На правах рукописи

ВОРОБЬЕВ НИКИТА ГРИГОРЬЕВИЧ

МЕТОДЫ И АЛГОРИТМЫ РУБРИКАЦИИ ДОКУМЕНТОВ НА ОСНОВЕ ВЕКТОРНО-ГРАФОВОЙ МОДЕЛИ

Специальность 2.3.8. – Информатика и информационные процессы

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

МОСКВА 2026

Работа выполнена в Федеральном государственном автономном образовательном учреждении высшего образования «Московский политехнический университет» на кафедре «Инфокогнитивные технологии».

Научный руководитель:	**Филиппович Юрий Николаевич**, кандидат технических наук, доцент, профессор кафедры «Инфокогнитивные технологии» Федерального государственного автономного образовательного учреждения высшего образования «Московский политехнический университет», г. Москва.
Официальные оппоненты:	**Конявский Валерий Аркадьевич** доктор технических наук, Академик РАЕН, академик АЭН РФ, заведующий кафедрой «Защита информации» Федерального государственного автономного образовательного учреждения высшего образования «Московский физико-технический институт (национальный исследовательский университет)», г. Долгопрудный.
	Кузнецов Владислав Владимирович кандидат технических наук, ведущий программист Федерального государственного учреждения «Федеральный исследовательский центр «Информатика и управление» Российской академии наук», г. Москва.
Ведущая организация:	Федеральное государственное автономное образовательное учреждение высшего образования "Московский государственный технический университет имени Н.Э. Баумана (Национальный исследовательский университет)", г. Москва.

Защита диссертации состоится «29» июня 2026 года в 13:30 часов, на заседании диссертационного совета 99.2.113.02 в ФГБОУ ВО «Рязанский государственный радиотехнический университет им. В.Ф. Уткина»., 390005, г. Рязань, ул. Гагарина, 59/1.

С диссертацией можно ознакомиться в научной библиотеке ФГБОУ ВО «Рязанский государственный радиотехнический университет им. В.Ф. Уткина», на сайте http://rsreu.ru/.

Автореферат разослан "____" _____ 2026 г.

Ученый секретарь
диссертационного совета
доктор технических наук, доцент А. Н. Колесенков

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы

Актуальность диссертационного исследования обусловлена быстрым ростом объёмов 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 обусловливает 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.

Степень проработанности темы

Степень проработанности темы исследования отражена в работах исследователей, посвящённых задачам рубрикации текстов, построению тезаурусов и развитию языковых систем. В трудах Н. В. Лукашевич отмечается, что интеллектуальные методы рубрикации обладают ограниченной применимостью при обработке больших массивов данных, тогда как информационно-поисковые тезаурусы обеспечивают высокую точность, но требуют значительных экспертных затрат. Проблемы масштабирования и сопровождения тезаурусов в условиях постоянно обновляющихся текстовых корпусов рассматриваются в работах С. А. Осокиной. В исследованиях Ю. Н. Караулова подчёркивается важность сохранения семантических и структурных связей между словами как основы адекватного представления языка. Работы М. С. Агеева и Б. В. Доброва посвящены анализу роста объёмов текстовых данных и необходимости автоматизации рубрикации документов; при этом указываются ограничения классических методов поиска и

распределённых языковых моделей. В целом, несмотря на наличие значительного числа исследований, задача автоматического построения масштабируемых и интерпретируемых семантических моделей, учитывающих структурную организацию языка, остаётся недостаточно решённой.

Цель исследования

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.

Задачи исследования

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.

Научная новизна

В работе получены следующие новые научные результаты:

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 ($D \to 0$, $D \to 1$);
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%.

Объект исследования

Процессы автоматической обработки и анализа естественно-языковых текстов в задачах рубрикации документов и диалоговых систем.

Предмет исследования

Методы и алгоритмы построения графовых языковых моделей, метрики оценки тезаурусных структур и методы рубрикации текстовых документов на основе графовых и гибридных векторно-графовых представлений.

Теоретическая значимость исследования

Теоретическая значимость диссертационной работы заключается в развитии графовых методов моделирования языка и формировании гибридного подхода к построению языковых моделей. Предложена формальная схема интеграции графовых и векторных представлений, позволяющая описывать лексико-семантические структуры с учётом как их топологии, так и вероятностных характеристик. Разработанная метрика эффективности обеспечивает возможность сравнения построенных тезаурусов. Полученные результаты вносят вклад в развитие теории графовых грамматик и компьютерной лингвистики, создавая основу для дальнейших исследований в области интерпретируемых и гибридных языковых представлений.

Практическая значимость исследования

Практическая значимость работы заключается в том, что результаты работы могут быть применены при разработке интеллектуальных систем анализа текстов, рекомендательных систем и поисковых сервисов. Разработанные методы обеспечивают повышение точности рубрикации документов и выбора ответов в

диалогах, а также позволяют интегрировать графовые модели в существующие интеллектуальные системы. Полученные решения могут быть использованы при создании обучающих программных комплексов и исследовательских систем в области компьютерной лингвистики и искусственного интеллекта.

