

На правах рукописи



Кравченко Даниил Юрьевич

**Модели и алгоритмы поиска, приобретения и
использования знаний в системах искусственного
интеллекта при обработке и анализе текстов на
естественном языке**

Специальность – 1.2.1. Искусственный интеллект и машинное обучение

Автореферат
диссертации на соискание ученой степени
кандидата технических наук

Таганрог – 2025

Работа выполнена в ФГАОУ ВО «Южный федеральный университет» на кафедре систем автоматизированного проектирования им. Виктора Михайловича Курейчика
Института компьютерных технологий и информационной безопасности

Научный руководитель:

доктор технических наук, профессор
Курейчик Владимир Викторович,

лауреат премии Правительства РФ в области образования, ФГАОУ ВО «Южный федеральный университет» (ЮФУ), заведующий кафедрой систем автоматизированного проектирования им. Виктора Михайловича Курейчика (г. Таганрог)

Официальные оппоненты:

доктор технических наук, профессор
Кравец Алла Григорьевна

ФГБОУ ВО «Волгоградский государственный технический университет» (ВолгГТУ), профессор кафедры «Системы автоматизированного проектирования и поискового конструирования» (г. Волгоград)

кандидат технических наук

Нагоев Залимхан Вячеславович

ФГБНУ «Федеральный научный центр «Кабардино-Балкарский научный центр Российской академии наук» (КБНЦ РАН), генеральный директор (г. Нальчик)

Ведущая организация:

ФГБОУ ВО «Ульяновский государственный технический университет» (УлГТУ, г. Ульяновск)

Защита диссертации состоится «03» апреля 2025 г. в 11⁰⁰ на заседании объединенного совета по защите диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук 99.2.107.02 на базе ФГАОУ ВО «Южный федеральный университет», ФГБОУ ВО «Южно-Российский государственный политехнический университет (НПИ) имени М.И. Платова» по адресу: 347922, г. Таганрог, пер. Некрасовский, 44, ауд. Г-439.

С диссертацией можно ознакомиться в зональной научной библиотеке ФГАОУ ВО «Южный федеральный университет» по адресу: 344015, г. Ростов-на-Дону, ул. Зорге, 21Ж, а также на портале электронных ресурсов ЮФУ: <https://hub.sfedu.ru/diss/show/1336486/>.

Отзыв на автореферат, заверенный гербовой печатью организации, просим направлять ученому секретарю объединенного диссертационного совета 99.2.107.02 по адресу: 347922, г. Таганрог, пер. Некрасовский, 44.

Автореферат разослан «____» января 2025 г.

Ученый секретарь объединенного
диссертационного совета 99.2.107.02,
доктор технических наук, доцент

Н. Е. Сергеев

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы диссертационного исследования.

Диссертационная работа посвящена решению актуальной научной проблемы искусственного интеллекта (ИИ), имеющей существенное научное и хозяйственное значение, которая заключается в необходимости повышения эффективности процессов *поиска, приобретения и использования знаний в системах искусственного интеллекта при обработке и анализе текстов на естественном языке*. Основной научной идеей данного исследования является переход от доминирования систем генеративного искусственного интеллекта, построенных на основе инструментов, представляющих собой «черный ящик», к более надежным интеллектуальным системам, созданным на основе *детерминированных методов и алгоритмов, позволяющих минимизировать время отклика системы на пользовательский запрос при условии обеспечения «прозрачности» процессов обработки входных данных*.

Поиск знаний (англ. *knowledge retrieval*) – процесс возврата информации к структурированной форме. Под *приобретением знаний* (англ. *knowledge acquisition*) для структурированной текстовой информации понимается систематизация полученных знаний через построение детализирующих семантику текста *гранул смыслла* – триплетов, состоящих из элементов «субъект» - «предикат» - «объект».

Степень разработанности темы диссертационного исследования.

Работы А.Е. Ермакова, И.А. Минакова, Е.А. Рабчевского, В.М. Курейчика, Т. Gruber, S.Lynn и D.W. Embley, J.Völker, D.Vrandečić, M. Sabou и Y.Sure и других исследователей в области решения задач построения онтологий вносят значимый вклад в решение обозначенной проблемы.

Наиболее известными работами в области решения задач поиска, приобретения и использования знаний являются труды следующих ученых: Э. Тоффлера; Д. Белла; М. Маклюэна; Ё. Масуды; Ю.А. Шрейдера; Р.С. Гиляревского; К. Виига; П. Сенге; И. Нонака; Х. Такеучи; Т. Давенпорта; Л. Прусака; К. Ньюэлла; Д. Смита; В.Л. Иноземцева; Б.З. Мильнера; А.Л. Гапоненко и многих других.

Объектом исследований в диссертации являются тексты на естественном языке.

Предметом исследований являются модели и алгоритмы поиска, приобретения и использования знаний в системах искусственного интеллекта при обработке и анализе текстов на естественном языке.

Целью диссертационного исследования является повышение эффективности моделей и алгоритмов поиска, приобретения и использования знаний в системах искусственного интеллекта при обработке и анализе текстов. Под эффективностью понимается минимизация времени отклика системы на запрос пользователя при условии обеспечения «прозрачности» процессов обработки входной текстовой информации.

Для достижения указанной цели в работе поставлены и решены следующие основные задачи:

1. Построена верхнеуровневая модель онтологии знаний, созданная на основе оригинальной компонентной архитектуры и применяемая при обработке и анализе текстов на естественном языке, которая позволяет обеспечить необходимую степень детализации анализируемой текстовой информации;

2. Построена нижнеуровневая модель онтологии знаний, созданная на основе оригинальной структуры отношений между понятиями и применяемая при обработке и анализе текстов на естественном языке, которая позволяет получить набор смысловых паттернов и проводить оценку их семантической близости;

3. Разработан алгоритм поиска знаний в текстах на естественном языке с применением графовых моделей при фильтрации информации на выходе парсера, извлекающий смысловую часть предложения для использования в процессах приобретения знаний;

4. Разработан алгоритм приобретения знаний в текстах на естественном языке с применением множества низкоуровневых правил семантического анализа полученных смысловых паттернов, позволяющий определить основные гранулы смысла для процессов использования знаний;

5. Разработан модифицированный биоинспирированный алгоритм использования приобретенных знаний в задачах генеративного искусственного интеллекта, основанный на улучшенных механизмах интенсификации поиска решений и процедурах выхода из локальных оптимумов для уменьшения времени отклика системы искусственного интеллекта и машинного обучения на пользовательский запрос при обработке и анализе текстов на естественном языке.

Методология и методы диссертационного исследования.

Методологической и теоретической основой проведенных исследований послужили положения теории искусственного интеллекта, биоинспирированной оптимизации, алгоритмов, графов, а также методы поиска, приобретения и использования знаний, имитационного, семантического и онтологического моделирования.

Научная новизна. Научной новизной проведенного исследования являются модели и алгоритмы поиска, приобретения и использования знаний в системах искусственного интеллекта при обработке и анализе текстов на естественном языке, которые направлены на решение научной задачи повышения эффективности и обеспечения «прозрачности» функционирования средств и инструментов генеративного искусственного интеллекта, что имеет важное значение для развития информатики, а именно:

1. Построена верхнеуровневая модель онтологии знаний, применяемая при обработке и анализе текстов на естественном языке, которая отличается включением в состав ее компонентов множеств понятий с различным уровнем нормализации, что позволяет обеспечить необходимую степень детализации анализируемой текстовой информации (пункт 4 паспорта специальности 1.2.1; страницы 61-68 диссертации);

2. Построена нижнеуровневая модель онтологии знаний, применяемая при обработке и анализе текстов на естественном языке, которая **отличается** использованием структуры отношений между понятиями, детализирующими семантику текстовой информации, что **позволяет** получить набор смысловых паттернов, а также проводить оценку их семантической близости (пункт 4 паспорта специальности 1.2.1; страницы 68-75 диссертации);

3. Разработан алгоритм поиска знаний в текстах на естественном языке, **отличающийся** созданием дополнительного фильтра на выходе парсера с применением графовых моделей, что **позволяет** извлечь смысловую часть предложения из полученной синтаксической схемы текстовой информации для использования в процессах приобретения знаний (пункт 5 паспорта специальности 1.2.1; страницы 82-99 диссертации);

4. Разработан алгоритм приобретения знаний в текстах на естественном языке, **отличающийся** применением множества низкоуровневых правил семантического анализа полученных смысловых паттернов, **позволяющий** определить основные гранулы смысла для процессов использования знаний (пункт 5 паспорта специальности 1.2.1; страницы 99-107 диссертации);

5. Разработан модифицированный биоинспирированный алгоритм использования приобретенных знаний в задачах генеративного искусственного интеллекта, **отличающийся** улучшенными механизмами интенсификации поиска решений и процедурами выхода из локальных оптимумов, что **позволило** уменьшить время отклика системы искусственного интеллекта и машинного обучения на пользовательский запрос при обработке и анализе текстов на естественном языке (пункт 5 паспорта специальности 1.2.1; страницы 107-120 диссертации).

