Министерство образования и науки Российской Федерации
Федеральное государственное бюджетное
образовательное учреждение высшего образования
«Комсомольский-на-Амуре государственный университет»

На правах рукописи

Галянт Владимир Павлович

Разработка интеллектуальной системы обработки заявок банковского сектора

Направление подготовки 09.04.03

«Прикладная информатика»

АВТОРЕФЕРАТ
МАГИСТЕРСКОЙ ДИССЕРТАЦИИ

2025

Работа выполнена в ФГБОУ ВО

«Комсомольский-на-Амуре государственный университет»

Научный руководитель: кандидат технических наук, профессор, сертифицированный специалист МА ТРИЗ Бердоносов Виктор Дмитриевич

Рецензент: программист высшей категории отдела мобильных решений ООО "Индорсофт", кандидат физико-математических наук Лошманов Антон Юрьевич

Защита состоится 20 июня 2025 года в 15 часов 10 мин на заседании государственной экзаменационной комиссии по направлению подготовки 02.04.03 «Прикладная информатика» в Комсомольском-на-Амуре государственном университете по адресу: 681000, г. Комсомольск-на-Амуре, пр. Ленина, 27, ауд. 204/5.

Автореферат разослан 20 июня 2025 г.

Секретарь ГЭК З.В. Широкова

ОБЩАЯ ХАРАКТЕРИСТИКА ДИССЕРТАЦИОННОЙ РАБОТЫ

Диссертационная работа посвящена разработке интеллектуальной системы обработки заявок в банковской сфере. В современных условиях цифровой трансформации банковских процессов значительно возрастает объем клиентских запросов и обращений в службы поддержки. Традиционные решения преимущественно регистрируют и маршрутизируют обращения, не учитывая смысловое содержание запросов. Применение семантического анализа и методов машинного обучения в обработке заявок позволяет повысить качество обслуживания клиентов, сократить время реакции и снизить нагрузку на операторов. Целью работы является разработка и реализация интеллектуальной системы, способной автоматически классифицировать и обрабатывать банковские заявки на основе семантического поиска и автоматических подсказок.

Для достижения поставленной цели сформулированы следующие задачи:

1. Анализ существующих подходов к автоматизированной обработке заявок в банковской и смежных областях, включая традиционные алгоритмы и нейросетевые методы.
2. Исследование методов применения векторных представлений текста и алгоритмов машинного обучения (в том числе нейронных сетей) для семантической обработки заявок.
3. Проектирование архитектуры системы обработки заявок с интеграцией семантического поиска и разработка прототипа на основе открытых технологий.

Объектом исследования является интеллектуальная система автоматизированной обработки заявок, использующая методы машинного обучения и векторный поиск, а **предметом исследования** - процесс обработки заявок в информационных системах банковского сектора.

В работе применяются методы системного и сравнительного анализа, проектирования информационных систем, технологии клиент-серверных архитектур, машинное обучение, методы обработки естественного языка и векторного поиска, а также элементы теории баз данных и программной инженерии.

Научная новизна исследования

- Предложена архитектура интеллектуальной системы обработки заявок, объединяющая традиционные подходы с механизмами семантического поиска, что позволяет повысить релевантность результатов и ускорить поиск решений.
- Разработана методика построения подсказок для пользователей на основе векторных представлений текстов заявок, что снижает количество дублирующих обращений и ускоряет время отклика системы.
- Проведена оценка экономической эффективности внедрения системы в реальных условиях, которая показала значительное сокращение операционных затрат и повышение производительности служб поддержки.

Практическая значимость работы

Разработанное решение может быть применено в службах технической поддержки банков, контакт-центрах и ИТ-отделах для автоматизации обработки внутренних и клиентских заявок. Внедрение семантических методов и механизма подсказок позволяет существенно сократить время реагирования и повысить точность перенаправления запросов к необходимым специалистам. Автоматизация обработки заявок приводит к снижению операционных издержек, снижению нагрузки на персонал и повышению удовлетворенности клиентов. Гибкость использованных технологий обеспечивает масштабируемость системы и возможность адаптации под задачи конкретной организации.