Методология и методы исследования

Методология исследования основана на сочетании методов обработки естественного языка, теории графов и машинного обучения. В работе использованы методы анализа текстовых корпусов, статистические и распределённые модели представления слов, а также подходы к построению и анализу графовых структур. Для моделирования семантических связей между терминами применялись формальные метрики связности, когерентности и энтропии, позволяющие количественно оценивать структуру формируемых языковых моделей.

В работе использовались следующие методы теории графов и графовых грамматик для построения и анализа структур языковых моделей. Применялись методы обработки естественного языка и компьютерной лингвистики для формирования текстовых корпусов и построения моделей. Использовались алгоритмы машинного обучения и нейросетевых технологий (Word2Vec, BERT) для сравнения эффективности различных языковых представлений. Проведены вычислительные эксперименты и эмпирические сравнения для проверки точности рубрикации документов и выбора ответов в диалоговых системах.

Положения, выносимые на защиту

1. Метод построения информационно-поискового тезауруса предметной области на основе статистики локальных контекстов и обратного индекса, обеспечивающий формирование связной семантической структуры, пригодной для последующего графового моделирования текстов;
2. Алгоритм автономной оценки эффективности тезаурусных структур на основе интегральной метрики, учитывающей топологические характеристики графа (мощность, средняя степень вершины, плотность, коэффициент кластеризации, энтропия), позволяющий выявлять некорректные структуры при предельных

значениях плотности (D→0, D→1) и использовать оценку в автоматическом режиме;

3. Метод формирования графовой языковой модели (GLM) с использованием графовых грамматик и локальных контекстов, обеспечивающий явное представление семантических связей (гипонимо-гиперонимических и ассоциативных) между лексическими единицами;
4. Метод насыщения графовой языковой модели тезаурусными рёбрами, обеспечивающий повышение структурной связности модели и улучшение её топологических характеристик (коэффициента кластеризации и энтропии связей);
5. Гибридная модель представления текстов, в которой вершинами графа выступают векторные представления документов размерности 300 признаков, обеспечивающая совместное использование структурных и вероятностных характеристик текста и применимая для обработки документов объёмом 3000–5000 символов (400–800 токенов);
6. Алгоритм приведения графовой языковой модели к векторному представлению на основе модифицированного алгоритма Вайсфейлера–Лемана с пороговой фильтрацией, обеспечивающий интеграцию графовых структур в нейросетевые методы обработки данных при сохранении не менее 90% структурной информации;
7. Алгоритм рубрикации текстовых документов на основе графовой языковой модели, использующий интегральную оценку структурной близости и обеспечивающий повышение точности рубрикации до 0.94 (прирост до 16% по сравнению с Word2Vec), а также улучшение значения F1 (macro) до 0.97 на корпусе из 1000 документов.

Полученные результаты соответствуют следующим пунктам паспорта специальности 2.3.8

1. Результаты 1, 3, 4 соответствуют п.4 «Разработка методов и технологий цифровой обработки аудиовизуальной информации с целью обнаружения закономерностей в данных, включая обработку текстовых и иных

изображений, видеоконтента. Разработка методов и моделей распознавания, понимания и синтеза речи, принципов и методов извлечения требуемой информации из текстов»;

2. Результаты 2, 5, 6, 7 соответствуют п. 1 «Разработка компьютерных методов и моделей описания, оценки и оптимизации информационных процессов и ресурсов, а также средств анализа и выявления закономерностей на основе обмена информацией пользователями и возможностей используемого программно-аппаратного обеспечения».

Степень достоверности и апробация результатов

Достоверность результатов обеспечивается корректным применением математического аппарата теории графов, формальных грамматик и методов машинного обучения; представительным объемом проанализированных текстовых корпусов; а также согласованностью результатов вычислительных экспериментов с данными, полученными при использовании классических векторных моделей (Word2Vec, BERT). Теоретические выводы подтверждаются успешным внедрением разработанных алгоритмов в практические программные комплексы.

Публикации. Результаты исследований и разработок опубликованы в 7 печатных статьях, в том числе в 2 журналах из списка ВАК. Статьи зарегистрированы в системе научного цитирования РИНЦ.

Апробация. Результаты диссертационного исследования были представлены на 9 научных конференциях, в том числе: «ИИАСУ 2024» (МГТУ им. Баумана); «Язык. Сознание. Коммуникация: методология и гуманитарные практики (вызовы современности)»; «VI Международный научный Форум профессорско-преподавательского состава и молодых ученых «Цифровые технологии: наука, образование, инновации» ФГБОУ ВО «МГТУ «СТАНКИН»»; «Международный научно-технический форум «Современные технологии в науке и образовании» (СТНО)»; «HETS 2022: II Ежегодный международный научно-практический форум по проблемам устойчивого развития в цифровом мире: Человек. Экономика. Технологии. Социум».

Внедрение. Результаты разработок подтверждены 4 свидетельствами на программу для ЭВМ в Федеральном институте промышленной собственности (ФИПС): 2025686741, 2025685835, 2025685987, 2025689874.