Теоретическая значимость работы. Полученные научные результаты развивают аппарат искусственного интеллекта и машинного обучения в области решения важной научной задачи повышения эффективности при условии обеспечения «прозрачности» процессов *поиска, приобретения и использования знаний в системах искусственного интеллекта при обработке и анализе текстов на естественном языке*.

Практическая значимость работы заключается в создании программного приложения, позволяющего использовать разработанные модели и алгоритмы поиска, приобретения и использования знаний в системах искусственного интеллекта при обработке и анализе текстов на естественном языке для снижения времени отклика системы на запрос пользователя.

Степень достоверности результатов. Достоверность научных результатов работы подтверждается непротиворечивостью и согласованностью с известными фактами и исследованиями в рассматриваемой области, высокой степенью сходимости теоретических результатов с данными экспериментов и определяется применением теоретических и методологических основ разработок ведущих ученых, а также корректным и обоснованным использованием математического аппарата, экспериментальными исследованиями разработанных моделей и алгоритмов.

Реализация и внедрение результатов работы. Внедрение теоретических и практических результатов работы проводилось в сотрудничестве с проектной организацией ООО «Газэксперт плюс» (г. Краснодар). Полученные в работе научные результаты позволили повысить качество процедур поиска и интеграции прототипов проектных решений в газотранспортной отрасли. Теоретические и практические результаты исследований вошли в материалы отчетов по грантам РНФ № 22-71-10121 (2022-2024) и № 23-21-00089 (2023-2024), а также РФФИ № 20-01-00148 (2020-2022). Также теоретические и практические результаты, полученные в диссертации, внедрены в учебный процесс Южного федерального университета.

Апробация результатов. Основные теоретические положения и практические результаты диссертационной работы докладывались и обсуждались в рамках ряда научных мероприятий, основными из которых являются следующие: 7th Computational Methods in Systems and Software (Чехия, октябрь 2023); «5th International Scientific Convention UCIENCIA» (Куба, сентябрь 2023); Международный конгресс по интеллектуальным системам и информационным технологиям (IS&IT) (п. Дивноморское, Краснодарский край, Россия, сентябрь 2022-2023); XII Международная научно-техническая конференция «Технологии разработки информационных систем» (ТРИС) (г. Феодосия, Республика Крым, Россия, сентябрь 2022); 7-я Международная конференция по информационным технологиям в инженерном образовании (г. Москва, Россия, апрель 2024).

Публикации. По теме диссертации опубликовано 19 научных работ, из которых: 5 статей опубликованы в научных изданиях, индексируемых международными базами данных, перечень которых определен в соответствии с рекомендациями ВАК; 4 статьи – в издании из перечня, утвержденного ВАК, рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук, в т.ч. 1 научная работа, принадлежащая лично автору. Имеется 2 свидетельства о государственной регистрации программ для ЭВМ. В трудах всероссийских и международных конгрессов и конференций опубликовано 7 работ.

Личный вклад автора. Все выносимые на защиту результаты и положения, составляющие основное содержание диссертационной работы, разработаны и получены лично автором или при его непосредственном участии. В работах, опубликованных в соавторстве, соискателю принадлежит определяющая роль при решении задач поиска, приобретения и использования знаний при обработке и анализе текстов на естественном языке.

Структура и объем работы. Диссертация состоит из введения, 4 разделов, заключения, списка сокращений и условных обозначений, списка литературы, содержащего 122 наименования, 2 приложений. Основная часть работы содержит 164 страницы, включая 39 рисунков и 10 таблиц.

Область исследования. Диссертационная работа соответствует пунктам 4 и 5 паспорта научной специальности 1.2.1. Искусственный интеллект и машинное обучение (технические науки).

Положения, выносимые на защиту:

1. Верхнеуровневая модель онтологии знаний, представленная в виде оригинальной компонентной архитектуры, **позволяет** обеспечить необходимую степень детализации анализируемой текстовой информации;

2. Нижнеуровневая модель онтологии знаний, представленная в виде оригинальной структуры отношений между понятиями, детализирующими семантику текстовой информации, **позволяет** получить набор смысловых паттернов и проводить оценку их семантической близости;

3. Алгоритм поиска знаний в текстах на естественном языке с применением графовых моделей для создания дополнительного фильтра на выходе парсера **позволяет** извлечь смысловую часть предложения из полученной синтаксической схемы текстовой информации для использования в процессах приобретения знаний;

4. Алгоритм приобретения знаний в текстах на естественном языке с применением множества низкоуровневых правил семантического анализа полученных смысловых паттернов **позволяет** определить основные гранулы смысла для процессов использования знаний;

5. Модифицированный биоинспирированный алгоритм использования приобретенных знаний в задачах генеративного искусственного интеллекта с применением улучшенных механизмов интенсификации поиска решений и процедур выхода из локальных оптимумов **позволяет** уменьшить время отклика системы искусственного интеллекта и машинного обучения на пользовательский запрос при обработке и анализе текстов на естественном языке.

Краткое содержание работы

Во введении содержится обоснование актуальности избранной темы и степень ее разработанности, описываются цель, основные задачи, объект и предмет, а также методология и методы проведенных исследований, представлены научная новизна, теоретическая и практическая значимость результатов диссертации, приведены основные положения, выносимые на защиту, указаны сведения о степени достоверности и апробации результатов, внедрении, публикациях автора, объеме и структуре работы.

Первый раздел посвящен анализу исследуемой предметной области. Исследованы основные современные модели, алгоритмы, механизмы и инструменты поиска, приобретения и использования знаний в системах искусственного интеллекта при обработке и анализе текстов на естественном языке.

Технологии обработки и анализа текстов на естественном языке (NLP, Natural Language Processing) предъявляют ряд требований к разработке систем искусственного интеллекта, обеспечивая тем самым конфигурируемость этих

инструментов. Одна из основных задач данного исследования – исключение избыточной информации из синтаксической схемы предложения, получаемой на выходе парсера. В некоторой степени решение данной задачи упрощает применение модулей «извлечения признаков» и «машинного вывода» (рис. 1), что позволяет классифицировать или кластеризовать элементы текстовой информации.

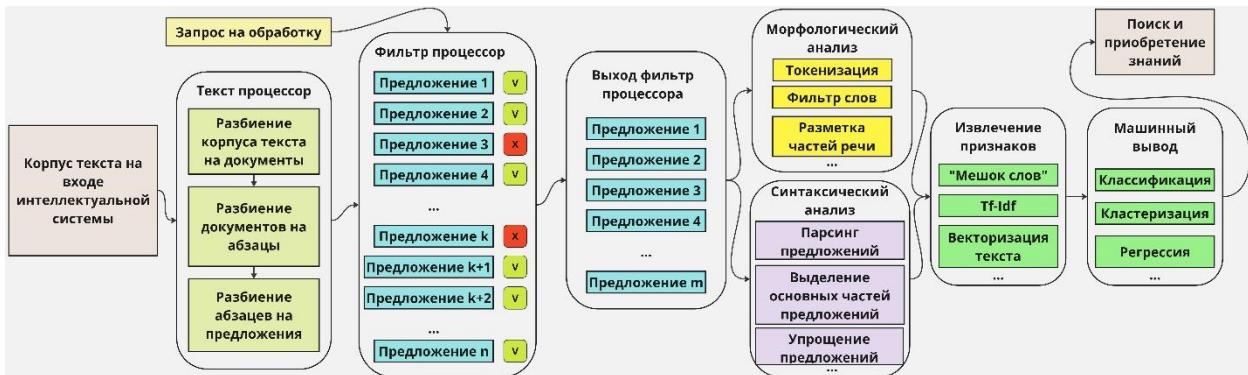


Рисунок 1 – Компонентная архитектура части системы искусственного интеллекта обработки и анализа текстов для этапа предобработки

Утверждение 1. Эффективное решение задачи исключения избыточной информации из текста требует применения низкоуровневых алгоритмов и правил, в том числе на основе использования графовых моделей для создания дополнительного фильтра на выходе парсера, позволяющего извлечь смысловую часть предложения из полученной синтаксической схемы.