Апробация результатов исследования

Основные положения диссертации были доложены на профильных научных конференциях по информационным технологиям и получили положительные экспертные отзывы. Разработанный прототип системы был интегрирован в лабораторное информационное окружение учебного банка и протестирован на специально подготовленной базе данных. Эксперименты подтвердили высокую скорость обработки запросов и точность классификации заявок, что свидетельствует о практической применимости предложенного решения.

Публикации

- Галянт В.П. Векторы как универсальная форма представления данных. – МОЛОДЕЖЬ И НАУКА: АКТУАЛЬНЫЕ ПРОБЛЕМЫ ФУНДАМЕНТАЛЬНЫХ И ПРИКЛАДНЫХ ИССЛЕДОВАНИЙ. Часть 2 Материалы VIII Всероссийской национальной научной конференции молодых учёных Комсомольск-на-Амуре, 07-11 апреля 2025 г. – С. 481-484.

- Галянт В.П. Системы управления векторными базами данных. – МОЛОДЕЖЬ И НАУКА: АКТУАЛЬНЫЕ ПРОБЛЕМЫ ФУНДАМЕНТАЛЬНЫХ И ПРИКЛАДНЫХ ИССЛЕДОВАНИЙ. Часть 2 Материалы VIII Всероссийской национальной научной конференции молодых учёных Комсомольск-на-Амуре, 07-11 апреля 2025 г. – С. 484-488.

Структура и объем диссертации

Работа состоит из введения, трех глав, заключения, списка использованных источников и приложений. Общий объем диссертации – 68 страниц основного текста, дополнительно приведены 3 рисунка и 2 таблицы.

Список литературы содержит 21 источник. Текст изложен в соответствии с требованиями КнАГУ к оформлению авторефератов.

ОСНОВНОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновывается актуальность разработки интеллектуальных систем обработки заявок в условиях цифровой трансформации банковской сферы. Определены цель исследования, поставленные задачи, объект и предмет работы. Кратко изложены методы исследования и сформулированы научная новизна и практическая значимость работы.

Первая глава посвящена теоретическим аспектам семантического поиска. Алгоритмы семантического поиска позволяют не только находить документы по ключевым словам, но и интерпретировать запросы пользователей на более глубоком уровне, учитывая их намерения и контекст. Рассмотрим несколько примеров таких алгоритмов. Первым из них можно назвать TF-IDF (Term Frequency-Inverse Document Frequency). Это классический метод, который, впрочем, не утратил своей актуальности. TF-IDF рассчитывает важность термина в документе относительно общей базы данных, выделяя ключевые слова. Однако он не учитывает семантические связи между словами, а значит, его возможности ограничены. На более продвинутом уровне находятся алгоритмы, основанные на векторных представлениях слов. Наиболее известные из них:

1) Word2Vec - предлагает две архитектуры: модель CBOW (предсказание слова по контексту) и предсказание контекста по слову. Формально максимизирует правдоподобие соседних слов, используя многоклассовую функцию softmax:

$$L = -\sum_{t=1}^{T} \quad \sum_{-c\ j\ c,j0} \quad ln\ p(w_{t+j}|\ w_t), \qquad (1.1)$$

где вероятность слова в позиции $t+j$ при условии слова w_t задаётся как

$$p(w_O \mid w_I) = \frac{\exp(v_{w_O} v_{w_I})}{\sum_{w \in V} \exp(v_w v_{w_I})}. \quad (1.2)$$

Такие модели обучаются эффективно и демонстрируют высокую точность при семантических аналогиях и сходстве [9]. Отмечается, что предложенные архитектуры дают качественные представления при низкой вычислительной сложности. Эмбеддинги Word2Vec широко применяются как базовые признаки при семантическом поиске и IR-задачах.

2) GloVe - модель глобальной векторизации, которая строится на счётной матрице совместной встречаемости слов. Цель GloVe - аппроксимировать логарифм вероятности совместной встречаемости пар слов линейной комбинацией эмбеддингов. Формально вводится функция потерь:

$$J = \sum_{\mathrm{i,j=1}}^{\mathrm{V}} f(X_{ij}) \left(w_{\mathrm{i}}^{\mathrm{top}} j + b_i + \ln X_{ij} \right)^2, \quad (1.3)$$

где X_{ij} - число встреч слов i и j,

f - уменьшает вклад крайне редких/частых пар.

