УДК 004.89

АЛГОРИТМ ТЕМАТИЧЕСКОГО МОДЕЛИРОВАНИЯ С ИСПОЛЬЗОВАНИЕМ КВАНТОВОЙ ТЕОРИИ ВЕРОЯТНОСТИ

Авдюшина А. Е. (Национальный исследовательский университет ИТМО (Университет ИТМО)),
Научный руководитель – д.т.н., проф. ФПИиКТ Бессмертный И.А.
(Национальный исследовательский университет ИТМО (Университет ИТМО))

В данном исследовании разработан новый подход к задачам тематического моделирования с применением аппарата квантовой теории вероятности. Алгоритм учитывает вероятность появления слов в скрытых темах и вероятность принадлежности документа им, геометрические проекции векторов слов и документов на базисные вектора, которыми являются скрытые темы. Численными исследованиями показана эффективность и достоверность разработанного алгоритма, возможность его применения для определения тематики документов, выделения контекста.

Введение. Развитие средств информатизации способствует росту количества хранимой информации, в том числе и текстовой. Большие текстовые данные не структурированы, не относятся к одной тематике, что затрудняет информационный поиск релевантной информации и требует разработки новых алгоритмов. Анализ литературы показал, что основными подходами к автоматическому выявлению смыслов в текстовых коллекциях являются тематическое моделирование и векторное представление слов. Оба подхода применяются для информационного поиска, классификации, категоризации, аннотирования, сегментации текстов. Тематическое моделирование строится на гипотезе независимости от порядка слов в документе. Наибольшее распространение получили модели, учитывающие вероятность появления слова. К ним относятся латентное размещение Дирихле, вероятностный латентный семантический анализ. В последние годы интенсивно развивается векторизация текста или векторное представление слов для лучшего отображения семантической близости, что в последствии улучшает анализ текстовой информации. Модель векторизации с упрощенным вариантом нейронной сети Word2Vec позволяет получить матричное разложение слов в контекстах. Из сравнительного анализа подходов к обработке больших объемов тестовой информации, выделении их общих характеристик и существующих отличий следует вывод о необходимости учета как вероятности появления необходимых характеристик документов, так и расстояний между ними, задаваемых метриками. Люди чаще всего в речи и написании текста манипулируют языком в определенной области или контексте, что сложно моделировать в интеллектуальных системах. Поэтому получают развитие методы квантовой математики, которые способны описывать состояние системы, состоящей из множества частиц, и учитывать влияние малейших изменений в одной из частей на состояние всей системы в целом. Использование вероятности и расстояний учитывается в полной мере в квантовой математике, поэтому ее применение целесообразно для поиска контекста, тематического моделирования и информационного поиска.

Основная часть. В данной работе представлена разработанная последовательность обработки множества документов для выделения релевантных тем. Тематического моделирования проводится для корпуса (набора обработанных документов, готовых для дальнейшего исследования), документов (неупорядоченных множеств термов) и тем (множеств слов, объединённых одной тематикой). Темы характеризуются наборами термов, которыми могут быть слова или словосочетания. Для разработки алгоритма вероятностный латентный семантический анализ, построенный на дискретном распределении слов по темам и тем по

документам на множестве скрытых параметров – тем, дополнен квантовым формализмом. Описаны правила предобработки документов, построение численных характеристик, демонстрирующих связь между словами и документами, которые являются важными составляющими предварительной обработки данных и получения корпуса. Термы ранее полученные предобработкой с помощью «мешка слов» имеют квантовую запутанность. Количество слов в документе вычисляется с использованием терм-документной матрицы. При этом используется гипотеза условной независимости, по которой вероятность появления слова не зависит от документа, а зависит от темы. На основе предположения о запутанности слов в дальнейшем строится запутанность принадлежности документа или слова теме, их взаимному влиянию при малейшем изменении каждой составляющей всей системы. Скрытые темы используются в качестве базиса. Состояние системы описывается с использованием суперпозиции, неопределенности, скрытое семантическое пространство определяется подпространством гильбертова пространства. Для соблюдение квантового формализма по методу Грама-Шмидта строится первоначальный ортонормированный базис, поэтому же принципу каждый вектор терма представляется в виде линейной комбинации базисных векторов с комплексными коэффициентами, на которые накладывается уловие нормировки. Аналогичным образом строятся соотношения через базисные вектора для каждого документа. Тогда вероятность распределение термов в документе представляется в виде матрицы проекций и вероятности выбора темы для терма. Проекции вектора документа на базис тем также представляются в виде матрицы проекций. Таким образом, задача тематического моделирования с использованием квантового формализма сведена к максимизации правдоподобия. Для поиска тем использован итерационный EM- алгоритм. На E-шаге по найденным на предыдущем шаге скрытым параметрам определяется количество слов, порожденных темой в документе; на этапе M происходит переход к новому базису латентных тем с использованием матрицы поворота, пересчет оценок терм и документов на новом базисе тем с использованием матриц проекций. Итерационный процесс продолжается пока не будет решена задача максимального правдоподобия. Численная апробация алгоритма проводилась для 5, 10, 15, 20 тем, поскольку на таком объеме можно явно увидеть релевантность полученных результатов. Исследования начаты с получения датасета с текстами различной тематики. Подобраны свыше тысячи двухсот документов по таким направлениям как программная инженерия, биология, география и другие. Первым этапом обработки текста является его преодобработка: токенизация, удаление стоп-слов, лемматизация и составление мешка слов. В результате применения разработанного алгоритма получены скрытые темы, точность и полнота которых измерялась по метрике mAP в зависимости от задаваемого количества тем. Несмотря на неравномерность выборки документов разработанный метод показывает результат лучше базового метода вероятностного семантического анализа. Предложенный алгоритм эффективнее справляется со скрытыми факторами для оцениваемых коллекций.

Выводы. В ходе исследования приведена аналогия между законами квантовой теории и практическими задачами информационного поиска. Программная реализация этого алгоритма позволила провести численные эксперименты по определению скрытых тем и проверки достоверности результатов. Показана большая релевантность полученных тем относительно результатов других методов исследованных в работе.

Авдюшина А.Е. (автор) Подпись

Бессмертный И.А. (научный руководитель) Подпись