На правах рукописи

Макарова Елена Андреевна

МОДЕЛИ И АЛГОРИТМЫ ОБРАБОТКИ СЛАБОСТРУКТУРИРОВАННЫХ ТЕКСТОВЫХ ДАННЫХ НА ОСНОВЕ МЕТОДОВ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА

Специальность 2.3.1 – Системный анализ, управление и обработка информации, статистика

АВТОРЕФЕРАТ

диссертации на соискание учёной степени
кандидата технических наук

Брянск – 2023

Работа выполнена на кафедре «Информатика и программное обеспечение» ФГБОУ ВО «Брянский государственный технический университет»

Научный руководитель — **Лагерев Дмитрий Григорьевич**
кандидат технических наук, доцент

Официальные оппоненты — **Суркова Анна Сергеевна**
доктор технических наук, доцент,
профессор кафедры «Вычислительные системы и технологии» ФГБОУ ВО «Нижегородский государственный технический университет им. Р.Е. Алексеева»

Клышинский Эдуард Станиславович
кандидат технических наук, доцент,
доцент факультета гуманитарных наук и школы лингвистики ФГАОУ ВО Национальный исследовательский университет «Высшая школа экономики»

Ведущая организация — ФГБОУ ВО «Уфимский университет науки и технологий»

Защита состоится 14 июня 2023 года в 12:00 часов на заседании диссертационного совета 24.2.375.01, созданного на базе ФГБОУ ВО «Рязанский государственный радиотехнический университет им. В.Ф. Уткина», по адресу: г. Рязань, ул. Гагарина, 59/1.

С диссертацией можно ознакомиться в библиотеке ФГБОУ ВО «Рязанский государственный радиотехнический университет им. В.Ф. Уткина» и на сайте по адресу: http://rsreu.ru/post-graduate/zashchita

Отзывы на автореферат присылать по адресу: 390005, г. Рязань, ул. Гагарина, 59/1, ФГБОУ ВО «Рязанский государственный радиотехнический университет им. В.Ф. Уткина», ученому секретарю диссертационного совета 24.2.375.01.

Автореферат разослан « ____ » _____ 2023 года.

Ученый секретарь диссертационного совета 24.2.375.01
доктор технических наук, доцент — Пруцков Александр Викторович

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы исследования.

С развитием информационных технологий, растёт объём данных, доступных для обработки. В 2020 году количество открытых данных достигает уже 50 000 Эксабайт, из которых около 90% составляют слабоструктурированные и неструктурированные данные (IDC The Digital Universe, 2020). Часть из этих данных являются текстовыми, создаваемыми людьми на естественных или формальных языках. В то же время, в связи с высокими темпами изменений в экономической и социальной сфере, количество и скорость решений, которые необходимо принимать управленцам в различных предметных областях, непрерывно растёт. Ценная информация, которая может повлиять на принятие данных решений, часто содержится в виде слабоструктурированных текстовых данных. Примерами таких данных могут быть как открытые данные в Интернет-СМИ и социальных сетях, так и данных из полей свободного ввода в профессиональных закрытых базах данных. Чтобы эти данные были полезны в процессе разработки решений, их необходимо особым образом собирать, классифицировать и обрабатывать.

Многие современные платформы для анализа данных (далее – АД) поддерживают обработку слабоструктурированных текстовых данных (далее – ССТД). Однако, вопросы выбора подходящих методов на этапе подготовки текстовых данных всё ещё остаются в зоне ответственности лица, принимающего решения или специалиста по анализу данных. По оценкам экспертов, данный этап может занимать до 70% трудозатрат в процессе анализа. Из-за этого важным шагом перед включением в модель анализа данных ССТД является разведочный анализ, одним из целей которого является определение ценности загружаемых данных в рамках решаемой задачи. Однако, большая часть методов и инструментов разведочного анализа, в том числе использующих визуализацию, не поддерживает работу с текстовыми данными.

ССТД из внутренних баз данных организаций, в тех случаях, где текст записывается в свободной форме, часто имеют особенности, которые затрудняют автоматическую обработку: специфические сокращения, ошибки и т.д. Без интеллектуальной обработки этих особенностей данные невозможно эффективно включать в модели АД, так как они порождают неоднозначности и могут снизить качество полученной модели. Кроме того, некоторые виды текстовых данных на естественном языке могут быть корректно обработаны только с привлечением эксперта в предметной области. Необходимость разметки этих данных ещё сильнее увеличивает трудозатраты на реализацию моделей АД. С учётом постоянных изменений ситуации, этот этап необходимо повторять регулярно при появлении новых данных. В свою очередь, увеличение времени на построение моделей АД влияет на скорость работы ряда систем, в которых они используются: систем искусственного интеллекта, систем поддержки принятия решений и т.д.

Таким образом, научной **задачей** диссертационной работы является разработка моделей и алгоритмов интеллектуальной **обработки регулярно обновляющихся слабоструктурированных текстовых данных** в системах,

использующих анализ данных. В данной работе под обработкой понимается процесс извлечения и трансформации данных для применения в системах, использующих анализ данных.

Объектом исследования являются слабоструктурированные текстовые данные на естественном языке.

Предметом исследования являются модели и алгоритмы интеллектуальной обработки регулярно обновляющихся слабоструктурированных текстовых данных на естественном языке, в том числе на русском.

Целью диссертационной работы является развитие моделей и алгоритмов интеллектуальной обработки слабоструктурированных текстовых данных в системах, использующих анализ данных.

Для достижения поставленной цели необходимо решить ряд **задач**:

1. Усовершенствовать модель интеллектуальной обработки ССТД в системах, использующих анализ данных.
2. Сформировать визуальные модели больших массивов ССТД для поддержки контроля извлечения и разведочного анализа.
3. Разработать алгоритм трансформации специфических сокращений в ССТД, учитывающий особенности русского языка.
4. Разработать алгоритм определения семантической близости ССТД, позволяющий вычислить степень близости текстов и минимизирующий количество запросов к эксперту в процессе обработки данных.
5. Выполнить программную реализацию предложенных моделей и алгоритмов и провести их апробацию на прикладных задачах обработки данных в системах, использующих анализ данных.