В учебный процесс в Московском политехническом университете по образовательным программам «Веб-технологии» (подготовка бакалавров по направлению 09.03.01) и «Интеллектуальные системы» (подготовка магистров по направлению 09.04.01) были внедрены следующие разработанные методики:

1. представления локальных контекстов в виде взвешенной графовой естественно-языковой модели предметной области. Реализация методики предусматривает учет различных типов отношений между терминами, формирование структурированной модели текста и повышение устойчивости представления при работе с семантически насыщенными документами. Реализация выполнена в виде программного модуля на языке python;
2. преобразования графового представления текста в векторное пространство без потери структурной информации, обеспечивающей учет локального и глобального контекста терминов. Метод основан на итеративной процедуре уточнения меток вершин графа и позволяет формировать признаковое пространство, пригодное для задач классификации и кластеризации текстов. Реализация выполнена в виде программного модуля на языке python;
3. проведения экспериментальных исследований, ориентированная на оценку эффективности различных естественно-языковых моделей в задаче обработки текстовых данных. Методика включает формирование обучающих и тестовых выборок, расчет стандартных метрик качества классификации (accuracy, precision, recall, F1), а также сравнительный анализ графовых и распределительных моделей с учетом структурных характеристик текстов.

Подготовлены 2 электронных образовательных ресурса: договор ЭОР/0235-ВМ/08-04/2024 и договор ЭОР/0332-ВМ/08-04/2025.

Метод формализации естественно-языковых текстовых данных в виде графовых моделей с возможностью насыщения различными типами семантических связей; алгоритм приведения графовой языковой модели к векторному представлению на основе алгоритма Вайсфейлера-Лемана, позволяющий сохранить структурные связи с множеством ребер в векторном формате, были внедрены автором диссертации при выполнении научного проекта на тему «Разработка методов сбора и оценки контекстных данных на основе машинного обучения» в рамках гранта им. В.Е. Фортова..

Личный вклад соискателя. Все основные результаты, выносимые на защиту, получены автором лично. В работах, опубликованных в соавторстве, соискателю принадлежит: разработка метода автоматического сбора локальных контекстов в режиме реального времени [Н. Г. Воробьев, Ю. Н. Филиппович // Язык. Сознание. Коммуникация: методология и гуманитарные практики (вызовы современности): Материалы II Всероссийской конференции с международным участием, Москва, 08–11 апреля 2024 года – Москва: ИД "Канцлер", 2024. – с. 131-133], разработка алгоритмов построения и оценивания информационно-поисковых тезаурусов на основе метрик полноты, связанности и энтропии [Н. Г. Воробьев, Ю. Н. Филиппович, Е. А. Пшехотская // Искусственный интеллект в автоматизированных системах управления и обработки данных: Сборник статей II Всероссийской научной конференции: в 5 томах, Москва, 27–28 апреля 2023 года. — Москва: КДУ, Добросвет, 2023. — с. 258-263], разработка метода создания тезауруса предметной области на основе обратного индекса статистических данных вхождения слов в локальный контекст [Н. Г. Воробьев, Ю. Н. Филиппович, Е. А. Пшехотская, А. Ю. Филиппович // Third International Conference on Optics, Computer Applications, and Materials Science (CMSD-III 2023). — Dushanbe, 20–22 декабря 2023 года. Vol. 13065. – Washington: SPIE-SOC PHOTO-OPTICAL INSTRUMENTATION ENGINEERS, 2024. — P. 1306514. – DOI 10.1117/12.3025183], а также проведение вычислительных экспериментов, анализ и интерпретация полученных результатов. Постановка задач исследования и обсуждение результатов проводились совместно с научным руководителем.

Структура и объем диссертации. Диссертация состоит из введения, четырех глав, заключения, списка литературы, включающего 93 источника, и 3 приложений. Общий объем работы составляет 178 страниц машинописного текста, включая 27 рисунков, 6 диаграмм и 9 таблиц.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность темы, формулируется цель работы, состав решаемых задач, приводится перечень основных результатов, и излагается краткое содержание глав диссертации.

В главе 1 рассматриваются основные подходы к автоматической обработке естественного языка, включая задачи рубрикации документов. Выполнен анализ существующих языковых моделей и тезаурусных систем, как экспертных, так и автоматически формируемых, с выделением их сильных сторон и ограничений. Рассматриваются векторные модели представления текста, в том числе Word2Vec и BERT, и разбираются их преимущества и ограничения. По результатам обзора обосновывается целесообразность разработки графовых языковых моделей, основанных на графовых грамматиках и способных более полно учитывать структуру и смысловые отношения в языке. В главе рассматриваются:

1. Современные задачи автоматической обработки текстов: рубрикация документов и диалоговые системы;
2. Сравнительный анализ экспертных и автоматически построенных тезаурусов;
3. Обзор векторных языковых моделей (Word2Vec, BERT) и их ограничений;
4. Обоснование выбора графовых грамматик как основы для построения языковых моделей.

