

ОТЗЫВ

научного руководителя кандидата технических наук, доцента

Филипповича Юрия Николаевича

на диссертационную работу

Воробьева Никиты Григорьевича

«МЕТОДЫ И АЛГОРИТМЫ РУБРИКАЦИИ ДОКУМЕНТОВ НА ОСНОВЕ
ВЕКТОРНО-ГРАФОВОЙ МОДЕЛИ»,

представленную на соискание ученой степени кандидата технических наук
по специальности 2.3.8. — Информатика и информационные процессы

Воробьев Никита Григорьевич обучался в Федеральном государственном (в н.в. автономном) бюджетном образовательном учреждении высшего образования «Московский политехнический университет» (Московский Политех). В 2019 г. окончил бакалавриат по направлению подготовки 09.03.02 -- Информационные системы и технологии, в 2021 г. окончил магистратуру по направлению 09.04.01 -- Информатика и вычислительная техника (образовательная программа «Компьютерная лингвистика и ИИ»), в 2025 г. окончил аспирантуру по направлению 09.06.01 -- Информатика и вычислительная техника (направленность «Теоретические основы информатики»). По направлениям подготовки 09.04.01 и 09.06.01 имеет дипломы с отличием.

Воробьев Никита Григорьевич с 2021 года по н.в. работает преподавателем на кафедре «Инфокогнитивных технологий», проводит лабораторные и практические занятия, ведет курсовое проектирование студентов по дисциплинам «Основы мобильной разработки», «Информационный поиск и индексирование», «Обработка больших данных» и др. Выполняет методическую работу, является соавтором двух электронных образовательных ресурсов по дисциплинам «Введение в психолингвистику» (2023 г.) и «Системы общения на естественном языке» (2024 г.). Пользуется уважением и авторитетом у преподавателей и студентов.

Актуальность диссертации. Актуальность диссертационного исследования обусловлена быстрым ростом объемов неструктурированных текстовых данных и

необходимостью повышения эффективности их автоматической обработки, включая задачи рубрикации документов, извлечения смысловых связей и поддержки диалоговых систем. Современные векторные языковые модели, обладая высокой вычислительной эффективностью, не обеспечивают явного представления семантической структуры языка, тогда как тезаурусные подходы отличаются высокой точностью и интерпретируемостью, но требуют значительных экспертных затрат и плохо масштабируются. Возникающее противоречие между масштабируемостью и структурной полнотой существующих методов обуславливает необходимость разработки гибридной графовой языковой модели, способной автоматически формировать структурированные контекстные связи и сочетать преимущества графовых и векторных представлений для повышения качества семантического анализа текстов.

Цель диссертации — Целью диссертационной работы является разработка методов и алгоритмов построения и применения графовой языковой модели для повышения точности рубрикации документов, а также интеграции этих моделей с векторными представлениями для использования в современных интеллектуальных системах.

Поставленная цель достигается автором путем последовательного решения следующих научных и практических задач:

1. Проведение анализа существующих подходов к построению языковых моделей (экспертные тезаурусы, Word2Vec, BERT) и выявление их ограничений;
2. На основе сравнения экспертных и автоматически построенных тезаурусов разработка метрики для оценки эффективности тезаурусов;
3. Разработка метода построения графовой языковой модели на основе графовых грамматик;
4. Оценка эффективности графовой модели в задачах рубрикации документов по сравнению с Word2Vec и BERT;
5. Разработка метода интеграции графовых моделей с векторными представлениями, где вершинами графа выступают векторные модели документов;
6. Разработка алгоритма приведения графовой модели к векторному представлению для применения в интеллектуальных системах;
7. Проведение вычислительных экспериментов и оценка практической применимости предложенных методов и алгоритмов.