Такое обучение эффективно использует глобальную статистику корпуса. В итоге GloVe-эмбеддинги имеют свойство сохранять лингвистические закономерности (аналогии, сходство) и дают «векторное пространство со значимой подструктурой» [10].

3) FastText - расширяет Word2Vec, моделируя не только слова, но и подслова (символьные n-граммы). Идея в том, что каждое слово представляется суммой векторов его n-грамм. Это позволяет лучше учитывать морфологию (особенно в языках с богатой морфологией), строить векторы редким и даже не встречавшимся словам (за счёт их n-грамм) [19]. Подчеркивается, что представление «каждое слово как мешок n-граммов символов» ускоряет обучение и позволяет получить эмбеддинги для OOV-слов (отсутствующих в тренировочном корпусе). FastText демонстрирует SOTA-качество для нескольких языков и задач семантики.

В результате этих методов каждое слово w кодируется вектором $v_w \in R^d$. Для сравнения похожести слов или документов обычно вычисляют косинусную меру схожести:

$$\cos(u, v) = \frac{u \cdot v}{|u||v|}, \quad (1.4)$$

или скалярное произведение. Эмбеддинги слов могут суммироваться или усредняться для представления коротких текстов. Однако эти модели фиксированы - контекст слов не учитывается при разборе конкретного предложения. Поэтому далее мы перейдём к контекстным (трансформерным) моделям.

С появлением архитектуры Transformer стало возможным эффективное моделирование длинных контекстов через механизм самовнимания. Во внимание берутся матрицы запросов Q, ключей K и значений V, и вычисляется функция внимания:

$$\text{Attention}(Q, K, V) = \text{soft}\left(\frac{QK}{\sqrt{\text{d}_\text{k}}}\right)V\,. \quad (1.5)$$

Модель Transformer состоит из чередующихся блоков многоголового внимания и полносвязных слоёв, что даёт возможность эффективно обучать очень глубокие языковые модели. Эти модели генерируют контекстно-зависимые эмбеддинги: вектор одного и того же слова меняется в зависимости от предложения.

1) BERT. Модель предобучается на двух задачах: замаскировано предсказывает слова и предсказывает, следует ли одно предложение за другим. После этого её можно обучать на любых задачах классификации или поиска с добавлением одного выходного слоя. BERT легко достигает SOTA на множестве задач [7]. Для поиска вектор эмбеддинга запроса или документа можно получить, например, по вектору токена или путём усреднения выходных эмбеддингов.

2) RoBERTa - это оптимизированная версия BERT. Авторы провели углублённое воспроизведение эксперимента BERT и показали, что многие гиперпараметры и размеры корпуса недоиспользованы в оригинальном

обучении. BERT был значительно переобучен, и может превзойти каждую последующую после него модель, если настроить обучения правильно. Эта улучшенная модель достигает SOTA на GLUE, RACE и SQuAD. Таким образом, RoBERTa достигает больших результатов за счёт увеличения объёма данных, отказа от задачи NSP и тщательной настройки параметров. Это показывает, что ключевыми факторами эффективности трансформеров являются грамотное предобучение и архитектура attention.

3) Sentence-BERT. Авторский обзор подчёркивает, что при стандартном BERT для поиска по семантике требуется пропустить в сеть пару запрос-документ, что экстраординарно затратно: сравнение 10 000 предложений займёт десятки часов на BERT. SBERT модифицирует BERT в сиамскую архитектуру: два входа (запрос и документ) проходят через тот же энкодер независимо, после чего их эмбеддинги сравниваются (например, косинусом). Такая архитектура позволяет заранее закешировать эмбеддинги коллекции документов и находить ближайшие по близости векторы быстро. Это сокращает время поиска самого похожего предложения с 65 часов до ~5 секунд с сохранением точности". На практике SBERT значительно ускоряет поиск по сравнению со стандартными трансформерами, сохраняя схожие метрики семантической близости.

Во второй главе анализируются векторные базы данных как ключевой компонент семантического поиска. Даны определения векторных представлений данных и описаны принципы их хранения и поиска. Проведен сравнительный анализ популярных векторных СУБД (Pinecone, Weaviate, Qdrant, FAISS и др.) по критериям производительности и функциональности. Обоснован выбор платформы Qdrant для разработки прототипа системы ввиду ее высокой скорости поиска и гибкости настройки.