Степень разработанности темы. Обработка ССТД – часть общего процесса анализа данных. Данное исследование опиралось на достижения отечественных и зарубежных исследователей в области интеллектуального анализа слабоструктурированных данных, таких как: В.Ф. Хорошевский, Д.А. Поспелов, Ю.И. Журавлев, К.В. Рудаков, А.Д. Наследов, Н.В. Клячкин, В.Е. Васильев, K. Borne, H. Jiawei, Ch. Faloutsos и др. Методы визуализации и разведочного анализа данных, в том числе ССТД, описывались в работах И.С. Бороздиной, А.А. Захаровой, G. Hinton, L. Maaten, O, Kaser, T. Allen и др. Особенности обработки текстовых данных на естественном языке рассматривали Э.В. Попов, Д.Г. Лахути, С.О. Кузнецов, Н.В. Лукашевич, К.В. Воронцов, Э.С. Клышинский, А.В. Пруцков, А.С. Суркова, St. Soderland, Y. Matsumoto, M. Kreuzthaler, T Mikolov, E. Hovy и др.

Научная новизна работы.

1. **Усовершенствована** модель интеллектуальной обработки данных, **за счёт** формализации задачи управления обработкой ССТД, в системах, использующих анализ данных. В рамках применения усовершенствованной модели возможно использование разработанных в диссертационной работе визуальных моделей, алгоритмов и интерактивных инструментов, основанных на сочетании методов искусственного интеллекта и классических подходов к визуализации и трансформации данных (*п. 2 паспорта специальности 2.3.1*).

2. Предложены визуальные модели больших массивов ССТД для поддержки контроля извлечения и разведочного анализа, **отличающиеся** применением методов машинного обучения для определения набора ключевых языковых единиц при построении визуализации типа «облако слов», а также **новым подходом** к визуализации статистических корреляций между языковыми единицами и другими переменными исследуемых данных (*п. 12 паспорта специальности 2.3.1*).

3. Впервые предложен алгоритм трансформации специфических сокращений в ССТД, учитывающий особенности сокращений на русском языке, основанный на статистических подходах и использовании методов машинного обучения для вычисления векторного представления слов (*пп. 4, 12 паспорта специальности 2.3.1*).

4. Создан алгоритм определения семантической близости ССТД с возможностью настройки необходимого уровня сходства на основе экспертной информации, **отличающийся** поддержкой в рамках решения одной задачи метрик, основанных на технологиях искусственного интеллекта, и классических метрик семантической близости, а также возможностью повторного использования экспертной информации для новых данных (*п. 4 паспорта специальности 2.3.1*).

Методы исследования. В ходе выполнения работы применялись методы системного анализа, методы анализа текстовых данных, методы машинного обучения, обработки естественного языка, получения и обработки экспертной информации, статистики, визуализации, определения семантической близости текстовых данных. При разработке программного комплекса использовались методологии и подходы объектно-ориентированного проектирования, разработки и автоматизированного тестирования программного обеспечения.

Теоретическая значимость работы заключается:

- в адаптации моделей и алгоритмов обработки ССТД к использованию в условиях постоянного накопления новых данных и необходимости привлечения экспертов в различных предметных областях;
- в развитии математического аппарата обработки ССТД за счёт формализации свойств, характеризующих эти данные;
- в развитии технологий обработки ССТД за счет создания новых моделей и алгоритмов, а также расширения и улучшения существующих свойств, характеризующих эти данные.

Практическая значимость работы:

1. Разработан программный комплекс для интеллектуальной обработки ССТД в системах анализа данных.
2. Созданы модели Word2Vec, обученные на наборах данных рынка труда и данных из региональной системы здравоохранения, которые могут быть использованы при обработке текстовых данных в системах анализа данных.
3. Решен ряд практических задач, таких как:

- интеллектуальная обработка обезличенных ССТД из интегрированных электронных медицинских карт (ИЭМК) пациентов для дальнейшего использования в моделях анализах данных;
- интеллектуальная обработка ССТД из описаний вакансий сферы ИТ с целью анализа актуальных технологий;
- интеллектуальная обработка данных о рынке труда с целью проведения социологических исследований.

Положения, выносимые на защиту.

1. Усовершенствованная модель интеллектуальной обработки ССТД, преимуществом которой является поэтапный отбор и обработка данных с привлечением эксперта в предметной области, для дальнейшего применения в системах, использующих анализ данных. Последовательное использование рекомендуемых моделью этапов обработки данных обеспечивает сокращение времени дальнейшей обработки данных на 14-28%, в зависимости от параметров данных.
2. Визуальные модели больших массивов ССТД для поддержки контроля извлечения и разведочного анализа данных, с целью их дальнейшего применения в системах, использующих анализ данных, что обеспечивает повышение эффективности решения задач, связанных с выбором значимых данных на этапе разведочного анализа и построением гипотез, за счет сокращения времени работы специалиста до 75%.
3. Алгоритм трансформации специфических сокращений в ССТД, учитывающий особенности сокращений на русском языке, позволяющий раскрывать до 90% сокращений в данных, насыщенных несловарными сокращениями, уменьшая количество обращений к эксперту в 9 раз, с обеспечением требуемого уровня качества.
4. Алгоритм определения семантической близости ССТД, позволяющий подобрать метрику и степень близости для предметной области и уменьшить количество запросов к эксперту в процессе обработки обновленных данных на 8-12%. Использование алгоритма позволяет определять от 19 до 28% процентов дублей, в зависимости от параметров данных.

Личный вклад соискателя. Все модели и алгоритмы, выносимые на защиту, а также реализующее их программное обеспечение, разработаны лично автором. Постановка задач исследования, формулировка положений научной новизны, а также выбор постановка задач экспериментальной проверки и апробации результатов исследования осуществлялись совместно с научным руководителем.

Апробация работы. Основные положения и результаты диссертационной работы докладывались и обсуждались на: VI Всероссийской научной конференции «Информационные технологии интеллектуальной поддержки принятия решений» ITIDS' 2018 (Уфа, 2018); 29-ой Международной конференции по компьютерной графике и машинному зрению «Графикон 2019» (Брянск, 2019); VII Международной конференции «Физико-техническая информатика – СРТ 2019» (Пущино, 2019); XXVI Международной научно-технической

конференции НГТУ им. Р.Е. Алексеева «Информационные системы и технологии. ИСТ-2022» (Нижний Новгород, 2020); 30-ая Международной конференции по компьютерной графике и машинному зрению «Графикон 2020» (Санкт-Петербург, 2020); X Международной научно–практической конференции имени А. И. Китова «Информационные технологии и математические методы в экономике и управлении» ИТиММ-2020 (Москва, 2020); 32-ой Международной конференции по компьютерной графике и машинному зрению «Графикон 2022» (Рязань, 2022).

