ОТЗЫВ

официального оппонента, доктора технических наук Краснова Федора Владимировича на диссертационную работу «Теоретико-графовые модели, методы и программные средства интеллектуального анализа текстовой информации на примере фольклорных и литературных произведений», представленную Москиным Николаем Дмитриевичем на соискание ученой степени доктора технических наук по специальности 05.13.18 – Математическое моделирование, численные методы и комплексы программ.

Актуальность избранной темы.

Исследовательская работа Н. Д. Москина направлена на разработку математических моделей и методов автоматизации выявления скрытых закономерностей в текстах. Текст представляет собой интересный объект для моделирования и дальнейшего анализа различными автоматизированными системами. Подобный анализ иногда называют Text Mining (интеллектуальный анализ текстов). При этом происходит преобразование текстовых данных в высококачественную информацию, на основе которой можно извлекать знания. Многочисленные публикации в этой области свидетельствуют об актуальности данного научного направления.

Отмечу также все возрастающую роль методов теории графов не только при автоматической обработке текстов, но и в других областях науки и техники. Упомянутый в работе семинар «Graph-Based Methods for Natural Language Processing» (TextGraphs), который последние годы проводится «Association for Computational Linguistics» (ACL), подтверждает актуальность выбранного автором научного направления. При этом теоретико-графовые модели можно использовать в тех теоретических и практических задачах, которые на сегодняшний день не нашли своего окончательного решения. Особенно это касается случаев, где прослеживаются определенные ограничения, связанные с формированием изучаемых текстовых выборок. Таких текстов много как среди литературных, так и фольклорных (народных) произведений. Поэтому результаты диссертационного исследования Н. Д. Москина будут, без сомнения, интересны как представителям технических наук, так и гуманитарных.

Структура и содержание диссертации.

Оформление и содержание диссертации отвечает общим требованиям Положения о порядке присуждения ученых степеней ВАК РФ. Работа объемом 370 страниц состоит из введения, семи глав, заключения, списка использованных источников и четырех приложений, содержит 91 рисунок и 101 таблицу. Список литературы состоит из 364 источников. В конце каждой главы сформулированы соответствующие выводы. Работа выполнена с учетом требований к оформлению диссертаций, изложенными в ГОСТ Р 7.0.11-2011 «Диссертация и автореферат диссертации. Структура и правила оформления».

Во введении обоснована актуальность темы диссертационной работы, выделен объект исследования, описана научная новизна, сформулированы цель, задачи и научные положения, выносимые на защиту.

В первой главе под названием «Обзор проблем и решений в области анализа текстов» представлен подход к исследованию текста как многоуровневой системе взаимосвязанных языковых элементов. Описаны задачи, привлекающие внимание специалистов, и методы их решения в области реферирования текста, машинного перевода, анализа тональности, генерации текстов, информационного поиска и др. Представляя задачу атрибуции литературных и фольклорных текстов, автор также описывает различные значимые лингвостатистические параметры.

Во второй главе «Теоретико-графовые модели фольклорных и литературных текстов» приводится обзор теоретико-графовых моделей языковой структуры текстов. Рассмотрев понятия нечеткого графа, темпорального графа и иерархического графа, автор представляет обобщенную модель, с помощью которой описываются теоретико-графовая модель синтаксической структуры поэтических и фольклорных текстов, теоретико-графовая модель составляющих, темпоральная теоретико-графовая модель сказочного сюжета, иерархическая теоретико-графовая модель синтаксической структуры предложения, нечеткая теоретико-графовая модель «сильных связей» грамматических классов, нечеткая теоретико-графовая модель зависимостей, теоретико-графовая модель сюжета фольклорных песен. На примерах показано, как происходит формирование вершин и ребер на основе текстовых фрагментов, какие значения принимает функция принадлежности и пр.

В третьей главе «Числовые характеристики текстов и их теоретико-графовых моделей» рассматриваются понятия лексического спектра и n-граммы для задачи

атрибуции литературных текстов. Лексический спектр аппроксимируется различными видами кривых, на основе которых формируются коэффициенты для классификации текстов. Выполнено построение ансамбля классификаторов на основе деревьев решений, в том числе при несбалансированности выборки. Представлена математическая модель поиска неоднородных фрагментов в текстах. После обзора ряда значимых числовых характеристик теоретико-графовых моделей проверяются гипотезы о влиянии жанра песен на некоторые из них (количество вершин и ребер, параметр связности, максимальная степень вершины). Описана аппроксимация распределений степеней вершин графов бесёдных песен с помощью гиперболической кривой. Построены классификаторы для разграничения фольклорных текстов и стилизованных текстов с помощью пяти методов интеллектуального анализа данных. Как деревья решений, так и дискриминантный анализ, метод опорных векторов, нейронная сеть, случайный лес показали среднюю точность распознавания около 80%.

В четвертой главе «Методы агрегации теоретико-графовых моделей» показано, как с помощью математических моделей можно количественно измерить значимость вершин, ребер и подграфов. Методика проиллюстрирована примерами для графов фольклорных песен. При этом излагается как подход по подсчету числовых характеристик с учетом иерархичности, нечеткости и темпоральности соответствующих графов, так и подход с использованием агрегированных моделей. Также в главе описывается алгоритм поиска мотивов на основе нечетких теоретико-графовых моделей.

Пятая глава «Метрики на множестве графов» посвящена исследованию расстояний на множестве графов. Помимо обзора известных метрик она содержит также описание нового расстояния, основанного на максимальном общем подграфе, где учитывается упорядоченность вершин. Приводится доказательство теоремы, что данное расстояние удовлетворяет всем свойствам метрики. Дан сравнительный анализ поэтических текстов П. А. Вяземского и И. А. Бродского с помощью подграфовых расстояний.