В работе рассмотрены три подхода к моделированию предметной области. Первый подход основан на использовании отдельных векторных моделей для каждого подкласса и эффективен при обработке небольших текстов с выраженным контекстом. Второй подход предполагает построение единой векторной модели на всём корпусе данных и обучение нейронной сети для распознавания подклассов, что позволяет учитывать сложные контекстные зависимости в больших текстах. Третий

подход использует графовую модель, в которой слова представлены вершинами, а их связанность — взвешенными рёбрами; для рубрикации пользовательский запрос преобразуется в граф и сопоставляется с графами подклассов по мере их структурной близости (1).

$$d(G,F) = 1 - min_{i=1,\dots,k}(\frac{|mcs(g_{\min(i,m)}, f_i|}{i}), \quad (1)$$

где d(G,F) – нормализованная величина близости графов; $|mcs(g_{\min(i,m)}, f_i|$ – мощность максимального общего подграфа контекстуальных графов; $min_{i=1,\dots,k}$ - минимальное число общих вершин; g и f – сравниваемые подграфы; i – число вершин в подграфе; k – число подграфов.

В таком случае формула графа определяется как: G=(N,C,W), где N – вершины графа (языковые единицы, составляющие грамматику); C – дуги, объединяющие вершины; W – веса дуг (вероятность перехода между вершинами). Предлагается метод автоматического создания информационно-поискового тезауруса:

1. Выделение ключевых терминов;
2. Выделение гипонимо-гиперонимических связей вхождением строки-суффикса в термин;
3. Выделение потенциальных синонимических связей;
4. Выделение ассоциативных связей с помощью контекстного графа, векторного представления или гибридного метода;
5. Нормализация представления.

Делаются следующие выводы: проведенные исследования показали, что использование нескольких векторных моделей обеспечивает приемлемую точность при обработке коротких запросов и может применяться для предварительной рубрикации текстов. Применение одной векторной модели в сочетании с нейронной сетью оказалось более эффективным и позволило корректно обрабатывать длинные и сложные тексты, что делает данный подход практичным для реальных задач. Графовая модель обеспечила наивысшую точность при анализе коротких запросов, однако её высокая вычислительная сложность ограничивает применение в прикладных системах. Установлено, что прямое обучение Word2Vec на текстах

предметной области является наиболее простым способом получения базовой функциональности, тогда как использование размеченных датасетов повышает качество рубрикации. Показано, что комбинирование векторных и графовых моделей создаёт основу для более структурированного и устойчивого семантического анализа.

В главе 2 «Методы оценки и экспериментальное сравнение тезаурусов» разрабатываются и обосновываются методы количественной оценки качества тезаурусов на основе метрик мощности и связности. Предложенные характеристики позволяют сравнивать экспертные и автоматически построенные тезаурусы с точки зрения структуры семантических связей, а не только объёма словаря. Результаты вычислительных экспериментов показывают, что экспертные тезаурусы, несмотря на меньший размер, обладают более устойчивой и связной структурой. Полученные выводы используются как обоснование необходимости разработки новых языковых моделей, сочетающих структурную точность экспертных подходов и масштабируемость автоматических методов. В главе рассматривается:

1. Разработка метрики и формул для измерения эффективности тезаурусов;
2. Экспериментальное сравнение экспертных и автоматически сгенерированных тезаурусов;
3. Выводы о преимуществах экспертных структур за счёт связности и качества семантических отношений.

Для сравнения были выбраны несколько систем сбора тезаурусов. Thesaurus.com, ABC Thesaurus, RuWordNet и тезаурус, который был составлен с использованием экспертного подхода выпускниками медицинских вузов.

Слово на английском	Синоним 1	Синоним 2	Синоним 3	Синоним 4	Синоним 5	Синоним 6	Синоним 7	Синоним 8	Синоним 9	Синоним 10	Синоним 11	Синоним 12	Синоним 13	Синоним 14	Синоним 15	Синоним 16	Синоним 17
disease	Cancer	bug	condition	contaminatic	defect	disorder	epidemic	fever	flu	illness	infection	inflammatio	malady	plague	sickness	stroke	syndrome
medicine	antibiotic	cure	drug	medication	pharmaceuti	pill	prescription	remedy	anesthetic	antidote	antiseptic	antitoxin	balm	capsule	dose	elixir	injection
molecular	atomic	infinitesimal	little	minute	subatomic												
biology	analysis	cytology	diagnosis	dissection	division	embryology	etiology	examination	genetics	histology	inquiry	investigation	medicine	morphology	physiology	zoology	
basic	elemental	essential	key	main	necessary	primary	primitive	underlying	vital	capital	central	chief	principal	radical	basal	indispensabl	inherent
human	animal	mortal	personal	anthropoid	biped	hominid	individual	anthropologi	anthropomo	bipedal	civilized	creatural	ethnologic	ethological	fallible	fleshly	forgivable
method	approach	arrangement	design	form	habit	manner	mechanism	mode	plan	practice	process	program	recipe	rule	scheme	style	system
cell	bacterium	egg	germ	unit	corpuscle	embryo	follicle	microorganis	spore	utricle	vacuole	cellule	haematid				
medicine	antibiotic	cure	drug	medication	pharmaceuti	pill	prescription	remedy	anesthetic	antidote	antiseptic	antitoxin	balm	capsule	dose	elixir	injection

Рисунок 1 — Фрагмент тезауруса, созданный с использованием Thesaurus.com.