Имеется 2 акта о внедрении: в ООО «Офисные технологии», являющихся разработчиком региональной информационной системы «МЕД-Комплит: Электронная медицина» и ООО «Айти Про», 1 справка о внедрении в Управлении государственной службы по труду и занятости населения Брянской области.

Публикации. По теме диссертационной работы опубликовано 14 печатных работ, в том числе 4 – в рецензируемых научных журналах из перечня ВАК, 3 – в изданиях, индексируемых в международной библиографической базе Scopus. Получено 2 свидетельства о регистрации программы для ЭВМ.

Структура и объем диссертации. Диссертационная работа состоит из введения, четырёх глав, общих выводов, библиографического списка и приложений. Работа изложена на 166 страницах, содержит 34 таблицы, 36 рисунков и библиографический список из 151 наименования.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Во введении приводится обоснование актуальности темы диссертации, ставятся цель и задачи исследования, определяются научная новизна, теоретическая и практическая значимость работы, рассматривается структура работы, приводятся основные положения, выносимые на защиту.

В первой главе проведён анализ задачи интеллектуальной обработки слабоструктурированных текстовых данных в системах анализа данных.

Дано определение слабоструктурированных текстовых данных (ССТД), описаны их свойства, такие как гибкость передачи информации и сложность автоматической обработки. Выявлены основные проблемы, с которыми сталкиваются системы обработки ССТД при увеличении количества данных. Описаны концепции и принципы, которые позволяют повысить оперативность обработки больших массивов ССТД. Изучен опыт анализа ССТД в различных предметных областях. В большинстве рассмотренных работ особый акцент делается на этапах отбора и подготовки данных, так как в случае ССТД нет универсального решения, подходящего каждому типу и источнику данных и часто требуется привлечение эксперта.

Рассмотрены задачи обработки текстов, которые могут быть использованы при построении моделей АД, такие как: извлечение информации, анализ тональностей, тематическое моделирование, классификация текстов и т.д. Выполнен обзор методов искусственного интеллекта, используемых для обработки СТТД.

Исходя из проведённого обзора, рассмотрены методы поддержки специалиста в предметной области в процессе принятия решений на основе предоставления инструментов визуализаций и интеллектуальной обработки ССТД. Проведен обзор аналогов программного обеспечения, поддерживающих обработку слабоструктурированных текстовых данных на русском языке. Исходя из результатов обзора, выявлены главные функциональные требования к перспективным инструментам.

Сделан вывод о целесообразности включения в модели АД слабоструктурированных текстовых данных. Решением проблемы сложности обработки подобных данных может быть автоматизация, которая позволит объединить сильные стороны алгоритмов обработки языка и эксперта в предметной области.

Сформулированы задачи диссертационного исследования.

Во второй главе описаны разработанные модели и алгоритмы для обработки слабоструктурированных текстовых с целью анализа.

Предложена **модель интеллектуальной обработки слабоструктурированных текстовых данных с целью анализа.**

Описаны этапы создания и использования моделей АД в рамках методологии интеллектуального анализа данных CRISP-DM (рисунок 1). Цветом выделены этапы, в которых используются результаты диссертационной работы.

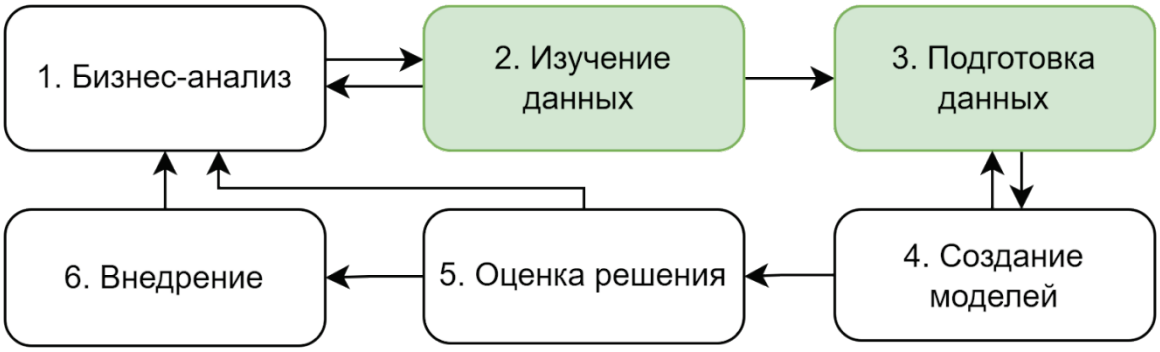


Рисунок 1 – Модель жизненного цикла исследования данных в методологии CRISP-DM

Описаны основные свойства модели обработки ССТД, которые формально задаются следующим образом:

$$A = < P, S_a, S, O, Pr; T >,$$

где P – параметры текстовых данных; S_a – общее количество текстовых данных в источнике, S – количество экземпляров текстовых данных, необходимых для обработки в рамках текущего анализа, O – накопленная информация о прошлых обработках, Pr – ряд преобразований, которые необходимо осуществить над данными; T – время на обработку массива данных, показатель, который стремится к уменьшению в данной модели.

В свою очередь, параметры текстовых данных (P) в данной модели описываются следующим образом:

$$P = <L, E, D >,$$

где L – Размер одного экземпляра; E – Наличие ошибок и специфических сокращений, D – Степень дублирования экземпляров. Определение параметров описано в диссертационной работе.

На рисунке 2 представлена модель интеллектуальной обработки ССТД для дальнейшего анализа, где этапы, улучшение которых производилось в рамках работы над диссертацией, выделены цветом.