Третья глава описывает интеграцию семантических механизмов в систему поддержки HelpDesk. Приведена архитектура разработанной интеллектуальной системы обработки заявок: описаны модули сбора и

предобработки данных, обучения моделей машинного обучения, поиска и ранжирования заявок, а также пользовательского интерфейса. Рассмотрены сценарии работы системы на примере тестовых запросов банка. Проведены эксперименты по оценке точности классификации заявок и времени ответа. Результаты показывают, что использование семантических методов значительно повышает качество поиска по сравнению с традиционными алгоритмами.

В заключении подводятся итоги работы и сформулированы основные выводы и рекомендации. Обсуждаются перспективы дальнейшего развития, включая расширение функциональности системы и интеграцию дополнительных источников данных.

ВЫВОДЫ

1. Разработана архитектура интеллектуальной системы обработки банковских заявок с интеграцией семантического поиска и автоматических подсказок. Реализованный прототип показал высокую точность классификации заявок (около 90–95 % на тестовых данных) и значительно сократил время обработки по сравнению с ручными методами.
2. Предложенные методы семантического сопоставления на основе векторных представлений текстов доказали свою эффективность: они существенно повысили полноту и релевантность поиска решений заявок по сравнению с классическими алгоритмами.
3. Проведена оценка экономической эффективности системы: автоматизация обработки заявок позволит банку сократить операционные издержки и повысить производительность служб поддержки.
4. Апробация результатов на учебной базе данных банка показала устойчивое функционирование системы и перспективность её применения в реальных условиях.

5. Определены направления дальнейшего развития: расширение функционала с учетом обратной связи пользователей, интеграция дополнительных источников данных и использование облачных технологий для повышения отказоустойчивости и масштабируемости системы.

Положительные стороны исследования

- Высокая актуальность темы и практическая значимость результатов для автоматизации работы банковских контакт-центров.
- Комплексный подход к разработке системы с применением современных нейросетевых моделей и векторных технологий.
- Наличие реального прототипа системы, успешно прошедшего тестирование и показавшего высокую точность обработки заявок.
- Подтверждена экономическая целесообразность внедрения: расчет эффективности показал снижение издержек и повышение производительности.
- Модульная архитектура системы обеспечивает её гибкость, масштабируемость и легкость адаптации под конкретные задачи.

Недостатки проведенного исследования

- Эксперименты проведены на ограниченном наборе тестовых данных (учебной базе банка), что может не полностью отражать разнообразие реальных заявок.
- В работе не затронуты вопросы безопасности и конфиденциальности при обработке персональных данных клиентов банка.
- Система требует значительных вычислительных ресурсов для обучения и поддержки моделей, что усложняет ее внедрение в условиях ограниченной инфраструктуры.

- Не учтены возможные изменения предметной области (эволюция запросов клиентов со временем), что может потребовать дополнительных механизмов обучения на ходу.
- Требуется дальнейшая валидация и пилотное внедрение на реальных данных банков с оценкой опыта конечных пользователей и показателей эффективности в промышленной эксплуатации.

Список публикаций автора

1 Галянт В.П. Векторы как универсальная форма представления данных. – МОЛОДЕЖЬ И НАУКА: АКТУАЛЬНЫЕ ПРОБЛЕМЫ ФУНДАМЕНТАЛЬНЫХ И ПРИКЛАДНЫХ ИССЛЕДОВАНИЙ. Часть 2 Материалы VIII Всероссийской национальной научной конференции молодых учёных Комсомольск-на-Амуре, 07-11 апреля 2025 г. – С. 481-484.

2 Галянт В.П. Системы управления векторными базами данных. – МОЛОДЕЖЬ И НАУКА: АКТУАЛЬНЫЕ ПРОБЛЕМЫ ФУНДАМЕНТАЛЬНЫХ И ПРИКЛАДНЫХ ИССЛЕДОВАНИЙ. Часть 2 Материалы VIII Всероссийской национальной научной конференции молодых учёных Комсомольск-на-Амуре, 07-11 апреля 2025 г. – С. 484-488.