Представим задачи поиска, приобретения и использования знаний в виде процесса извлечения знаний (*knowledge extraction*). Данный процесс в самой общей форме представляет собой последовательность следующих этапов, показанных на рисунке 2: поступление корпуса текста на вход системы искусственного интеллекта и машинного обучения; решение задач предобработки текста; применение оператора преобразования полученной после предобработки входных данных текстовой информации.

Формально данное отображение запишем в следующем виде:

$$F_{\text{extraction}}: [\text{Doc}_x, x = 1 \dots T] \rightarrow \text{knowledge}, \quad (1)$$

где $F_{\text{extraction}}$ – оператор преобразования полученной после предобработки входных данных текстовой информации; Doc_x – множество текстовых документов; knowledge – знания, необходимые для решения последующих, иерархически вышестоящих задач, например, для решения задачи информационной поддержки процессов предупреждения и/или ликвидации последствий чрезвычайных ситуаций.

Одним из способов представления полученных знаний является построение модели, описывающей предметную область в виде множества концептов (понятий) ($K_i, i = 1 \dots N$, где N – количество концептов) с заданной на нём системой связей ($R_q, q = 1 \dots M$, где M – количество связей), являющихся по своей сути несимметричными семантическими отношениями. Таким образом, элементом подобной модели является кортеж длины три: $\langle k_i, r_q, k_j \rangle$, $i \neq j$.

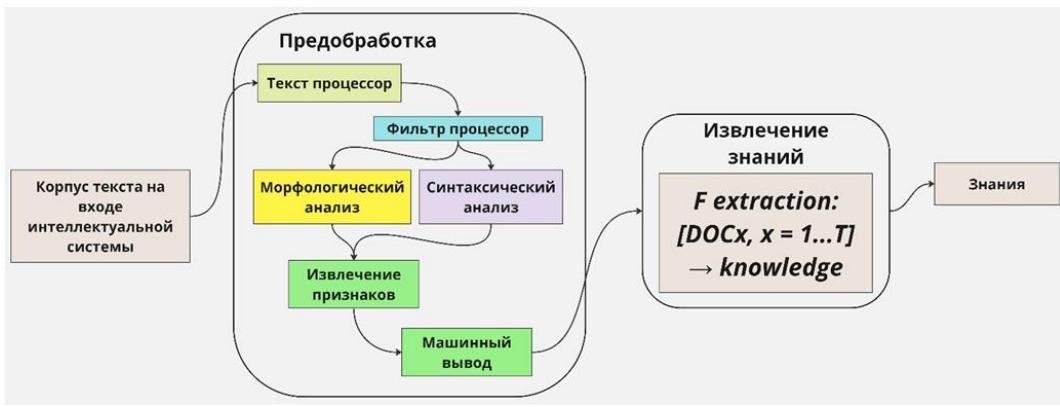


Рисунок 2 – Процесс извлечения знаний в общем виде

Утверждение 2. Комбинированное использование методов извлечения концептов, основанных на низкоуровневых правилах, совместно с методами машинного обучения, повышает качество решения задач поиска, приобретения и использования знаний при обработке и анализе текстов на естественном языке.

Развитие подходов к построению многоуровневых информационных моделей, в том числе онтологических, привело к появлению знаний со сложной внутренней структурой. Таким образом, знания заданы в виде следующей многоуровневой структуры, показанной на рисунке 3.

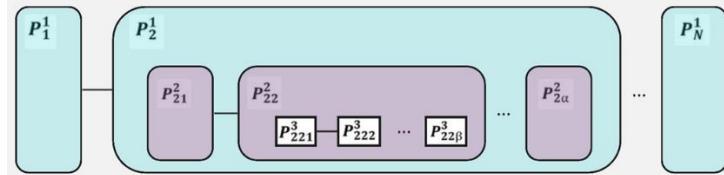


Рисунок 3 – Многоуровневая структура объектов знаний

Между понятиями P разных уровней, представленных на рисунке 3, устанавливаются отношения различного типа. Каждый такой объект знаний имеет разную глубину вложения уровней декомпозиции, определяемую в зависимости от условий решаемой задачи.

$$P_i^\gamma = \{P_{i1}^{\gamma+1}, P_{i2}^{\gamma+1}, \dots, P_{i\alpha}^{\gamma+1}\}; \quad (2)$$

$$P_{ij}^{\gamma+1} = \{P_{ij1}^{\gamma+2}, P_{ij2}^{\gamma+2}, \dots, P_{ij\beta}^{\gamma+2}\}, \quad (3)$$

где P_i^γ – понятие информационной модели знаний под номером i на уровне иерархии γ ; i – номер понятия на уровне γ ($i = 1, \dots, N$); N – количество понятий на уровне γ ; γ – номер уровня иерархии структуры знаний ($\gamma = 1, \dots, L-2$); L – количество уровней иерархии представленной структуры знаний; j – номер понятия на уровне $\gamma+1$ ($j = 1, \dots, \alpha$); α – количество понятий на уровне $\gamma+1$; β – количество понятий на уровне $\gamma+2$.

Обработка такой многоуровневой системы опосредованных отношений между элементами знаний с помощью систем генеративного искусственного интеллекта возможна только при условии грамотно построенной последовательности запросов (промптов). ChatGPT4, например, не может «понимать» смысл обрабатываемого контента, это «понимание» ему дает

специалист (промптер), который через логически построенную схему запроса определяет траекторию принятия решений. Таким образом, извлечение смысла из источников знаний является гораздо более сложным процессом, требующим, помимо сведений о контексте (context) и семантической близости (semantic similarity), учитывать множество других особенностей.

В связи с этим в диссертационной работе решаются задачи поиска, приобретения и использования знаний. Дадим постановки перечисленных задач.

Задача *поиска знаний* в системах искусственного интеллекта при обработке и анализе текстов на естественном языке в большей степени связана с качественным построением синтаксической схемы текста на основе применения парсера.

На основе полученной синтаксической схемы строится ориентированный граф $G = (V, E)$, где V – множество вершин, а E – множество ребер. Оценка или важность вершины определяется как:

$$S(V_i) = (1 - d) + d * \sum_{j \in In(V_i)} \frac{1}{|Out(V_j)|} S(V_j), \quad (4)$$

где $In(V_i)$ – это набор вершин, которые указывают на V_i , а $Out(V_i)$ – это набор вершин, на которые указывает V_i . При этом d – это коэффициент затухания, который устанавливается в диапазоне от 0 до 1.

Задача *приобретения знаний* сводится к оценке семантической близости понятий (концептов), позволяющей распределить их по классам для дальнейшего построения онтологии предметной области. Дадим постановку задачи *классификации знаний* на основе использования признакового описания исследуемых объектов знаний в виде векторного представления. Предположим, что имеется N объектов знаний, M признаков, а также K наименований классов. Рассмотрим объект знаний \vec{X}_n и класс \vec{Y}_k . Определим расстояние между рассматриваемым объектом и классом:

$$\|X_n - Y_k\| = \sqrt{(x_{1,n} - y_{1,k})^2 + (x_{2,n} - y_{2,k})^2 + \dots + (x_{M,n} - y_{M,k})^2}. \quad (5)$$

Введём пороговое значение Δ максимального расстояния, превышение которого исключает возможность отнесения объекта \vec{X}_n к классу \vec{Y}_k . В этом случае оптимизационная задача классификации примет следующий вид:

$$Y^* = \arg \min_k \|X_n - Y_k\|; \quad (6)$$

$$\begin{cases} \|X_n - Y^*\| \leq \Delta, & Y^* \text{ – определен для } X_n \\ \text{otherwise}, & X_n \text{ – не классифицирован} \end{cases} \quad (7)$$

Эффективное *использование знаний* при обработке и анализе текстов на естественном языке предполагает построение системы опосредованных отношений между гетерогенными элементами знаний, поэтому актуальной задачей является *интеграция* онтологий различных предметных областей. Дадим постановку задачи *интеграции знаний* множества онтологий:

$$O^\cup = \bigcup_i O_i, i = \overline{1, N}, \quad (8)$$

где O_i – онтограф i -той предметной области; i – номер рассматриваемой предметной области; N – общее число предметных областей.

Выполнение поставленных задач обеспечивает построение упорядоченной структуры извлекаемых знаний в информационном пространстве поиска решений на основе детерминированных моделей и алгоритмов, что позволяет в итоге минимизировать время отклика системы искусственного интеллекта на запрос пользователя ($T_{\text{response}} \rightarrow \min$).

Второй раздел диссертационной работы посвящен построению верхнеуровневой и нижеуровневой моделей онтологии знаний, применяемых при обработке и анализе текстов на естественном языке. Верхнеуровневая модель онтологии знаний отличается применением оригинальной компонентной архитектуры (рис. 4), которая позволяет обеспечить необходимую степень детализации анализируемой текстовой информации.