Слово на русском	Синоним 1	Синоним 2	Синоним 3	Синоним 4	Синоним 5	Синоним 6	Синоним 7	Синоним 8	Синоним 9
болезнь	недомогание	хворь	страдание	недуг	хвороба	расстройство	зараза	слабость	заболевание
медицина	наука о жизни	помощь	спасение	врачевание	лечение				
молекулярный	моляльный	состоящий из частиц							
биология	наука	естествознание							
основа	база	фундамент	основа	базис	столп	каркас			
человек	личность	тело	организм	пациент	индивид	существо	единица		
метод	тактика	подход	техника	способ	инструмент				
клетка	единица	структура	ячейка	макромер					
лекарство	препарат	медикамент	панацея	антидот	средство				

Рисунок 2 — Фрагмент тезауруса, созданный экспертами.

Для оценки эффективности тезауруса были предложены следующие параметры: мощность, связанность, плотность, коэффициент кластеризации и энтропия терминов. Мощность тезауруса характеризует общий объём информации и определяется как сумма числа вершин и рёбер графа, представленная формулой 2:

$$P = |V| + |E|, \qquad (2)$$

где P – мощность тезауруса; V – число вершин графа; E – число ребер графа.

Для оценки характера связности используется средняя степень вершины, представленная формулой 3:

$$\bar{d} = \frac{2|E|}{|V|}, \qquad (3)$$

где $\bar{d}$ - средняя степень вершины; |E| — число рёбер; а |V| — число вершин, показывающая, насколько тесно элементы сети связаны между собой. В экспертных тезаурусах этот показатель, как правило, имеет умеренные значения, что свидетельствует о сбалансированности структуры, тогда как в автоматически построенных тезаурусах часто наблюдается его завышение за счёт слабых статистических связей.

Дополнительно для нормированной оценки структуры применяется плотность графа, представленная формулой 4:

$$D = \frac{2|E|}{|V|(|V|-1)}, \qquad (4)$$

где D – плотность графа; |E| — число рёбер; |V| — число вершин. Значения метрики лежат в диапазоне $0<D\leq1$. Если $D\rightarrow1$, граф близок к полному, что говорит о сильной связности, но при этом может означать отсутствие структурной дифференциации. Если $D\rightarrow0$, граф разрежен, что может указывать на неполноту или избыточную фрагментированность семантической сети.

Коэффициент кластеризации отражает склонность вершин к образованию локальных групп — семантических кластеров. Для каждой вершины он вычисляется по формуле 5:

$$C_i = \frac{2e_i}{k_i(k_i-1)}, \qquad (5)$$

где C_i - коэффициент кластеризации; e_i — количество рёбер между соседями вершины; k_i — степень вершины (число её соседей).

Средний коэффициент кластеризации для всего графа определяется формулой 6:

$$C = \frac{1}{|V|}\sum_{i=1}^{|V|} C_i, \quad (6)$$

где |V| — число вершин; C_i - коэффициент кластеризации для выбранной вершины. Высокое значение C свидетельствует о том, что слова, связанные с одним понятием, также тесно связаны между собой, образуя устойчивые тематические группы.

В контексте NLP это соответствует свойству когерентности понятий — способности тезауруса группировать близкие по смыслу элементы.

Для оценки степени упорядоченности и разнообразия связей используется показатель энтропии.

Энтропия вычисляется по формуле 7:

$$H = -\sum_{i=1}^{|V|} p_i \log_2 p_i, \quad (7)$$

где

$$p_i = \frac{d_i}{\sum_{j=1}^{|V|} d_i}$$

где d_i — степень вершины v_i.

Показатель энтропии отражает то, насколько равномерно распределены семантические связи между вершинами графа.

Таким образом, для создания вектора признаков тезауруса T формируется вектор:

$$f(T) = \left(\tilde{P}(T), \tilde{\tilde{d}}(T), \widetilde{D}(T), \tilde{C}(T), \widetilde{H}(T)\right), \quad (8)$$

где $\tilde{P}(T)$ – нормализованное значение мощности, $\tilde{\tilde{d}}(T)$ – нормализованное значение средней степени вершины, $\widetilde{D}(T)$ – нормализованное значение плотности, $\tilde{C}(T)$ – нормализованное значение среднего коэффициента кластеризации, $\widetilde{H}(T)$ – нормализованное значение энтропии.

Итоговая метрика эффективности:

$$E(T_1, T_2) = \frac{1}{1+\sqrt{(f(T_1)-f(T_2))^{\top} W (f(T_1)-f(T_2))}}, \quad (9)$$

где $f(T_n)$ – вектор признака тезауруса, W – диагональная матрица весов, $^{\top}$ – операция транспонирования.

Алгоритм определения эффективности тезауруса следующий:

1. Выделение ключевых терминов;
2. Построение графа тезауруса;
    a. Вершины – леммы;
    b. Ребра – связи;
    c. Веса – сила связей, определенная с помощью обратного индекса;
3. Определение основных метрик, представленных выше;
4. Нормализация показателей;
5. Вычисление эффективности.