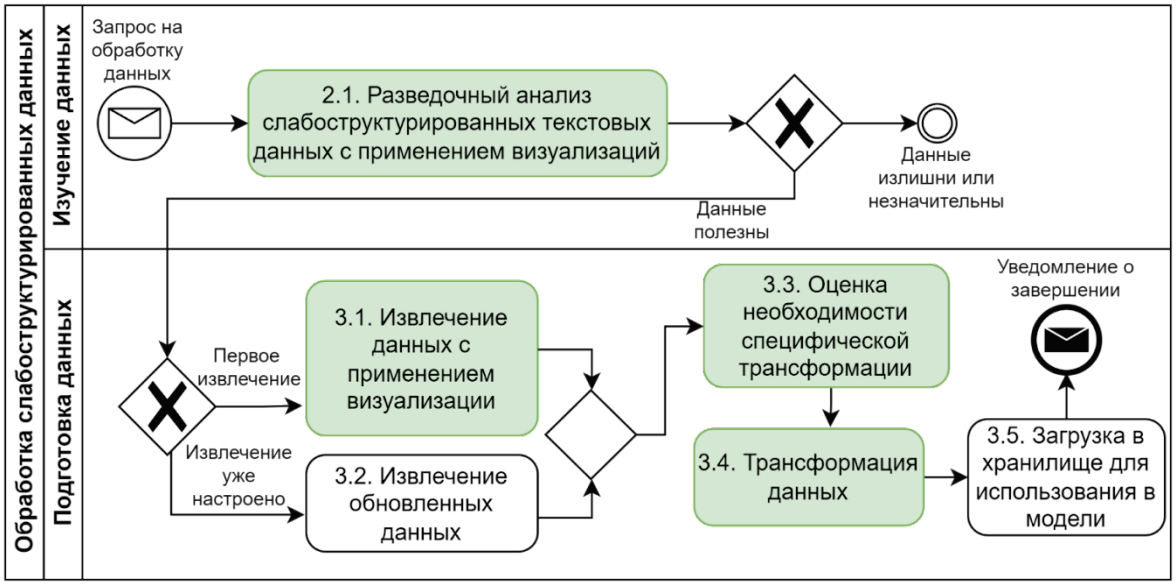


Рисунок 2 – Модель интеллектуальной обработки слабоструктурированных текстовых данных для дальнейшего анализа

Рассмотрены условия включения каждого из этих этапов в процесс обработки ССТД, исходя из количества и структуры данных. Только при соблюдении ряда условий время специалиста, затраченное на подготовительный этап обработки, будет оправданно.

Формально, целесообразность применения модели можно описать следующим образом:

$$T_a = Pre + \sum_{j=1}^{c} S_j * T_E * V \; ; T_m = \sum_{j=1}^{c} S_j * T_E \; ; T_a < T_m,$$

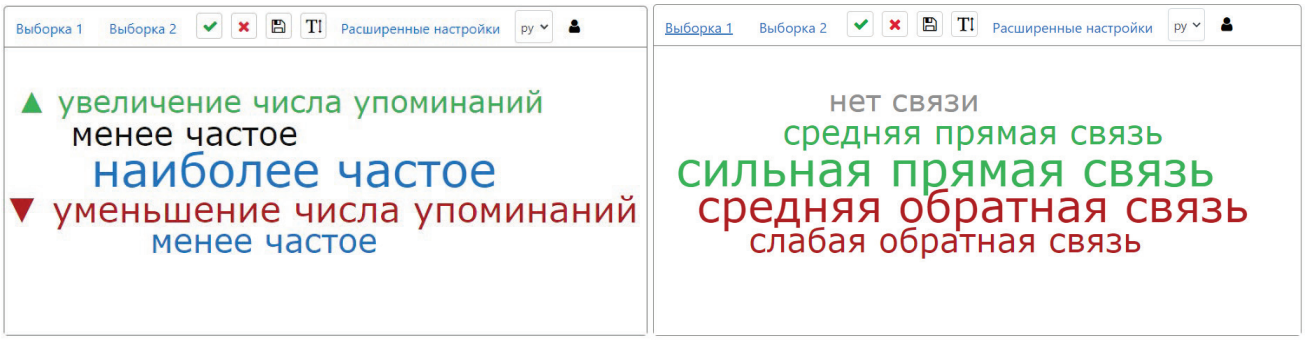
где T_a – общее время обработки при полностью ручном подходе; Pre – время, необходимое на первичную настройку; c – количество циклов обработки; j – номер цикла предобработки; T_E – среднее время работы эксперта над одной единицей данных (строкой или документом, в зависимости от задачи), включая время простоя; S_j – множество вариаций текстовых данных в цикле обработки j; V – доля данных для ручной валидации; T_m – общее время обработки при полностью ручном подходе.

Этап трансформации ССТД рассмотрен подробнее в следующих результатах диссертационной работы.

Рассмотрены существующие подходы к визуализации больших массивов текстов. Предложены **визуальные модели больших массивов слабоструктурированных текстовых данных для поддержки контроля извлечения и разведочного анализа данных**.

Предложены две визуальные модели (рисунок 3):

1. Интерактивная «количественная» визуализация.
2. Визуализация связей между языковыми единицами.



а) б)

Рисунок 3 – Интерактивная «количественная» визуализация (а); Визуализация связей между языковыми единицами (б)

Описан процесс использования визуальных моделей для поддержки контроля извлечения данных (блок 3.1 на рисунке 2), схема представлена на рисунке 4.

Рисунок 4 – Извлечение данных с применение визуализации

Описано использование методов обработки текста для подготовки данных к визуализации, в том числе с целью сокращения пространства ключевых языковых единиц. Описаны условия эффективного использования визуальных моделей, которые позволят сократить время эксперта на решение задач. Рассмотрены способы визуализации различий между выборками и математическое обеспечение, позволяющее определять разницу упоминаний языковых единиц в слабоструктурированных текстовых данных:

$$\text{если } \frac{t_2}{t_1} \geq 1{,}5, \text{то изменение положительное, если } 0{,}66 < \frac{t_2}{t_1} < 1{,}5,$$

$$\text{то изменения незначительное, при условии, что } |t_1 - t_2| \geq 10,$$

где t_1, t_2 – общее количество упоминаний для языковой единицы из первой и второй выборки соответственно.

Описана визуальная модель, позволяющая анализировать корреляции между упоминаниями языковых единиц и другими переменными объектов в выборке, которая может решать следующие задачи: поиск зависимостей, которые необходимо учесть при детальном исследовании и построении гипотез; отбор переменных для использования в моделях машинного обучения.

Предложен **алгоритм автоматизированной трансформации специфических сокращений в слабоструктурированных текстовых данных, учитывающий особенности сокращений на русском языке**, основанный на статистических подходах и использовании методов машинного обучения для вычисления векторного представления слов.