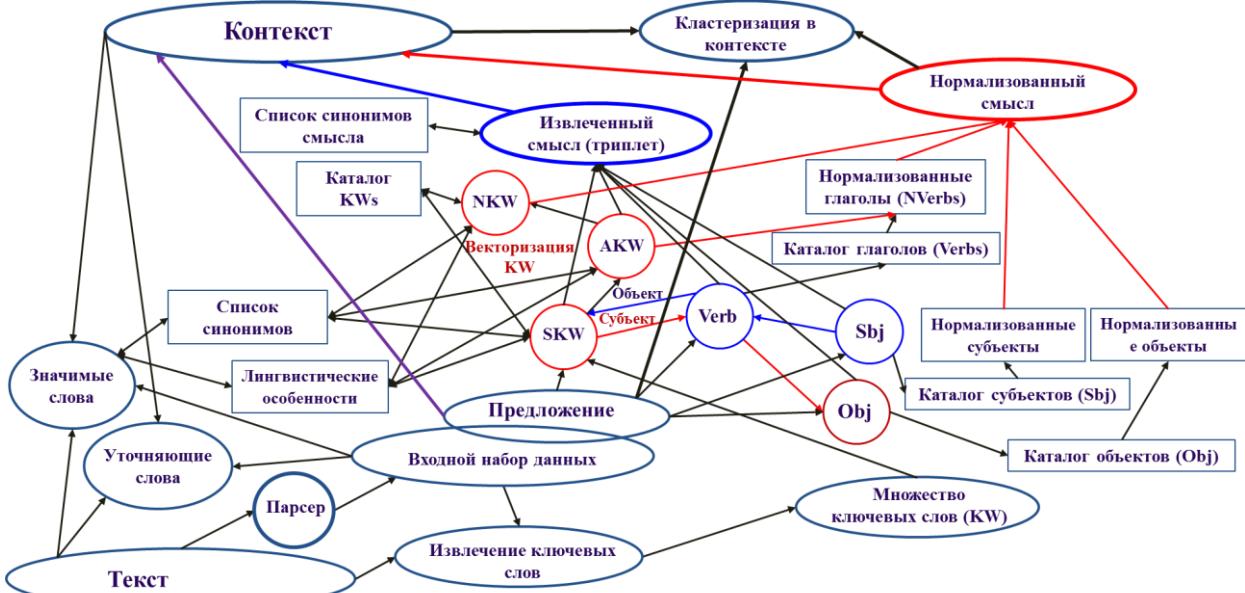


Рисунок 4 – Верхнеуровневая модель онтологии знаний для интеллектуальных систем обработки и анализа текстов

Имеющиеся в онтологии субъекты (Sbj) и объекты (Obj) составляют множество концептов (понятий) ($K_i, i = 1\dots N$, где N – количество концептов) с заданной на нём системой связей (бинарных отношений) ($R_q, q = 1\dots M$, где M – количество связей), являющихся по своей сути несимметричными семантическими отношениями.

Конечное множество лексических меток (словарь онтологии) представлено множеством L (9):

$$L = \{L_1, \dots, L_u\}. \quad (9)$$

Выражением (10) представлено антисимметричное, транзитивное, нерефлексивное бинарное отношение, являющееся отношением частичного порядка на множестве понятий (концептов) K :

$$Rel_K \subseteq K \times K, Rel_K \in K. \quad (10)$$

Бинарное отношение инцидентности между множествами L и K задано следующим выражением (11):

$$Rel_{LK} \subseteq L \times K. \quad (11)$$

Бинарное отношение инцидентности между множествами L и R задано следующим выражением (12):

$$Rel_{LR} \subseteq L \times R. \quad (12)$$

Формализованной моделью онтологии верхнего уровня описания является следующий кортеж (13):

$$O = \langle K, R, L, I, Sent, KW, Word, Verb, Rel, Syn, A \rangle, \quad (13)$$

где I – множество экземпляров понятий; $KW = \{KWs, SKWs, AKWs, NKWs\}$ – интегрированное множество поисковых, актуальных и нормализованных ключевых слов; $Word = \{MWs, CWs\}$ – интегрированное множество значимых и уточняющих слов в онтологии; $Verb = \{Verbs, NVerbs\}$ – интегрированное множество оригинальных и нормализованных глаголов в онтологии; $Rel = \{Rel_k, Rel_{LK}, Rel_{LR}, Rel_{SentKWs}, Rel_{SentMWs}, Rel_{SentCWs}\}$ – интегрированное множество бинарных отношений между элементами; $Syn = \{Syn_{SKWSAKWs}, Syn_{SKWSNKWs}, Syn_{AKWSNKWs}, Syn_{AKWSAKWs}, Syn_{MWSMWs}, Syn_{VerbsNVerbs}\}$ – интегрированное множество бинарных отношений синонимии между элементами; A – аксиомы онтологии.

Предложенная нижеуровневая модель онтологии знаний отличается использованием структуры отношений между понятиями, детализирующими семантику текстовой информации, что позволяет получить набор смысловых паттернов, а также проводить оценку их семантической близости (рис. 5).

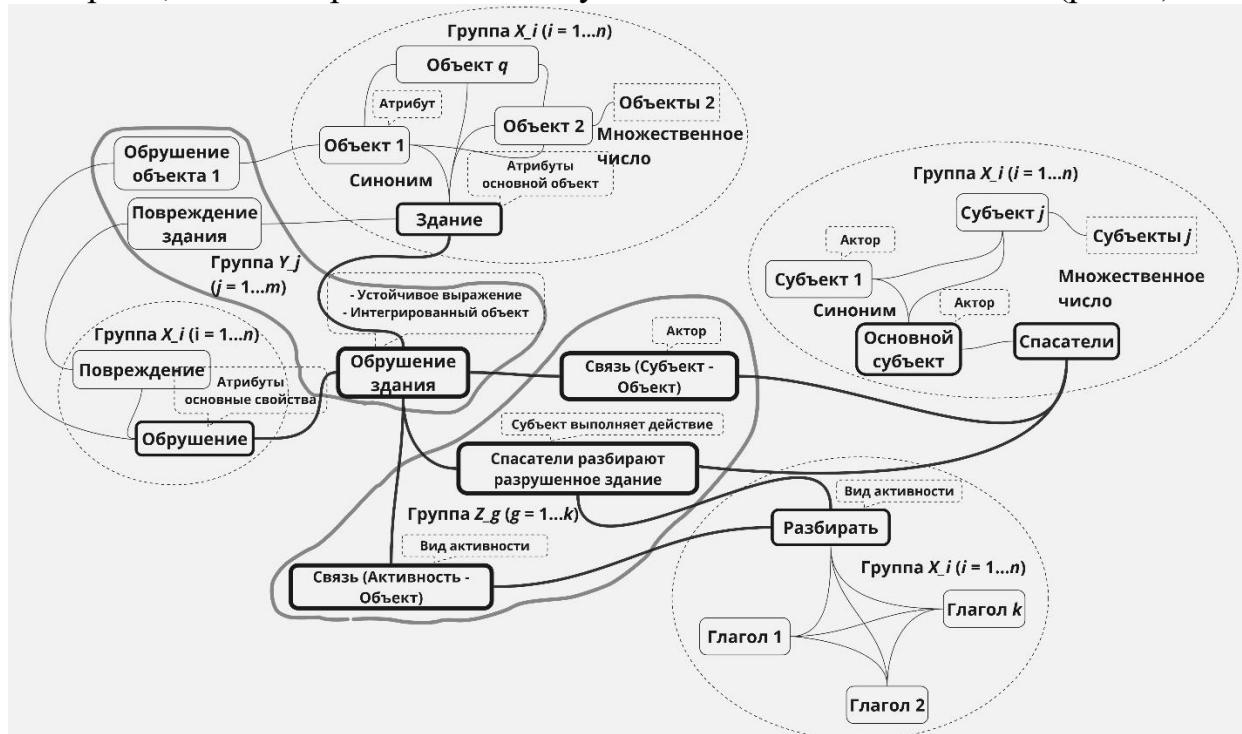


Рисунок 5 – Нижнеуровневая модель онтологии знаний для паттерна информационного пространства при мониторинге чрезвычайных ситуаций

Элементам множества понятий (*концептов*) K ставится в соответствие набор векторов, значения компонентов которых определяются их атрибутами. Веса понятий определяются как на основе экспертных оценок, так и частотных характеристик появления в информационных ресурсах исследуемых предметных областей, а также контекста исследуемого информационного процесса. Веса отношений характеризуют их важность для определения тематики информационного процесса. Критерием оценки принадлежности определенному классу является аргумент минимизации расстояния между

элементами информации в векторном пространстве, как это показано в выражении (6).

