критикуемых допущений в тематическом моделировании. Разрабатывается модель интерпретируемых семантических векторных представления текста и обобщается на случай нетекстовых данных, таких как авторы, метки времени и др. В проведенных экспериментах демонстрируется качество, сопоставимое или превосходящее известные методы семейства word2vec, при этом дополнительно удастся добиться высокой разреженности и интерпретируемости компонент построенных векторов.

Основные результаты и новизна

В рамках диссертационной работы Потапенко А.А. получены следующие новые результаты:

1. Предложено обобщенное семейство ЕМ-алгоритмов, позволяющее обучать как известные, так и новые тематические модели, комбинируя эвристики робастности, сэмплирования тем и регуляризации Дирихле.
2. Предложена комбинация регуляризаторов, позволяющая увеличивать различность, интерпретируемость и разреженность предметных тем в тематической модели.
3. Предложен разреживающий регуляризатор отбора тем, позволяющий постепенно удалять из модели незначимые темы.
4. Разработан алгоритм построения семантических векторных представлений слов, обладающих свойством интерпретируемости компонент.
5. Разработан алгоритм построения единого векторного пространства для слов, документов, временных меток, категорий, авторов и других мета-данных документов.

Теоретическая и практическая значимость

В работе Потапенко А.А. систематизированы алгоритмы векторного представления текстов на основе гипотезы дистрибутивной семантики и на основе вероятностного тематического моделирования. Предложены новые методы, позволяющие совмещать преимущества известных методов из данных областей.

Существенно разработан подход аддитивной регуляризации вероятностных тематических моделей, являющийся альтернативой байесовским методам. В рамках данного подхода предложены новые регуляризаторы разреживания, сглаживания, декоррелирования, отбора тем; формализована гипотеза дистрибутивной семантики, получены векторные представления мета-данных документов.

В экспериментах на пяти русскоязычных и англоязычных корпусах различного объема и жанра (новости, научные статьи, статьи Википедии) продемонстрирована высокая практическая значимость полученных результатов. Результаты научной работы Потапенко А.А. легли в основу библиотеки тематического моделирования с открытым кодом BigARTM, которая была использована в проектах по анализу текстов различными научными группами и компаниями, в частности, для анализа этно-релевантных новостей в социальных сетях, обращений клиентов в службы поддержки в банках, построения карты научных областей по статьям конференций.

Достоверность результатов

Математические результаты диссертационной работы Потапенко А.А. сформулированы в виде утверждений и теорем с корректными доказательствами. На основе данных теорем предложены алгоритмы, которые были реализованы автором и применены к корпусам текстов на русском и английском языках. В экспериментах были продемонстрированы преимущества предложенных методов по сравнению с известными методами в области тематического моделирования и дистрибутивной семантики.

Результаты работы докладывались на конференциях «Диалог», ММРО, ECIR, AINL, AIST, SLDS, на семинарах конференций ACL и EMNLP, в исследовательских отделах компаний Google и Microsoft, в научных группах университетов ETH Zurich и TU Darmstadt. Таким образом, алгоритмы и модели, предложенные в диссертации Потапенко А.А., признаются сообществом в области автоматической обработки текстов, как в научной среде, так и в крупных компаниях.

Замечания

1. В работе уделяется мало внимания проблеме многозначности. Использование только одного векторного представления для многозначного слова не в полной мере отражает эту особенность естественных языков.
2. В тексте работы сложно отделить исследования автора от обзора релевантных работ, что затрудняет понимание личного вклада автора. Для этого требуется постоянная работа со списком литературы, отсылки к которому содержат только порядковые номера без дополнительных пояснений.
3. Рисунки в третьей главе перегружены, а обозначения описаны не полностью. Например, догадаться, что по оси ординат на всех рисунках отложены итерации EM-алгоритма, можно только из контекста при внимательном чтении основного текста. Это затрудняет восприятие работы. На рис. 3.7 левая ось ординат подписана как «итерации», что не соответствует действительности: должно быть «перплексия».
4. Некоторые утверждения нуждаются в пояснениях:
 - На странице 82 приводится утверждение, что при уменьшении коэффициента регуляризации «наступает момент, когда условие вырожденности темы перестает выполняться хотя бы для одного терма». Необходимо пояснить, почему так происходит.
 - На странице 92 делается вывод, «что комбинирование регуляризаторов позволяет улучшить все критерии качества при незначительном ухудшении перплексии», однако не приводится пояснений, какое изменение перплексии можно считать значительным.

Заключительная оценка

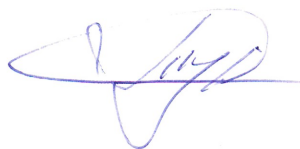
Диссертационная работа Потапенко Анны Александровны «Семантические векторные представления текста на основе вероятностного тематического моделирования», выполненная под руководством д.ф.-м.н., профессора Воронцова К.В., является законченной научно-квалификационной работой, содержащей новые научные результаты в области семантического анализа текстов. Все результаты, выносимые на защиту, строго обоснованы и подтверждены в ходе вычислительных экспериментов.

Результаты работы опубликованы в 12 печатных трудах, 7 из которых – в изданиях, рекомендованных ВАК. Автореферат диссертации, опубликованные статьи и тезисы докладов достаточно полно отражают основное содержание работы. Отмеченные недостатки работы не снижают общей положительной оценки.

Диссертация соответствует всем критериям, установленным Положением о присуждении ученых степеней, предъявляемым к диссертациям на соискание ученой степени кандидата физико-математических наук по специальности 05.13.17 – «Теоретические основы информатики», а ее автор – Потапенко Анна Александровна заслуживает присуждения ученой степени кандидата физико-математических наук по указанной специальности.

Настоящий отзыв обсуждался и был одобрен на заседании отдела информационных систем Федерального государственного бюджетного учреждения науки Института системного программирования им. В.П. Иванникова РАН 4.02.2019 г., протокол № 1.

Заведующий отделом
информационных систем ИСП РАН
к.ф.-м.н.



Д.Ю. Турдаков