Описан статистический подход, позволяющий выделить специфические сокращения в тексте. Описан алгоритм поиска исходной формы сокращённого слова с целью дальнейшей трансформации. Автоматическая замена сокращений на найденное исходное слово возможно в случае, если:

1) выполняется одно из условий: сокращение совпадает с началом исходного слова; сокращение совпадает с началом исходного слова, при условии удаления гласных: начало и конец исходного слова совпадает с сокращением, середина заменяется дефисом;

2) исходное слово не является сокращением и присутствует в словаре употребимых слов русского языка;

3) исходное слово, удовлетворяющее вышеперечисленным условиям, только одно в диапазоне близости k от 0,8 до 0,99.

Предложены инструменты поддержки эксперта в процессе работы с текстами, насыщенными специфическими сокращениями.

Предложен **алгоритм определения семантической близости слабоструктурированных текстовых документов**, сочетающий ряд метрик определения семантической близости. Алгоритм позволяет определить для решаемой задачи подходящую метрику и порог семантической близости для поиска дублей или группировки результатов, основываясь на экспертной информации о разметке на предмет схожести специальным образом собранной выборки. Алгоритм подбирает метрику таким образом, чтобы уменьшить количество ошибок первого и второго рода, или выполняет вычисления, которые помогут подобрать приемлемый порог близости самому эксперту в предметной области. Одной из используемых в алгоритме метрик является Word Mover's Distance. Пример работы метрики приведен на рисунке 5. Приведение результатов работы метрики Word Mover Distance к классическим методам выполняется по формуле:

$$k = 1 - \frac{c}{w_m},$$

где c – евклидово расстояние между единицами ССТД, k – коэффициент подобия; w_m – количество слов в предложении (за исключением стоп-слов). Предложены подходы для совместного применения нескольких метрик семантической близости. Описан процесс валидации экспертом результатов работы выбранной на первом этапе алгоритма метрики, а также описан процесс обработки данных о валидации экспертом с целью повторного использования и корректировки результатов работы алгоритма.

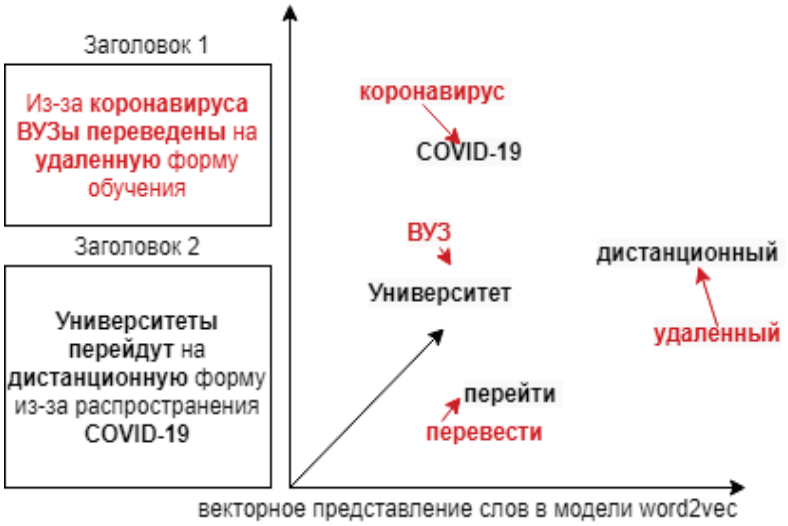


Рисунок 5 – Визуализация работы метрики Word Mover’s Distance

В третьей главе рассмотрена программная реализация разработанных моделей и алгоритмов.

Описаны функциональные и аппаратные требования к программному комплексу и варианты использования программного комплекса (ПК). Разработана архитектура ПК обработки ССТД. ПК построен на клиент-серверной архитектуре с использование фреймворка Django и REST API для связи клиента и сервера. На серверной части происходят все процессы обработки данных, на клиентской – интерфейс пользователя для настройки процесса обработки и инструменты контроля результата, включая визуализации. В качестве СУБД использована PostgreSQL.

Описаны основные библиотеки для языка программирования Python, используемые для решения задач по обработки текстовых данных, такие как NLTK, Gensim, Mystem3. Для управления очередью задач используется библиотеки Celery. Приведено обоснование выбора всех основных технологий и библиотек. Подробно описаны детали программной реализации ключевых подсистем, представлены макеты пользовательского интерфейса. Рассмотрены особенности создания интерфейсов, включающих визуализацию. Описан процесс функционального и нагрузочного тестирование подсистем ПК и его результаты. В результате функционального тестирования доказана устойчивость работы комплекса, в том числе для работы с большими массивами данных, и достоверность полученных результатов обработки данных по сравнению с полностью ручной обработкой данных. Описаны аппаратные и программные требования к серверу, на котором ПК может быть размещен. Описаны

возможности интеграции результатов работы комплекса в сервисы анализа данных в различных форматах, таких как JSON и XSL. Таким образом, решена задача создания ПК, реализующего модели и алгоритмы интеллектуальной обработки слабоструктурированных текстовых данных.

В четвертой главе описаны методы экспериментальной апробации разработанных моделей и методов. Разработанные модели и алгоритмы апробированы в трёх предметных областях. Предметные области подобраны таким образом, чтобы экспериментально доказать возможности работы модели интеллектуальной обработки с данными разного объёма и степени структурированности. Помимо демонстрации инвариантности применения самой модели к разным предметным областям, в ходе различных экспериментов проверено каждое из положений, выносимых на защиту.

Первая предметная область – интеллектуальная обработка слабоструктурированных текстовых данных из ИЭМК пациентов для дальнейшего использования в моделях анализах данных (блок 3.4 на рисунке 2, этап «Трансформация данных»). В рамках исследования было проведено два эксперимента обработки данных и апробирован алгоритм трансформации специфических сокращений на подзадаче улучшения данных.

Для апробации алгоритма была сформирована выборка из 60 000 записей диагнозов в слабоструктурированном текстовом виде из ИЭМК пациентов Брянской области. Была собрана группа экспертов для выделения и раскрытия вручную сокращений из 3000 записей (5% от общей выборки) описания диагноза. Примеры сокращений представлены в таблице 1.