Третий раздел посвящен разработке комплекса алгоритмов поиска, приобретения и использования знаний при обработке и анализе текстов. Предложенный автором алгоритм поиска знаний опишем на основе обработки и анализа следующего предложения: «Нашей миссией является задача, помочь вам ориентироваться в постоянно растущем множестве правил безопасности». Для построения графовой модели данного предложения был выбран ациклический ориентированный граф (рис. 6).

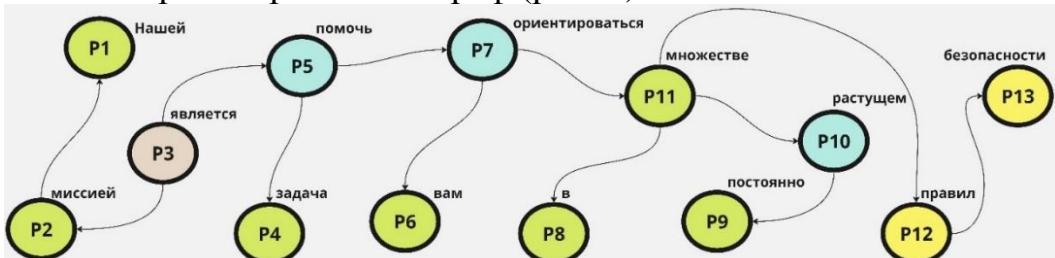


Рисунок 6 – Ациклический ориентированный граф, моделирующий систему отношений между словами анализируемого предложения

Используя информационный граф структуры текстового документа, содержащий верхнеуровневое описание контекста, были восстановлены субъект и объект – «МЧС России» и «население» соответственно. Значение предиката для ключевой фразы в вершине P_7 «ориентироваться» изменяется на определенное выше нормализованное значение глагола «узнать».

Все «тупиковые» вершины (P_4 , P_8 , P_9), не замененные ранее и/или не получившие собственных имен, удаляются из графа. «Висячая» корневая вершина P_3 , имеющая только выходные ребра, является последовательным отношением между субъектом «МЧС России» и предикатом «помогает». При упрощении структуры предложения подобная вершина заменяется ребром. На выходе разработанного алгоритма поиска знаний граф анализируемого предложения приобретает вид, представленный на рисунке 7.

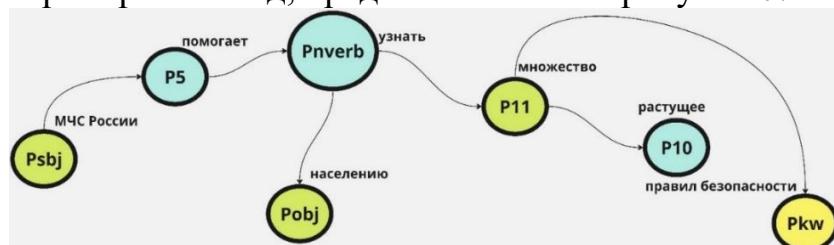


Рисунок 7 – Упрощенный ациклический ориентированный граф, моделирующий систему отношений между элементами и компонентами смысла

Поиск «изолированных», «висячих» и «тупиковых» вершин происходит на основе анализа матрицы смежности графа $A = \|a_{ij}\|$, где для всех вершин k ($k = \overline{1, n}$, n – общее число вершин в графе) находится вектор $a(k) = (a_k, a^k)$ со следующими компонентами:

$$a_k = \sum_{j=1}^n a_{kj}, a^k = \sum_{i=1}^n a_{ik}, \quad (14)$$

где a_k – сумма элементов k -ой строки матрицы смежности, a^k – сумма элементов k -го столбца матрицы смежности.

Оставшиеся инвариантные к смыслу вершины P_{10} «растущее» и P_{11} «множество» исключаются из рассмотрения разработанным автором алгоритмом приобретения знаний в текстах на естественном языке с применением множества низкоуровневых правил семантического анализа уже полученных смысловых паттернов, позволяющим построить основные гранулы смысла. На рисунке 8 представлена его укрупненная схема.

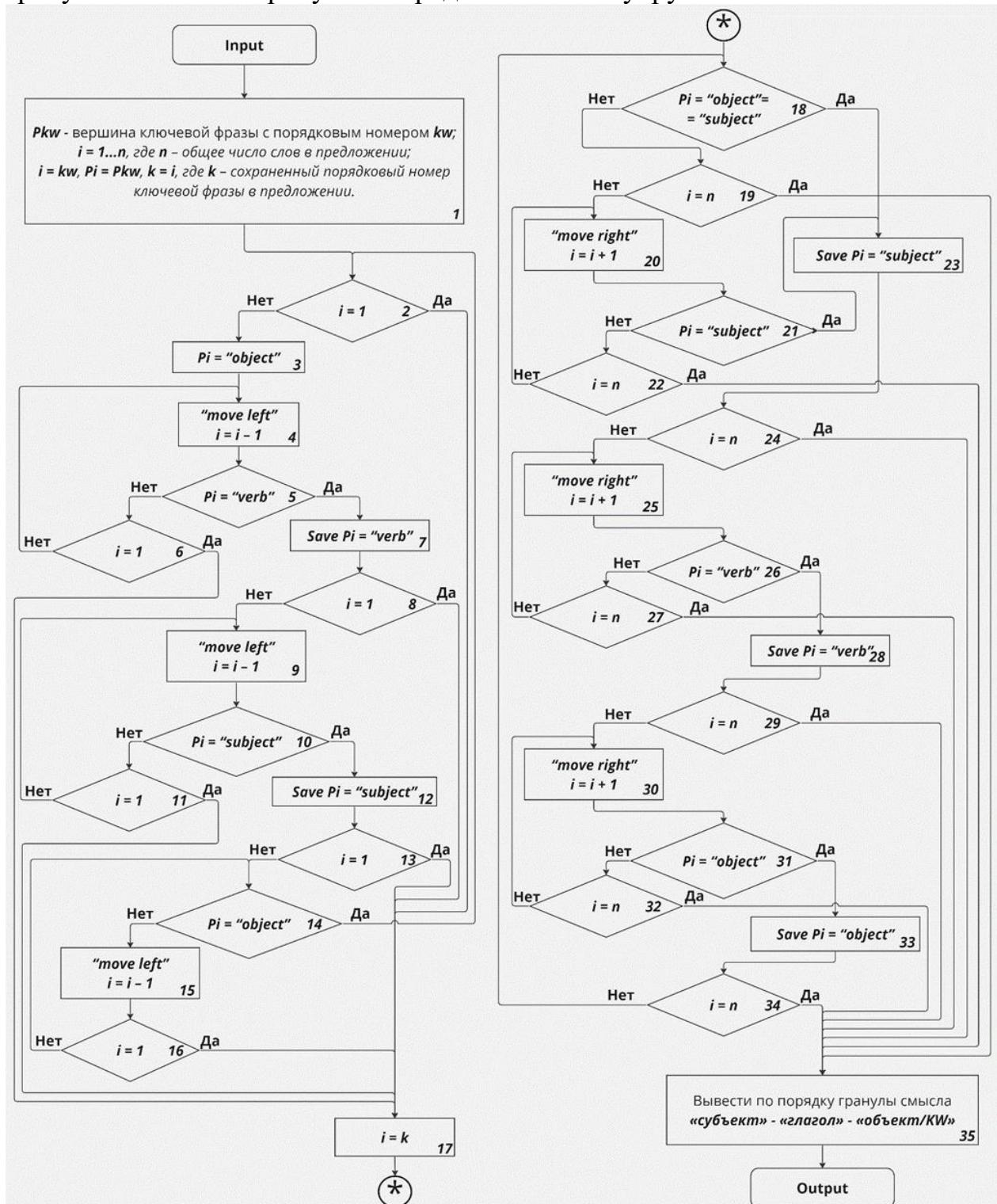


Рисунок 8 – Укрупненная схема алгоритма приобретения знаний в текстах на естественном языке с применением множества низкоуровневых правил

Разработанные алгоритмы поиска и приобретения знаний используются для наполнения смысловыми паттернами онтологической структуры, которая является базой прецедентов и применяется в качестве информационного пространства для поиска решений при обработке пользовательского запроса на основе биоинспирированного алгоритма использования знаний. В качестве биоинспирированного алгоритма использования знаний автор предложил модификацию алгоритма бактериальной оптимизации.