Итоговый набор характеристик позволяет всесторонне оценить качество тезауруса, на основании которого были проанализированы экспертные и автоматически собранные тезаурусы и сделаны выводы об их эффективности.

В результате применения разработанного методического аппарата выполнено экспериментальное сравнение экспертных и автоматически сформированных тезаурусов в ряде предметных областей. Разработана метрика эффективности тезауруса.

В главе 3 «Разработка графовой языковой модели» предлагается метод построения графовой языковой модели на основе графовых грамматик. Подробно описываются принципы формирования графа, где вершины соответствуют словам или контекстным полям, а рёбра отражают семантические связи между ними. В главе рассматривается:

1. Метод создания графовой языковой модели на основе графовых грамматик;
2. Разработка метода гибридизации: построение графа, вершинами которого являются векторные модели документов;
3. Алгоритм приведения графовой модели к векторной форме;
4. Возможность использования разработанных моделей в существующих интеллектуальных системах.

Метод сбора и использования локальных контекстов для формирования графовой естественно-языковой модели предметной области:

1. **Инициализация:** загрузка корпуса текстов и подготовка лексической базы;
2. **Предобработка:** лемматизация, фильтрация и выделение ключевых единиц;
3. **Формирование графа:** создание множества вершин и рёбер, вычисление весов;
    a) Выбор типов семантических связей для дополнения модели;
    b) Заполнение модели дополнительными ребрами и весами, соответствующими выбранным параметрам;
4. **Установка локальных весов:** уточнение идентификаторов слов с учётом структуры графа;
5. **Оптимизация структуры:** итерационное обновление весов и связей для повышения когерентности графа;
6. **Векторизация графа:** обучение оператора для получения финальных представлений;
7. **Экспорт модели:** формирование итоговых матриц эмбеддингов для дальнейшей интеграции в нейронные и гибридные языковые модели.

Для векторизации модели был применен алгоритм Вайсфейлера–Лемана в модифицированной форме с пороговой фильтрацией, что обеспечивает корректный механизм измерения структурно-семантической близости графов, на основании которого выполняется рубрикация документов.

Алгоритм создания и приведения графовой естественно-языковой модели к векторному представлению:

1. Предобработка документа;
2. Построение графа документа. Формируется граф локальных контекстов$G_d = (V_d, E_d, W_d)$, где V_d — множество терминов документа; E_d — рёбра совместной встречаемости; $W_d(u, v)$ — вес связи между терминами u и v;
3. Релабелинг вершин;
4. Формирование мультисета меток графа;
5. Построение словаря признаков;

6. Векторизация графа как гистограммы меток;
7. Нормализация.

Таким образом, графовая модель приводится к векторному представлению, которое можно затем интегрировать в интеллектуальную систему для обучения или работы аналогично классическим эмбеддингам.

Метод рубрикации документов с использованием GLM:

1. Построение графов классов. Для каждой рубрики предварительно формируется новый агрегированный граф;
2. Пороговая фильтрация рёбер;
3. Вычисление структурной близости (WL-ядро);
4. Спектральная оценка близости. Вычисляются собственные значения лапласианов графов: $\lambda(G) = (\lambda_1, \lambda_2, \dots, \lambda_k)$;
5. Интегральная метрика соответствия. Итоговая мера близости определяется как комбинация WL-ядра и спектральной метрики;
6. Выбор рубрики. Документ рубрицируется в класс, для которого достигнуто максимальное значение интегральной метрики.

Таким образом можно провести рубрикацию набора документов, сохраняя высокую точность графового метода. Тем не менее, подразумевается, что тексты набора соответствуют общей тематике.

В разделе разработана графовая языковая модель (GLM), объединяющая графовые и векторные подходы к представлению лексико-семантических структур. Показано, что использование формализма графовых грамматик и методов структурного анализа позволяет сохранять топологические свойства языка и учитывать контекстные зависимости, выявляемые статистическими и нейросетевыми методами.

В главе 4 «**Экспериментальная оценка эффективности графовой языковой модели**» рассматриваются вопросы интеграции разработанных графовых моделей с современными нейросетевыми архитектурами и приводятся результаты их применения в задачах рубрикации документов и выбора ответов в диалоговых системах. Показано, что использование графовых представлений позволяет

повысить точность решений при сохранении приемлемой вычислительной сложности. Также обсуждаются возможности практического применения предложенного подхода в интеллектуальных поисковых системах и чат-ботах на примере предметных областей «медицина» и «разработка программного обеспечения».

В рамках исследования был проведён эксперимент по апробации предложенных алгоритмов рубрикации.

Первый эксперимент был направлен на оценку эффективности графовой языковой модели (GLM) в задаче рубрикации текстов и её сравнение с традиционными векторными моделями. Для этого был сформирован корпус документов по предметным областям «информатика» и «медицина», включающий научные тексты, распределённые по тематическим подклассам. На основе корпуса были построены векторные представления и обучена графовая модель, после чего выполнена рубрикация документов. Полученные результаты использовались для анализа влияния структурной информации, зафиксированной в графовой модели, на качество выделения тематически однородных классов. По результатам эксперимента повышение эффективности составило до 26% по метрике accuracy.