Таблица 1 – Примеры сокращений в данных из системы здравоохранения

Исходное слово	**Сокращение 1**	**Сокращение 2**	**Сокращение 3**	**Сокращение 4**
стадия	«2 СТ.»	«ст.4»	«I ст.»	«ССТД»
отрицательный	«отриц»	«отрицат»	«отрц»	«отр»
метастазы	«метаст.»	«метастаз»	«MTS»	«МТС»

Для имеющейся выборки было подсчитано время и качество для трёх вариантов трансформации данных. Использование алгоритма позволило сократить время работы экспертов с приемлемой потерей качества трансформации, результаты приведены в таблице 2.

Таблица 2 – Результаты тестирования алгоритма

Описание подхода	**Время обработки 60 000 записей**	**Количество трансф. сокращений**
Полностью ручное исправление	Приблизительно 410 часов	Приближена к 100%
Полностью автоматический подход	От 5 до 10 минут*	До 53%
Алгоритм трансформации специфических сокращений	От 20 до 67 часов*	83-90%

Первый эксперимент – интеллектуальная обработка слабо-структурированных данных из описания диагнозов для использования в моделях АД, обрабатывающих данные пациентов, имеющих злокачественные новообразования. Классификация МКБ-10 не позволяет различать важные детали некоторых заболеваний, объединяя ряд схожих случаев под одним кодом, поэтому часто врачи пишут важные уточнения для диагноза в поле примечаний в свободной форме. Если не проведена дополнительная обработка, важные уточнения к диагнозу объединяются в модели АД под одним значением. В изучаемых данных по злокачественным новообразованиям лишь в меньшей части записей присутствовало указание стадии новообразования в явном виде, легко поддающимся автоматической обработке. Примеры записей представлены в таблице 3. Для достижения улучшения качества была применена модель интеллектуальной обработки, в рамках которой скорректирован процесс трансформации данных, с учетом наличия специфических сокращений. Результаты представлены в таблице 4.

Таблица 3 – Примеры записей стадий онкологических заболеваний

Пример записи	**Стадия**
Bl corpori uteri st 1 T1N0M0 после комб лечения в 2013г МТС в кости ,после луч терапии в процессе лечения б/фосф кл гр 4 (С53.9)	1
bl ovariorum st 3c T3cNxMO асцитная форма после комб лечения и 8 курсов пхт.Прогрессирование, кл гр 2	3
Bl.OV ARIORUM I V st Асцитная форма MTS плеврит	4

Таблица 4 – Результаты извлечения стадий онкологических заболеваний

Стадия / **Этап**	**I**	**II**	**III**	**IV**	**Стадия не определена**
До обработки экспертом	313	882	154	624	7610
После обработки	375	2971	425	1316	4496

При использовании модели обработки с привлечением эксперта становится возможным уменьшить трудоемкость задач, выполняемых экспертом в предметной области. Полученные результаты можно с небольшими трудозатратами масштабировать на другие нозологии (применялось также для обработки данных с целью использования в модели оценки рисков развития онкологических заболеваний).

Второй эксперимент – интеллектуальная обработка данных об анамнезе из ИЭМК пациента с целью дальнейшего использования в моделях АД. Даны обоснования важности включения данных об анамнезе пациента в модели АД, применяемые в процессе разработки управленческих решений в сфере здравоохранения. Описанные данные проанализированы с точки зрения обработки в предложенной модели интеллектуального анализа. В результате применены поэтапные преобразования над выборкой текстовых данных

об анамнезе пациента, включая очистку данных, трансформацию сокращений и ошибок и группировку по семантической близости. Результаты представлены в таблице 5.

Суммарное сокращение выборки перед классификацией: на 14,7%, что позволило сократить общее время обработки, с учетом валидации экспертом не менее 5% процентов записей, на 14,1%. Также расчеты показали, что от 8 до 12% записей, которые добавятся в последующие годы, будет обработано автоматически. В рамках данного эксперимента доказана важность соблюдения последовательности этапов, описанных в модели интеллектуальной обработки ССТД.

Таблица 5 – Результаты операций по обработке данных

Операция над данными	**Количество уникальных записей**	**Сокращение выборки (в %)**
Подготовка данных	72 597	2,2
Трансформация сокращений	65 696	9,3
Удаление достоверных дублей	63 315	3,2

Вторая предметная область – обработка слабоструктурированных текстовых данных из описаний вакансий ИТ-сферы с целью анализа актуальных технологий. На данной предметной области был использован контроль извлечения из источников и удаление дублирующийся информации (блоки 3.1 («Извлечение данных») и 3.4 («Трансформация данных») на рисунке 2).

Регулярная оценка технологических трендов – задача, которая стоит перед большинством ИТ-компаний. Описано, какие задачи может помочь решить анализ вакансий в сфере информационных технологий. Проведён анализ литературы по анализу вакансий и применению визуализации для решения этой задачи. Для решения этой задачи на основе разработанной модели использовались данные с сайта «Работа в России», из которых были выбраны вакансии в сфере информационных технологий. В данном примере важную роль будет играть корректное извлечение данных для анализа. С помощью визуальной модели контроля извлечения данных возможно настроить запрос нужным образом. Результаты скорректированного извлечения представлены в таблице 6.

Таблица 6 – Количество вакансий из области «Frontend разработка», найденных по разным типам запроса

Тип запроса	**2019**	**2020**	**2021**
Простой по заголовкам	15	12	18
Простой по требованиям	301	265	232
Сложный запрос, составленный с помощью визуального редактора	182	144	123

Из визуальной модели можно сделать вывод, что в выборке исследуемых вакансий значительно увеличился спрос на разработку для высоконагруженных систем и применение инструментов Docker для развертывания Frontend-приложений. Затраченное время на ручной и автоматизированный анализ трендов