Ранее для интенсификации поиска при бактериальной оптимизации производилась настройка показателя скорости передвижения агента-бактерии в пространстве поиска решений. Данная особенность алгоритма повышала эффективность биоинспирированного поиска на множестве необработанной предварительно текстовой информации. В разработанном автором диссертации модифицированном алгоритме бактериальной оптимизации (МАБО) нет необходимости в подобной постоянной интенсификации процедур поиска решений, так как предложенные в диссертации эвристические алгоритмы поиска и приобретения знаний позволяют строить онтологическую структуру, содержащую только отфильтрованные гранулы смысла в паттернах прецедентов. Поэтому при модификации алгоритма бактериальной оптимизации автор использовал для интенсификации поиска решений механизм «локального прыжка», взятый из одной разновидности алгоритма обезьяньего поиска (рис. 9).

Данный механизм, помимо поисковых процедур, также позволяет диверсифицировать пространство поиска, исключив траекторию передвижения агента-бактерии, например, по ребрам онтографа через вершины $P_6 - P_7 - P_8$, тем самым снижая временные затраты на обработку и анализ используемых в информационной модели знаний.

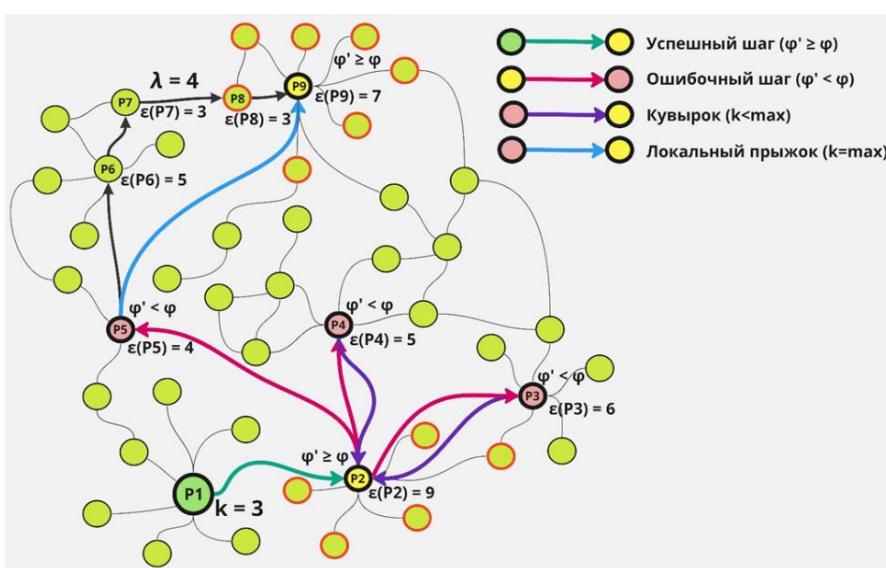


Рисунок 9 – Иллюстрация «локального прыжка»
бактерии

В случае попадания агента-бактерии в локальный оптимум, что характеризуется локализацией показателя уровня текущего здоровья h_i бактерии s_i в средней части допустимых значений. Вычисление текущего уровня здоровья происходит на основе канонического выражения (15), в котором данный параметр определяется суммой значений целе-

вой функции, рассчитанной во всех точках пройденной траектории перемещений:

$$h_i = \sum_{\tau=1}^T \varphi_{i,r,l}(\tau), i \in [1, |S|]. \quad (15)$$

Для достижения баланса между скоростью сходимости и диверсификацией пространства поиска в разработанном модифицированном алгоритме бактериальной оптимизации реализованы механизмы «репродукции», позволяющей интенсифицировать протекание процесса поиска в локальных областях, а также – «ликвидации и рассеивания», позволяющих агентам-бактериям покинуть локальные оптимумы.

Четвертый раздел посвящен описанию разработки программного приложения, а также подготовки и проведения вычислительного эксперимента для оценки эффективности созданных моделей и алгоритмов.

Для реализации предложенных автором моделей и алгоритмов поиска, приобретения и использования знаний при обработке и анализе текстов на естественном языке разработано программное приложение, компонентная архитектура которого представлена на рисунке 10.

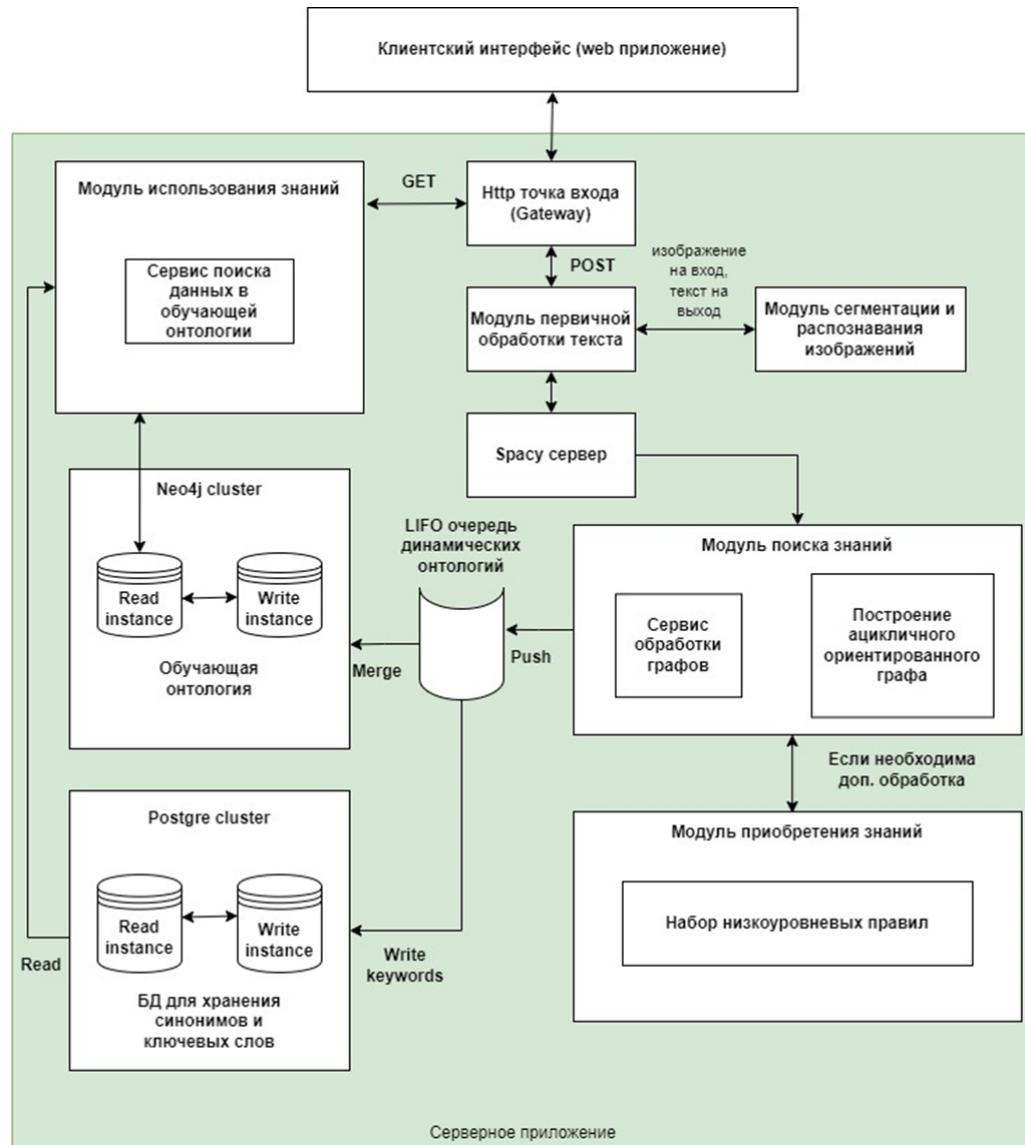


Рисунок 10 – Компонентная архитектура программного приложения

С применением созданного программного приложения был проведен вычислительный эксперимент, подтверждающий эффективность предложенных в работе моделей и алгоритмов поиска, приобретения и использования знаний в процессе генерации ответа на запрос пользователя системой искусственного интеллекта и машинного обучения при обработке и анализе текстов на естественном языке.

Учитывая специфику решаемой задачи поддержки принятия решений по предупреждению и ликвидации последствий ЧС, для обеспечения необходимой оперативности работы приложения критерием оценки эффективности было выбрано время отклика системы на запрос пользователя с определением соответствующих ограничений по качеству принимаемых решений. Для сравнительного анализа были выбраны известные алгоритмы роя частиц (АРЧ), обезьяньего поиска (АОП) и поиска кукушки (АПК). В среднем при значительном размере пространства поиска решений (более 1 миллиона вершин) предложенный автором модифицированный алгоритм бактериальной оптимизации снижает время отклика системы искусственного интеллекта на 5 – 7 %.