Следующий эксперимент состоял в оценке эффективности графовой языковой модели (GLM), приведенной к векторному представлению, при рубрикации сложноразделимых медицинских текстов, описывающих различные заболевания. Для проведения эксперимента был сформирован датасет, включающий 20 классов заболеваний с несколькими примерами научных статей для каждого. Дополнительно использовался расширенный корпус из 200 публикаций для проверки качества рубрикации. Полученные результаты применялись для анализа способности модели корректно различать близкие по смыслу медицинские тексты.

Использование относительно компактного корпуса позволило минимизировать влияние избыточных статистических зависимостей и выявить вклад структурных характеристик, таких как связность, плотность и энтропия семантического графа. При этом классы формировались тематически однородно, а тексты проходили единообразную предобработку, что обеспечило сопоставимость

результатов. Следует отметить, что полученные результаты интерпретируются как подтверждение качественных закономерностей, связанных с влиянием структуры графа на эффективность моделей, и не претендуют на абсолютную оценку качества рубрикации в масштабных прикладных системах.

В итоге, при прочих равных параметрах вектора с использованием векторных моделей и векторного представления модели GLM удалось верно рубрицировать 18 из 20 подклассов, используя GLM, 17 из 20, используя BERT, и 14 из 20, используя Word2Vec.

В разделе экспериментально подтверждена эффективность графовой языковой модели (GLM) в задачах рубрикации документов и выбора ответов в диалоговых системах. Показано, что GLM превосходит модели Word2Vec и BERT по основным метрикам качества за счёт учёта как контекстных, так и структурных семантических связей. Сохранение топологии языка обеспечивает более высокую интерпретируемость и согласованность формируемого лексико-семантического пространства, что подтверждает перспективность графового подхода для практических задач обработки естественного языка и его интеграции в интеллектуальные системы.

ЗАКЛЮЧЕНИЕ

В диссертационной работе решена актуальная научно-техническая задача повышения эффективности автоматической рубрикации текстовых документов за счёт разработки методов и алгоритмов построения графовой языковой модели и её интеграции с векторными представлениями.

Поставленная цель, заключающаяся в разработке методов и алгоритмов построения и применения графовой языковой модели для повышения точности рубрикации документов, достигнута. Все сформулированные в работе задачи решены в полном объёме.

1. Проведён анализ существующих подходов к построению языковых моделей (экспертных, статистических и нейросетевых), выявивший ограниченность векторных моделей с точки зрения явного представления семантической структуры и трудности масштабирования экспертных тезаурусов;

2. Разработана интегральная метрика оценки эффективности тезаурусных структур на основе пяти топологических характеристик графа, а также алгоритм её автономного вычисления, позволяющие количественно сравнивать различные типы тезаурусов и выявлять некорректные структуры;
3. Предложен метод построения графовой языковой модели (GLM) на основе графовых грамматик и локальных контекстов, обеспечивающий явное представление семантических связей между лексическими единицами;
4. Разработан метод насыщения графовой языковой модели тезаурусными рёбрами, повышающий структурную связность модели и улучшающий её топологические характеристики;
5. Реализован гибридный подход к представлению текстов, в котором вершинами графа выступают векторные представления документов размерности 300 признаков, что обеспечивает совместное использование структурных и вероятностных характеристик и позволяет обрабатывать тексты объёмом 3000–5000 символов (≈400–800 токенов);
6. Разработан алгоритм приведения графовой модели к векторному представлению на основе модифицированного алгоритма Вайсфейлера–Лемана, обеспечивающий интеграцию графовых структур в нейросетевые методы обработки данных при сохранении не менее 90% структурной информации;
7. Проведены вычислительные эксперименты на корпусе из 1000 документов, показавшие преимущество предложенной графовой языковой модели по сравнению с векторными аналогами. Достигнуто повышение точности рубрикации до 0.94 (прирост до 16% по сравнению с Word2Vec), увеличение значения F1 (macro) до 0.97, а обобщённый прирост точности рубрикации составил до 26%.

Полученные результаты подтверждают эффективность предложенного подхода и его применимость для решения задач автоматической рубрикации текстов, анализа семантических связей и построения интеллектуальных информационных систем. Результаты работы апробированы на научных конференциях, опубликованы

в научных изданиях и внедрены в учебный процесс и научную деятельность, что подтверждает их практическую значимость.

Опубликованные статьи в журналах из списка ВАК:

1. Воробьёв Н. Г. Сравнение эффективности методов сбора и использования естественно-языковых моделей в медицине // Вестник РГРТУ. — 2025. — № 93. — С. 200–212.

2. Воробьёв Н. Г. Применение методов retrieval-augmented generation для автоматизации анализа сетевой инфраструктуры // XXI век: итоги прошлого и проблемы настоящего плюс. — 2025. — № 3(71). — Т. 14. — С. 65–70.