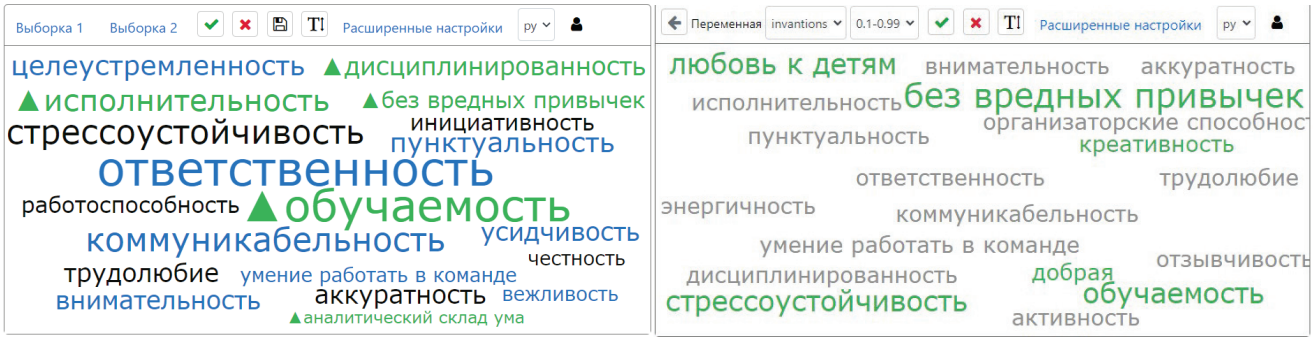
упоминаний в вакансиях представлено в таблице 7. В выборке 670 вакансий 149 были определены как дубли (повтор вакансии слово в слово) и 38 вакансий с высоким уровнем схожести по требованиям. Таким образом, выборка перед анализом или ручной классификацией ЛПР отдельных вакансий (если таковая требуется), **уменьшается на 28%**. Сокращение времени работы эксперта на этапе обработки данных позволяет повысить оперативность управленческих решений, которые необходимо принимать быстро и регулярно, учитывая обновленную информацию. При дальнейшем использовании полученных данных в СППР отсутствие дублирующийся информации в выборке способствует построению более качественных моделей АД.

Таблица 7 – Время ручного и автоматизированного анализа трендов упоминаний в вакансиях

Тип анализа	2019	2020	2021
Ручной	4,6 ч.	4,53 ч.	4,73 ч.
Автоматизированный	3 ч.	0,5 ч.	0,5 ч.

Третья предметная область – разведочный анализ слабоструктурированных текстовых данных из вакансий с целью отбора данных для дальнейшего исследования (пункт 2.1 «Создание и анализ визуальных моделей» на рисунке 2).

Описано, каким образом разведочный анализ позволяет сократить время на выбор значимых переменных на начальном этапе исследования и, в дальнейшем, избежать обработки излишних или незначительных. Описана апробация предложенных визуальных моделей на примере анализа рынка труда. Созданы визуальные модели различных текстовых данных из вакансий и резюме, таких как: требования к соискателю, должностные обязанности, навыки, описание опыта работы и т.д. Представлены примеры визуализации содержимого поля «гибкие навыки» из резюме соискателей и вакансий, отображающие как наиболее часто упоминаемые соискателями из различных профессиональных областей навыки, так и влияние упоминания этих навыков на приглашения соискателей на собеседования. Примеры визуализаций на рисунке 6.



а) б)

Рисунок 6 – Различия в упоминании «гибких навыков» среди резюме соискателей, получивших приглашение на собеседование, и случайными резюме (а); выявленные корреляции между указанием «гибких навыков» и приглашениями на собеседования для соискателей по профессии «Вожатый» (б)

Проведенная апробация показала, что использование разработанных визуальных моделей позволяет принять решение, необходимо ли включать текстовую переменную в модель АД на этапе разведочного анализа. В примере, рассматривающем резюме на должность «Программист» среди 3000 записей поле «гибкие навыки» было заполнено у 57% соискателей и составляли около 18% объема всех слабоструктурированных данных, заполненных в резюме. Временные затраты на разведочный анализ и возможности сокращения выборки в результате будет зависеть от специфики исследуемых данных.

ЗАКЛЮЧЕНИЕ

В диссертационной работе решена актуальная задача разработки моделей и алгоритмов интеллектуальной обработки регулярно обновляющихся ССТД на естественном языке, в том числе на русском, имеющая существенное значение для повышения эффективности процесса обработки ССТД в системах, использующих АД, и для повышения качества моделей АД.

Основные выводы и результаты работы заключаются в следующем.

1. Усовершенствована модель интеллектуальной обработки данных за счёт формализации задачи управления обработкой ССТД, в системах, использующих анализ данных, и использования разработанных в диссертационной работе визуальных моделей, алгоритмов и интерактивных инструментов, основанных на сочетании методов искусственного интеллекта и классических подходов к визуализации и трансформации данных; использование усовершенствованной модели обеспечивает сокращение времени дальнейшей обработки данных на 14-28%, в зависимости от параметров ССТД.
2. Предложены визуальные модели больших массивов ССТД для поддержки контроля извлечения и разведочного анализа, обеспечивающие повышение эффективности решения задач, связанных с выбором значимых данных за счет сокращения времени работы специалиста до 75%, и отличающиеся применением методов машинного обучения для определения набора ключевых языковых единиц при построении визуализации типа «облако слов», а также новым подходом к визуализации статистических корреляций между языковыми единицами и другими переменными исследуемых данных.
3. Впервые предложен алгоритм трансформации специфических сокращений в ССТД, учитывающий особенности сокращений на русском языке и позволяющий раскрывать до 90% сокращений в данных, насыщенных несловарными сокращениями, уменьшая количество обращений к эксперту в 9 раз, за счет использования статистических подходов и методов машинного обучения для вычисления векторного представления слов.
4. Создан алгоритм определения семантической близости ССТД, позволяющий определять от 19 до 28% процентов дублей, с возможностью настройки необходимого уровня сходства на основе экспертной информации, отличающийся поддержкой в рамках решения одной задачи метрик, основанных на технологиях искусственного интеллекта, и классических метрик семантической близости, а также возможностью повторного использования экспертной

информации для новых данных, что позволяет сократить количество обращений к эксперту на 8-12% для обновлённых данных.

5. Разработан программный комплекс «Text Preparation Wizard» для интеллектуальной обработки ССТД с целью дальнейшего применения в системах, использующих АД.

6. Выполнена апробация предложенных моделей и алгоритмов в трёх различных предметных областях на данных разного объема и степени структурированности: таких как: обработка обезличенных ССТД из интегрированных электронных медицинских карт (ИЭМК) пациентов для дальнейшего использования в моделях анализах данных; обработка ССТД из описаний вакансий ИТ-сферы с целью анализа актуальных технологий; обработка данных о рынке труда с целью проведения социологических исследований.