Временная сложность всех рассмотренных биоинспирированных алгоритмов является полиномиальной и в худшем случае составляет $O(n^2)$. Диаграмма сравнения временных характеристик представлена на рисунке 11.

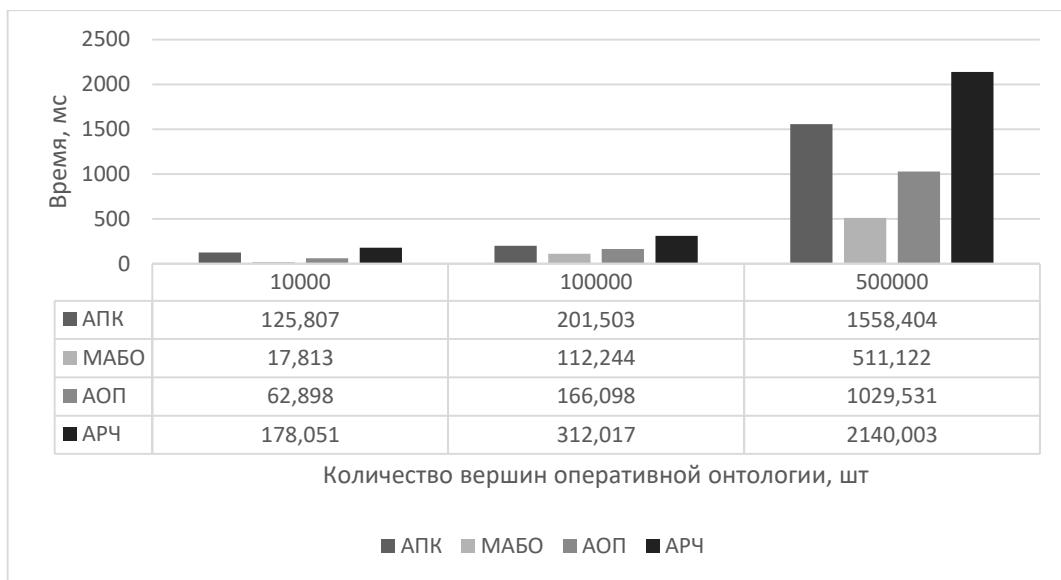


Рисунок 11 – Диаграмма сравнения времени работы исследуемых алгоритмов

Как видно из представленной на рисунке 11 диаграммы, разработанный автором модифицированный алгоритм бактериальной оптимизации (МАБО) превосходит остальные исследуемые алгоритмы по скорости работы. Высокая скорость работы предложенного алгоритма обусловлена применением более эффективных механизмов интенсификации поисковых процедур и диверсификации пространства поиска решений, что позволяет проводить оценку семантической близости прецедентов и анализируемых знаний с уменьшением времени генерации ответа на запрос пользователя системой искусственного интеллекта и машинного обучения при обработке и анализе текстов на естественном языке.

Описанный вычислительный эксперимент подтверждает, что предложенный автором модифицированный алгоритм бактериальной

оптимизации получает набор квазиоптимальных решений за полиномиальное время, это указывает на необходимость продолжения исследований применимости биоинспирированных алгоритмов для решения задач поиска, приобретения и использования знаний в процессе генерации ответа на запрос пользователя системой искусственного интеллекта и машинного обучения при обработке и анализе текстов на естественном языке.

В заключении изложены итоги выполненного исследования, рекомендации и перспективы дальнейшей разработки темы.

В приложениях приведены свидетельства об официальной регистрации программ для ЭВМ и копии актов внедрения результатов работы.

ЗАКЛЮЧЕНИЕ

Основной целью выполненной диссертационной работы являлось повышение эффективности моделей и алгоритмов поиска, приобретения и использования знаний в системах искусственного интеллекта при обработке и анализе текстов. Под эффективностью понимается минимизация времени отклика системы на запрос пользователя при условии обеспечения «прозрачности» процессов обработки входной текстовой информации.

Созданные в представленном исследовании модели и алгоритмы поиска, приобретения и использования знаний указывают на перспективность дальнейшей разработки выбранной темы в направлении развития подходов искусственного интеллекта для построения гибких механизмов обработки и анализа текстов на естественном языке с учетом семантики и контекста информационных ресурсов.

Результаты диссертации имеют важное значение для развития технологий искусственного интеллекта, получили высокую оценку научного сообщества при аprobации и положительные рекомендации для внедрения в информационные процессы предприятий, учреждений и организаций различного профиля деятельности.

Основными результатами проведенных исследований являются следующие:

1. Проведен аналитический обзор особенностей создания систем искусственного интеллекта и машинного обучения для обработки и анализа текстов на естественном языке. Сформулированы постановки основных задач исследования;

2. Построена верхнеуровневая модель онтологии знаний, отличающаяся включением в состав ее компонентов множеств понятий с различным уровнем нормализации, что позволяет обеспечить необходимую степень детализации анализируемой текстовой информации;

3. Построена нижнеуровневая модель онтологии знаний, отличающаяся использованием структуры отношений между понятиями, детализирующими семантику текстовой информации, что позволяет получить набор смысловых паттернов, а также проводить оценку их семантической близости;

4. Разработан алгоритм поиска знаний в текстах на естественном языке, отличающийся созданием дополнительного фильтра на выходе парсера с применением графовых моделей, что позволяет извлечь смысловую часть предложения из полученной синтаксической схемы текстовой информации для использования в процессах приобретения знаний;

5. Разработан алгоритм приобретения знаний в текстах на естественном языке, отличающийся применением множества низкоуровневых правил семантического анализа полученных смысловых паттернов, позволяющий определить основные гранулы смысла для процессов использования знаний;

6. Разработан модифицированный биоинспирированный алгоритм использования приобретенных знаний в задачах генеративного искусственного интеллекта, отличающийся улучшенными механизмами интенсификации поиска решений и процедурами выхода из локальных оптимумов, что позволило уменьшить время отклика системы искусственного интеллекта и машинного обучения на пользовательский запрос при обработке и анализе текстов на естественном языке;

7. Разработано программное приложение, реализующее разработанные алгоритмы поиска, приобретения и использования знаний в системах искусственного интеллекта и машинного обучения при обработке и анализе текстов на естественном языке, позволяющее проводить сравнительный анализ предложенных моделей и алгоритмов с существующими аналогами;

8. Выполнен вычислительный эксперимент, который подтвердил эффективность полученных решений. В среднем при значительном размере пространства поиска решений (более 1 миллиона вершин) предложенный автором модифицированный алгоритм бактериальной оптимизации снижает время отклика системы искусственного интеллекта на 5 – 7 %.

В целом, предложенные в работе модели и алгоритмы поиска, приобретения и использования знаний находятся на лидирующих позициях. Перечисленные результаты получили высокую оценку научного сообщества при апробации и положительные рекомендации для внедрения.

Таким образом, при выполнении диссертационного исследования решены все поставленные задачи, цель работы достигнута. Построены модели и разработаны алгоритмы поиска, приобретения и использования знаний при обработке и анализе текстов, которые направлены на решение актуальной научной задачи, связанной с необходимостью повышения эффективности функционирования систем генеративного искусственного интеллекта, где под эффективностью понимается минимизация времени отклика системы на запрос пользователя при условии обеспечения «прозрачности» процессов обработки входной текстовой информации, что имеет важное значение для развития информатики.

Дальнейшим направлением исследований является масштабирование предложенных в данной работе онтологических моделей отдельных предметных областей в общую «Модель Мира», что позволит затем перейти к разработке универсальных алгоритмов и методов искусственного интеллекта для поиска, приобретения и использования знаний в глобальном информационном пространстве.