Научную новизну имеют следующие результаты, полученные лично автором и представленные в диссертационной работе:

1. Разработана интегральная метрика оценки эффективности тезаурусных структур, основанная на совокупности топологических характеристик графа (мощность, средняя степень вершины, плотность, коэффициент кластеризации, энтропия), которая позволяет количественно сравнивать экспертные и автоматически сформированные тезаурусы и обеспечивает дифференциацию структур по нормализованному вектору признаков;

2. Предложен алгоритм автономной оценки качества тезауруса на основе интегральной метрики и весовой матрицы, обеспечивающий выявление некорректных структур при предельных значениях плотности ($D \rightarrow 0$, $D \rightarrow 1$);

3. Разработан метод построения графовой языковой модели (GLM) на основе графовых грамматик и локальных контекстов, обеспечивающий явное представление семантических связей при обработке текстовых корпусов объёмом до 1000 документов;

4. Предложен метод насыщения графовой языковой модели дополнительными тезаурусными рёбрами, повышающий структурную связность модели и улучшающий её топологические характеристики (в частности, коэффициент кластеризации и энтропию связей) по сравнению с базовой графовой структурой;

5. Разработан гибридный подход к моделированию контекста, в котором вершинами графа выступают векторные представления документов размерности 300 признаков, что обеспечивает сопоставимость с существующими векторными моделями и позволяет обрабатывать текстовые документы объёмом 3000–5000 символов (400–800 токенов) без снижения точности;

6. Предложен алгоритм приведения графовой модели к векторной форме на основе модифицированного алгоритма Вайсфейлера–Лемана с пороговой фильтрацией, позволяющий интегрировать графовые структуры в нейросетевые архитектуры и обеспечивающий сохранение не менее 90% структурной информации при преобразовании;

7. Экспериментально установлено преимущество разработанной графовой языковой модели в задачах рубрикации документов по сравнению с векторными моделями (Word2Vec, BERT), выражающееся в повышении точности рубрикации (accuracy до 0.94 против 0.78, прирост до 16%), увеличении значения F1 (macro) до 0.97 и устойчивом улучшении качества рубрикации на корпусе из 1000 документов. Обобщённый прирост точности составляет до 26%.

Практическая ценность разработок диссертационной работы состоит в конкретном их использовании в учебном процессе. Также имеется возможность более широкого применения разработанных в рамках диссертации и зарегистрированных в реестрах Федерального института промышленной собственности (ФИПС) 4-х программ для ЭВМ: «Генератор моделей медицинских текстов», «Сравнение контекстной близости языковых медицинских моделей», «Конструктор контекстных графов», «Конвертер графовых языковых моделей».

В диссертации отсутствуют случаи заимствования разработок, принадлежащих другим авторам. Приведены сведения об источниках использованных в исследовательских целях материалов.

Результаты проведенных и представленных в диссертации исследований достоверны и полно представлены в 7 печатных публикациях, в том числе 2-х в журналах из списка ВАК, а также апробированы на 9-ти научно-технических конференциях.

Диссертация Воробьева Никиты Григорьевича является научной квалификационной работой, в которой содержится решение задачи, имеющей существенное значение для разработки автоматизированных информационных технологий обработки текстовой информации, отвечает требованиям, предъявляемым к диссертациям на соискание ученой степени кандидата технических наук. Содержание диссертации соответствует специальности 2.3.8. — Информатика и информационные процессы (технические науки), а ее автор заслуживает присуждения ученой степени кандидата технических наук по данной специальности.

Кандидат технических наук, доцент, профессор кафедры «Инфокогнитивные технологии» Федерального государственного автономного образовательного учреждения высшего образования «Московский политехнический университет»
Адрес: 107023, Г.МОСКВА, УЛ. БОЛЬШАЯ СЕМЁНОВСКАЯ, Д. 38, Телефон: +7 (495) 223-05-23, E-mail: mospolytech@mospolytech.ru
Филиппович Юрий Николаевич

Дата: «11» марта 2026 г.

ПОДПИСЬ Филипповича Ю.Н. заверяю

СПЕЦИАЛИСТ ПО
КАДРОВОМУ
ДЕЛОПРОИЗВОДСТВУ
ПОГОРЕЛОВА А.В.