СПИСОК РАБОТ, ОПУБЛИКОВАННЫХ АВТОРОМ ПО ТЕМЕ ДИССЕРТАЦИИ

Публикации в изданиях, входящих в Перечень рецензируемых научных изданий ВАК

1. Лагерев, Д.Г. Поиск и раскрытие сокращений в русскоязычных данных медицинских информационных систем / Д.Г. Лагерев, **Е.А. Макарова //** Вестник компьютерных и информационных технологий, 2020 – № 7– С. 44-54.
2. **Макарова, Е.А.** Оценка семантической близости новостных сообщений на основе анализа заголовков / **Е.А. Макарова**, Д.Г. Лагерев // Вестник компьютерных и информационных технологий, 2021. – Т.18. – № 7(205). – С. 46-56.
3. **Макарова, Е.А.** Модель обработки слабоструктурированных текстовых данных на русском языке для интеллектуальной поддержки информационного управления в динамических организационных системах / **Е.А. Макарова**, Д.Г. Лагерев // Модели, системы, сети в экономике, технике, природе и обществе, 2022. – № 3. – С. 104-125.
4. **Макарова, Е.А.** Обработка слабоструктурированных текстовых данных для использования в моделях анализа / **Е.А. Макарова //** Информационные и математические технологии в науке и управлении, 2023. – № 1(29). – С. 178-189.

Публикации в изданиях, индексируемых в международной библиографической базе Scopus

5. **Makarova, E.A.** Approaches to visualizing big text data at the stage of collection and pre–processing / **E.A. Makarova**, D.G. Lagerev, F.Y. Lozbinev // Scientific Visualization, 2019. – Vol. 11 (4). – p. 13-26.
6. **Makarova, E.A.** Features of big text data visualization for managerial decision making / **E.A. Makarova**, D.G. Lagerev, F.Y. Lozbinev // CEUR Workshop Proceedings of the 29th International Conference on Computer Graphics and Vision (GraphiCon 2019), 2019. – Vol. 2485. – p. 99-102.
7. **Makarova, E.A**. Methodology for Preprocessing Semi–Structured Data for Making Managerial Decisions in the Healthcare / **E.A. Makarova**, D.G. Lagerev // CEUR Workshop Proceedings of the 30th International Conference on Computer Graphics and Vision (GraphiCon 2020), 2020 . –Vol. 2744. – https://ceur-ws.org/Vol-2744/paper78.pdf.

Публикации в прочих изданиях

8. **Макарова, Е.А.** Анализ неструктурированных данных с целью получения дополнительной информации при оценке кредитоспособности юридических лиц / **Е.А. Макарова**, Д.Г. Лагерев // Информационные технологии интеллектуальной поддержки принятия решений (ITIDS'2018): труды VI междунар. конф. – Уфа: УГАТУ, 2018 – Т. 1. – С. 189-195.
9. **Макарова, Е.А.** Оценка семантической ценности текстовой информации для разработки управленческих решений / **Е.А. Макарова**, Д.Г. Лагерев, А.А. Захарова // Междунар. конф. СРТ2019 (13-17 мая 2019 г., ЦарьГрад, Моск. обл.). – Нижний Новгород: Изд-во ННГАСУ и НИЦФТИ, 2019. – С. 356-360.
10. **Макарова, Е.А.**. Применение автоматизированной системы интеллектуального анализа текстовых данных для управления процессом формирования индивидуальных образовательных траекторий / **Е.А. Макарова**, Д.Г. Лагерев // [Электронный ресурс]: сборник материалов XXVI Международной научно–технической конференции – Электрон. дан. – Н. Новгород: Нижегород. гос. техн. ун-т им. Р.Е. Алексеева, 2020. – С. 362-367.
11. **Макарова, Е.А.** Автоматизация извлечения признаков из слабоструктурированных медицинских данных / **Е.А. Макарова**, Д.Г. Лагерев // X Международная научно-практическая конференция имени А. И. Китова «Информационные технологии и математические методы в экономике и управлении» (ИТиММ-2020). 15-16 октября 2020 г.: сб. статей. – М.: ФГБОУ ВО «РЭУ им. Г. В. Плеханова», 2020. – С 56-62.
12. **Макарова, Е.А.** Формирование индивидуальных образовательных траекторий студентов с учетом результатов анализа описаний вакансий / **Е.А. Макарова** // Современные технологии в науке и образовании – СТНО-2022 [текст]: сб. тр. V межд. науч.-техн. форума: в 10 т. Т.10./ под общ. ред. О.В. Миловзорова. – Рязань: Рязан. гос. радиотехн. ун-т, 2022. – С 101-106.
13. **Макарова, Е.А**. Поддержка процессов информационного управления с помощью программного сервиса для автоматизированной обработки слабоструктурированных текстовых данных на русском языке / **Е.А. Макарова** // Информационные технологии. Проблемы и решения, 2022. – № 3(20) . – С 50-56.
14. **Макарова, Е.А**. Использование визуальных моделей для разведочного анализа слабоструктурированных текстовых данных / **Е.А. Макарова**, Д.Г. Лагерев // GraphiCon 2022: труды 32-й Межд. конф. по компьютерной графике и машинному зрению (Рязань, 19-22 сент. 2022 г.). – М.: Институт прикладной математики им. М.В. Келдыша РАН, 2022. – С. 1094-1105.

Свидетельства о регистрации программы для ЭВМ

15. Свидетельство о государственной регистрации программы для ЭВМ № 2022660138. Подсистема для визуализации больших массивов слабоструктурированных текстовых данных: Российская Федерация / **Е.А. Макарова.** № 2022619487; заявл. 22.05.2022; опубл. 31.05.2022.
16. Свидетельство о государственной регистрации программы для ЭВМ № 2022662584. Подсистема для обработки слабоструктурированных текстовых данных на русском языке: Российская Федерация / **Е.А. Макарова.** № 2022662234; заявл. 28.06.2022; опубл. 05.07.2022.

Макарова Елена Андреевна

Модели и алгоритмы обработки слабоструктурированных текстовых данных на основе методов искусственного интеллекта

Автореферат
диссертации на соискание ученой степени кандидата технических наук

Подписано в печать: 05.04.2023 г.
Объем: 1,16 усл.п.л.
Тираж: 100 экз. Заказ № 0010

Отпечатано в типографии «Артель»
ИП Чуканов А.С., ИНН 712302042912, ОГРНиП 322710000010992
г. Брянск, пр-т Станке Димитрова, 28, оф. 205
+7 (4832) 77-02-72 http://art-pk.ru/