ПУБЛИКАЦИИ ПО ТЕМЕ ДИССЕРТАЦИОННОЙ РАБОТЫ

Публикации в научных изданиях, индексируемых международными базами данных, перечень которых определен в соответствии с рекомендациями ВАК РФ, опубликовано 5 работ:

1. Kravchenko, D.Y. Architecture and Method of Integrating Information and Knowledge on the Basis of the Ontological Structure / D.Y. Kravchenko, Y.A. Kravchenko, I.O. Kursitys // Advances in Intelligent Systems and Computing. – 2018. – Vol. 658. – P. 93-103. (Scopus, Q3)

2. Kravchenko, D.Y. Ontological Approach for Designing a Multi-agent Behavior Model in the Internet Environment / D.Y. Kravchenko, Yu.A. Kravchenko, I.O. Kursitys // Journal of Physics: Conference Series. – 2019. – Vol. 1333(3): 032043. (Scopus, Q4)

3. Kravchenko, D.Y. Harnessing key phrases in constructing a concept-based semantic representation of text using clustering techniques / D.Y. Kravchenko, Ali Mansour, Juman Mohammad, Nemury Silega, Yury Kravchenko // Lecture Notes in Computer Science. – 2023. – Vol. 14335. – P. 190-201. (Scopus, Q2)

4. Kravchenko, D.Y. The Swarm Bacterial Algorithm Based on New Attractive Operators and Patterns of Agent Behavior / D.Y. Kravchenko, Yu.A. Kravchenko, E.V. Kuliev, S.I. Rodzin, L.S. Rodzina // Lecture Notes in Networks and Systems. – 2024. – Vol. 934. – Paper ID: 202183. – P. 147-168. (Scopus, Q4)

5. Kravchenko, D. Algorithm for Optimization of Keyword Extraction Based on the Application of a Linguistic Parser / D. Kravchenko, Yu. Kravchenko, A. Mansour, J. Mohammad, N. Pavlov // Informatics and Automation. – 2024. – Vol. 23. – No 2. – P. 467-494. (Scopus, Q4)

Публикации в издании из перечня, утвержденного ВАК РФ, рецензируемых научных изданий, в которых должны быть опубликованы основные научные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук, опубликовано 4 работы:

6. Кравченко, Д.Ю. Семантический поиск с использованием генетических операторов / Д.Ю. Кравченко, В.В. Марков, А.А. Новиков, Ю.С. Старкова // Известия ЮФУ. Технические науки. – 2017. – № 7 (192). – С. 122-133. (ВАК, К2)

7. Кравченко, Д.Ю. Гибридный биоинспирированный алгоритм отображения онтологий в задачах извлечения и управления знаниями / Д.Ю. Кравченко, Ю.А. Кравченко, В.В. Марков // Известия ЮФУ. Технические науки. – Таганрог: Изд-во ЮФУ. – 2020. – № 2 (212). – С. 16-28. (ВАК, К2)

8. Кравченко, Д.Ю. Поддержка принятия решений по предупреждению и ликвидации последствий чрезвычайных ситуаций на основе нечеткого метода структурирования информации / Д.Ю. Кравченко, Е.М. Герасименко, Ю.А. Кравченко, Э.В. Кулиев // Известия ЮФУ. Технические науки. – Таганрог: Изд-во ЮФУ. – 2023. – № 2 (232). – С. 201-212. (ВАК, К2)

9. Кравченко, Д.Ю. Модель онтологии знаний для интеллектуальных систем обработки и анализа текстов / Д.Ю. Кравченко // Известия ЮФУ. Технические науки. – 2024. – № 2 (238). – С. 38-50. (ВАК, К2)

Публикации в других изданиях, опубликовано 10 работ:

10. Кравченко, Д.Ю. Модель онтологии электронного учебного курса для систем дистанционного обучения / Д.Ю. Кравченко, Д.В. Лещанов, Ю.Ю. Запорожец // Студенческая наука для развития информационного общества: сборник материалов V Всероссийской научно-технической конференции. – Ставрополь: Изд-во СКФУ. – 2016. – С. 352-356.

11. Кравченко, Д.Ю. Перспективы интеграции технологии коллективного семантического поиска и формата rdf / Д.Ю. Кравченко, А.А. Новиков, В.А. Козачков // 1-я Международная научно-практическая конференция «Программная инженерия: методы и технологии разработки информационно-вычислительных систем» (ПИИВС-2016). Сборник научных трудов. ГОУ ВПО «Донецкий национальный технический университет». – Донецк: 2016. – С. 149-152.

12. Кравченко, Д.Ю. Структуризация информации на основе комбинации генетического, роевого и обезьяньего алгоритмов / Д.Ю. Кравченко, Н.В. Кулиева, Ю.С. Новикова, М.И. Анчеков // Российская академия наук. Научный журнал. Известия КБНЦ РАН. – Нальчик: Изд-во КБНЦ РАН. – 2019. – № 5(91). – С.5-13.

13. Кравченко, Д.Ю. Математическое описание процесса поддержки принятия решений при оценке семантической близости знаний в конкретизированной модели онтологии / Д.Ю. Кравченко, Ю.А. Кравченко, В.В. Курейчик, В.В. Марков // Современные компьютерные технологии: материалы II научно-методической конференции НПР. – Таганрог: Изд-во ЮФУ. – 2021. – С. 25-28.

14. Кравченко, Д.Ю. Метод автоматического извлечения ключевых слов / Д.Ю. Кравченко, Ю.А. Кравченко, А.М. Мансур, Ж.Х. Мохаммад // Труды международного научно-технического конгресса «Интеллектуальные системы и информационные технологии – 2022». Научное издание. – Таганрог: Изд-во Ступина С.А., Т.1. – 2022. – С. 90-97.

15. Кравченко, Д.Ю. Модифицированный метод построения семантического представления текста на основе методов кластеризации и взвешивания терминов / Д.Ю. Кравченко, Ю.А. Кравченко, А.М. Мансур, Ж.Х. Мохаммад // Труды XII международной научно-технической конференции «Технологии разработки информационных систем (ТРИС-2022)». – Таганрог: 2022. – С. 94-100.

16. Кравченко, Д.Ю. Модифицированный биоинспирированный метод поддержки принятия решений по предупреждению и ликвидации последствий чрезвычайных ситуаций / Д.Ю. Кравченко, Е.М. Герасименко, В.В. Курейчик, Э.В. Кулиев, Ю.А. Кравченко, С.И. Родзин // Информационные технологии. – 2023. – Т. 29. – № 8. – С. 423-436. (RSCI, K1)

17. Кравченко, Д.Ю. Алгоритм идентификации предпосылок возникновения чрезвычайных ситуаций на основе правил «если-то» / Д.Ю. Кравченко, Ю.А. Кравченко, Э.В. Кулиев, С.И. Родзин // Научный журнал «Информатизация и связь», № 2, 2023. – С. 11-17. (ВАК, К2)

18. Кравченко, Д.Ю. Баланс между скоростью сходимости биоэвристики и диверсификацией пространства поиска решений (на примере модели роя саранчи) / Д.Ю. Кравченко, О.Н. Родзина // Труды международного научно-технического конгресса «Интеллектуальные системы и информационные технологии – 2023». Научное издание. – Таганрог: Изд-во Ступина С.А. – 2023. – Т.1. – С. 208-217.

19. Kravchenko, Daniil Computational Model of Swarm Algorithm for Optimizing Process of Keywords Extraction from Text Information Presented as Graph / Daniil Kravchenko, Sergey Rodzin, Nikolay Pavlov, Lada Rodzina, Elmar Kuliev, Yury Kravchenko // 7th International Conference on Information Technologies in Engineering Education, Inforino. – IEEE. – 2024, Page(s): 1 – 10. (Scopus)

Свидетельства о государственной регистрации программ для ЭВМ:

1. Кравченко, Д.Ю. Программный модуль сегментации изображений на основе модели «хищники – травоядные» / Д.Ю. Кравченко, Д.Ю. Запорожец, В.В. Курейчик, С.И. Родзин // Свидетельство регистрации программы для ЭВМ. – 17.10.2023. – № 2023681642.

2. Кравченко, Д.Ю. Программный модуль оптимизации работы классификатора при векторизации текста на основе биоэвристик / Д.Ю. Кравченко, Ю.А. Кравченко, А. Мансур, М. Жуман // Свидетельство регистрации программы для ЭВМ. – 12.12.2023. – № 2023687185.

Личный вклад автора.

В опубликованных трудах:

[1, 2, 10] – автором построены онтологические модели для решения задач поиска, приобретения и использования знаний;

[3, 5, 14, 19] – автором разработаны методы и алгоритмы извлечения ключевых слов при обработке и анализе текстов на естественном языке;

[4, 7, 8, 12, 16-18] – автором разработаны биоинспирированные и нечеткие методы и алгоритмы для решения задач обработки и анализа текстов на естественном языке, в том числе при поддержке принятия решений по предупреждению и ликвидации последствий чрезвычайных ситуаций;

[6, 11, 13, 15] – автором разработаны методы и алгоритмы семантического поиска знаний при обработке и анализе текстов на естественном языке.

При создании указанных программ для ЭВМ, в том числе использованы теоретические результаты, полученные лично автором.

Автореферат

Подписано в печать «16» января 2025 г.

Бумага офсетная. Печать офсетная. Формат 60 × 84 1/16.

Усл. печ. л. 1,375. Уч.-изд. л. 1,1. Заказ № 17. Тираж 100 экз.

Отпечатано в отделе полиграфической, корпоративной и сувенирной продукции
Издательско-полиграфического комплекса КИБИ МЕДИА ЦЕНТРА ЮФУ.

344090, г. Ростов-на-Дону, пр. Ставки, 200/1, тел. 8 (863) 243-41-66