В шестой главе «Метрики для сравнения деревьев решений» более углубленно описываются расстояния на множестве деревьев. При этом, с одной стороны, дерево рассматривается как частный случай графа. С другой стороны, специфика деревьев позволяет построить более эффективные методы их анализа. Автор исследует

закономерности между параметрами для задания выборки текстов (размер фрагмента и шаг скользящего окна) и деревьями решений (на примере произведений Ф. М. Достоевского). Разделяя подобные деревья на обучающие и тестирующие, в работе предлагается построение классификаторов, основанных на подграфовых расстояниях.

В седьмой главе «Программная поддержка исследований текстовых коллекций» рассмотрены основные возможности и структура разработанной информационной системы «Фольклор», позволяющей исследовать тексты с теоретико-графовой формализацией. Завершает главу история разработки и основная структура информационной системы СМАЛТ («Статистические методы анализа литературных текстов»). Приводится также описание модулей данных программных систем.

В заключении сформулированы основные результаты работы.

Научная новизна работы заключается в том, что автором предложена совокупность теоретико-графовых моделей для анализа литературных и фольклорных текстов, которые обладают свойствами нечеткости, иерархичности и темпоральности (1 и 2 главы). Также предложено рекурсивное определение обобщенной контекстно-зависимой теоретико-графовой модели для решения задачи атрибуции (2 глава). На основе подобных моделей автор предлагает набор числовых характеристик, которые можно в некоторых случаях аппроксимировать гиперболическими кривыми, что подтверждается вычислительными экспериментами с использованием расстояния хи-квадрат (3 глава). Предложены три модификации методов агрегации с использованием факторов нечеткости, иерархичности и темпоральности с заданным «простым графом» (4 глава). Предложено расстояние для сравнения теоретико-графовых моделей, которое основано на максимальном общем подграфе и учитывает упорядоченность вершин (5 глава). При этом убедительно доказаны все свойства метрики, а именно тождественность, неотрицательность, нечеткость и неравенство треугольника. Для повышения качества результатов классификации текстов предложены методы сравнения деревьев решений, которые базируются на использовании подграфовых расстояний (6 глава). Разработан комплекс, состоящий из двух взаимосвязанных программ «Фольклор» и СМАЛТ, где были реализованы алгоритмы анализа текстов, изложенные в диссертации (7 глава).

Достоверность и степень обоснованности научных положений, выводов и рекомендаций.

Обоснованность и достоверность научных положений, выводов и рекомендаций диссертационного исследования не вызывают сомнения, что подтверждается использованием современных методов численного моделирования и вычислительных алгоритмов.

Научные положения и теоретические выводы получены путем корректного применения математического аппарата теории графов, теории вероятностей и прикладной статистики, интеллектуального анализа данных, математической лингвистики и др. Основные результаты опубликованы в 73 научных работах, из которых 14 в научных журналах, рекомендованных ВАК РФ, и в изданиях, индексируемых в базе данных Scopus и Web of Science, 5 свидетельствах на регистрацию программы для ЭВМ и БД. Среди публикаций можно отметить две монографии: «Теоретико-графовые модели фольклорных текстов и методы их анализа» (Петрозаводск, 2013) и «Проблема атрибуции в журналах «Время», «Эпоха» и еженедельнике «Гражданин»» (Петрозаводск, 2021).

Замечания по диссертационной работе.

1. Диссертация содержит достаточно много математических терминов, однако в ней отсутствует предметный указатель, что несколько усложняет чтение и понимание (особенно для тех, кто не является специалистами по теории графов).

2. Отсутствуют необходимые пояснения для ряда обозначений, встречающихся в диссертационной работе. Например, в начале четвертой главы при определении графа горизонтальной видимости непонятно, что скрывается за изображенными на рисунке обозначениями B^{n}_{3}, E^{n+1}_{7}, A^{n+2}_{1} и т. п.

3. Некоторые алгоритмы (например, алгоритм поиска расстояния между деревьями с использованием операций редактирования, *tree edit distance*) недостаточно полно изложены в диссертации. Полагаю, что алгоритмы можно было бы представить, например, в псевдокоде, а также более четко прописать используемые структуры данных.

4. В седьмой главе дано описание программных систем «Фольклор» и СМАЛТ, однако отсутствуют требования к аппаратному и лингвистическому программному обеспечению.

Приведенные замечания не снижают теоретического и практического значения работы и не ставят под сомнение её положительную оценку.

Заключение о соответствии диссертации критериям Положения о присуждении ученых степеней ВАК РФ.

Анализ диссертационной работы, автореферата и опубликованных работ показывает, что диссертация Москина Николая Дмитриевича «Теоретико-графовые модели, методы и программные средства интеллектуального анализа текстовой информации на примере фольклорных и литературных произведений» соответствует паспорту специальности 05.13.18 – «Математическое моделирование, численные методы и комплексы программ», является законченной научно-квалификационной работой, в которой на основании выполненных автором исследований содержится научно обоснованное решение задач автоматической обработки текстов, что соответствует п. 9 Положения о присуждении ученых степеней.

Считаю, что диссертационная работа удовлетворяет требованиям Положения о присуждении ученых степеней, предъявляемым к докторским диссертациям, а ее автор, Москин Николай Дмитриевич, заслуживает присуждения ученой степени доктора технических наук по специальности 05.13.18 – Математическое моделирование, численные методы и комплексы программ.

Аналитик больших данных в отделе разработки ООО «Ваше Цифровое Издательство», доктор технических наук

Краснов Федор Владимирович

30 августа 2022 г.

Адрес: 109263, г. Москва,
ул. Шкулёва, д. 9, к. 2
Телефон: +7 (499) 754-99-93
E-mail: fkrasnov2@yandex.ru