Опубликованные статьи РИНЦ:

1. Воробьёв Н. Г., Филиппович Ю. Н. Автоматический сбор локальных контекстов в системе диалогового взаимодействия // «Язык. Сознание. Коммуникация: методология и гуманитарные практики (вызовы современности)»: Материалы II Всероссийской конференции с международным участием, Москва, 08–11 апреля 2024 года – Москва: ИД "Канцлер", 2024. – с. 131-133.
2. Воробьёв Н. Г., Филиппович Ю. Н., Пшехотская Е. А. Методы оценивания качества тезаурусов // Искусственный интеллект в автоматизированных системах управления и обработки данных: сб. ст. II Всерос. науч. конф.: в 5 т. (Москва, 27–28 апр. 2023 г.). — М.: КДУ; Добросвет, 2023. — С. 258–263.
3. Vorobyev N. G., Philippovich Y. N., Pshehotskaya E. A., Philippovich A. Y. The comparison of local context collection systems and methods // Third International Conference on Optics, Computer Applications, and Materials Science (CMSD-III 2023) (Dushanbe, Dec. 20–22, 2023). — Vol. 13065. — Washington: SPIE, 2024. — P. 1306514. — DOI: 10.1117/12.3025183.
4. Воробьёв Н. Г. Методы создания контекстных моделей предметных областей // Искусственный интеллект в автоматизированных системах управления и обработки данных: сб. ст. III Всерос. науч. конф.: в 3 т. (Москва, 30 окт. — 1 нояб. 2024 г.). — М.: Издательский дом КДУ, 2025. — С. 183–189.

5. Воробьёв Н. Г. Методы автономного сбора и оценивания качества локальных контекстов // Современная наука: актуальные проблемы теории и практики. Сер.: Естественные и технические науки. — 2022. — № 5. — С. 46–50. — DOI: 10.37882/2223-2966.2022.05.06.

Регистрации программ ЭВМ:

1. Воробьёв Н. Г. Генератор моделей медицинских текстов: свидетельство о гос. регистрации программы для ЭВМ № 2025686741 (Рос. Федерация). — 2025.

2. Воробьёв Н. Г. Сравнение контекстной близости языковых медицинских моделей: свидетельство о гос. регистрации программы для ЭВМ № 2025685835 (Рос. Федерация). — 2025.

3. Воробьёв Н. Г. Конструктор контекстных графов: свидетельство о гос. регистрации программы для ЭВМ № 2025685987 (Рос. Федерация). — 2025.

4. Воробьёв Н. Г. Конвертер графовых языковых моделей: свидетельство о гос. регистрации программы для ЭВМ № 2025689874 (Рос. Федерация). — 2025.

Результаты работы были представлены на конференциях:

1. Воробьёв Н. Г. Методы автономного сбора и оценивания качества локальных контекстов — VIII Междунар. науч.-практ. конф. «Информационные технологии и машиностроение *(г. Москва, 2021 г.)*».
2. Воробьёв Н. Г. Сбор и оценивание локальных контекстов — науч.-практ. конф. «Промышленность 4.0 в России: технологии, материалы, опыт внедрения *(г. Барнаул, 2022 г.)*».
3. Воробьёв Н. Г. Сравнение систем сбора и использования локальных контекстов — HETS 2022: «II ежегодный междунар. науч.-практ. форум по проблемам устойчивого развития в цифровом мире: Человек. Экономика. Технологии. Социум *(г. Москва, 2022 г.)*».
4. Воробьёв Н. Г., Пшехотская Е. А., Филиппович Ю. Н. Методы оценивания качества тезаурусов — ИИАСУ’23 «Искусственный интеллект в автоматизированных системах управления и обработки данных *(г. Москва, 2023 г.)*» (МГТУ им. Н. Э. Баумана).

5. Воробьёв Н. Г. Метод проектирования чат-ботов для интерфейсов программных систем — VI Междунар. науч. форум профессорско-преподавательского состава и молодых учёных «Цифровые технологии: наука, образование, инновации *(г. Москва, 2023 г.)*» (МГТУ «СТАНКИН»).
6. Воробьёв Н. Г., Филиппович Ю. Н. Моделирование локальных контекстов в системе диалогового взаимодействия — науч. конф. «Язык. Сознание. Коммуникация: методология и гуманитарные практики (вызовы современности) *(г. Москва, 2024 г.)*».
7. Воробьёв Н. Г. Моделирование локальных контекстов в системе диалогового взаимодействия — ИИАСУ’24 «Искусственный интеллект в автоматизированных системах управления и обработки данных *(г. Москва, 2024 г.)*» (МГТУ им. Н. Э. Баумана).
8. Воробьёв Н. Г. Методы формализации естественного языка — Междунар. науч.-техн. форум «Современные технологии в науке и образовании *(г. Рязань, 2025 г.)*».
9. Воробьёв Н. Г. Приведение векторной графовой модели локальных контекстов к векторному представлению — ИИАСУ’25 IV Всерос. науч. конф. «Искусственный интеллект в автоматизированных системах управления и обработки данных *(г. Москва, 2025 г.)*» (МГТУ им. Н. Э. Баумана).