Федеральное государственное автономное образовательное учреждение высшего образования
«Национальный исследовательский ядерный университет «МИФИ»
(НИЯУ МИФИ)

На правах рукописи

АРТАМОНОВ Алексей Анатольевич

Модели, методы и технологии интеллектуального анализа информационных объектов в научно-технических и социально значимых задачах

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

Диссертация на соискание ученой степени
доктора технических наук

Научный консультант
д.т.н, с.н.с. Кореньков Владимир Васильевич

Москва – 2025

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Актуальность работы

В современных условиях стремительного роста объёмов информации и развития технологий искусственного интеллекта необходима разработка новых подходов к интеграции и анализу данных для поддержки принятия решений в научно-технической и социальной сферах. В различных предметных областях накапливаются огромные массивы разнородных данных: научные публикации, патенты, отчёты, а также цифровые следы человеческой деятельности в социальных сетях и других онлайн-сервисах. Интеллектуальный анализ данных (ИАД) стал одним из ключевых инструментов для извлечения знаний из таких массивов. Он сочетает методы машинного обучения, статистического анализа, обработки неструктурированных данных и визуальной аналитики, что открывает возможности выявления скрытых закономерностей и принятия более обоснованных решений на основе имеющихся данных. Классические подходы зачастую не справляются с актуальными требованиями: данные поступают непрерывно и в разнообразных форматах, содержащаяся в них информация слабо структурирована, а существующие системы нередко ориентированы на заданные типы данных. Это приводит к необходимости разработки новых методологических принципов и комплексных инструментов для интеллектуального анализа информации.

В научно-технической сфере задача анализа данных важна для отслеживания развития научных направлений, выявления технических трендов и управления знаниями. В социальной сфере интеллектуальный анализ цифровых профилей позволяет решать социально значимые задачи, например, раннее обнаружение групп риска (по поведенческим индикаторам в соцсетях) или противодействие информационным угрозам. Однако междисциплинарный характер подобных задач выявляет ряд проблем.

Во-первых, необходимо обеспечить единый подход к описанию и хранению разнородных данных, чтобы можно было интегрировать

информацию из разных источников (научные статьи, посты в соцсетях, базы данных и т.д.) в единую систему.

Во-вторых, требуется поддерживать высокое качество и достоверность данных: без очистки, нормализации и верификации исходной информации выводы анализа могут оказаться неверными.

В-третьих, сложность алгоритмов ИАД порождает проблему интерпретируемости и доверия – пользователям и экспертам важно понимать, на каких фактах основаны те или иные прогнозы или рекомендации.

В-четвертых, при работе с социальными данными возникают дополнительные ограничения, связанные с конфиденциальностью и этикой использования персональной информации.

Актуальность работы обусловлена потребностью в новых моделях, методах и технологиях, обеспечивающих эффективный интеллектуальный анализ информационных объектов различной природы, учитывая перечисленные вызовы. Разработка интегративной системы интеллектуального анализа данных позволит систематизировать и объединять неструктурированные данные из научно-технической и социальной областей, извлекать из них новые знания с высокой степенью достоверности, а также автоматизировать значительную часть трудоёмких аналитических процессов, что будет способствовать повышению качества управления научными исследованиями и социальными проектами, раннему выявлению важных тенденций и решению практических задач в интересах науки, экономики и общества. Фактологически подтверждается, что объём и разнообразие данных будут только расти, а потому создание интеллектуальных инструментов для их анализа является актуальной и стратегически значимой научной задачей.

Степень научной разработанности темы исследования. Фундаментальные основы интеллектуального анализа данных заложены в трудах G. Piatetsky-Shapiro, R. Agrawal, J. Han, T.M. Mitchell, W.F. Frawley, B. А. Дюка, В. В. Коренькова, Ю. С. Сахарова, И. Г. Благовещенского и А. В. Замятина. Эти исследователи сформировали концептуальный аппарат и

методологические принципы Data Mining, разработали базовые алгоритмы поиска ассоциативных правил, классификации и кластеризации, создав теоретический фундамент для извлечения знаний из данных.

В области интеграции и управления разнородными данными значительный вклад внесли А. И. Аветисян, В. И. Будзко, Ю. А. Зеленков, А. Е. Янковская, M. Stonebreaker, J. Gray, R. Kimball, I.F. Ilyas, D. Abaid. Разработанные этим научным сообществом подходы, включая концепций озер данных и хранилищ больших объемов данных, онтологического инжиниринга и управления метаданными, обеспечивают технические основы для консолидации разноформатных данных, однако остаются ограниченными в решении задач семантической интеграции разнородных информационных объектов.

Качество и предобработка данных исследовались в работах C. Fan, E. E. Simoudis, H.G. Miller, J.M. Hellerstein, D. Barbara, Д. Барбары, В. Б. Яковлева, С. В. Ключарева, А. Н. Тихонова, М. Г. Шеразадишвили и П. С. Бондаренко. Разработанные этим сообществом методы очистки, восстановления и верификации информации создают важный задел для обеспечения достоверности анализа, однако носят фрагментарный характер и не образуют целостной технологии сквозной предобработки.

Проблемы интерпретируемости и этики в интеллектуальном анализе данных активно исследуются К. В. Воронцовым, А. Ю. Дьяконовым, С. В. Мирошниковым, K.T. Lundberg, M.T. Ribeiro, S.M. Fatemi, S.M. Hosseini, C.D. Waitz. Их работы в области объяснимого искусственного интеллекта и дифференциальной приватности создают теоретическую базу для построения доверяемых аналитических систем, однако не решают в полной мере проблему «семантического разрыва» между сложными моделями ИАД и потребностями экспертов в интерпретируемых результатах.

В диссертационной работе предложено комплексное решение, включающее формальную модель представления цифровых информационных объектов, методы автоматизированного сбора, обработки, насыщения и

хранения данных, а также специализированные программные средства для их анализа и визуализации. Интеграция разнородных источников информации на основе единой модели и применение совокупности методов ИАД позволит получать новые знания и решения в междисциплинарных задачах быстрее и надёжнее, чем при использовании разрозненных подходов.

Объектом исследования являются цифровые информационные объекты, формируемые в процессе научно-технической деятельности и социальной коммуникации, а именно данные научных публикаций, патентных описаний, отчетов, записей социальных сетей и иных источников, значимые для решения научно-технических и социальных задач.

Предметом исследования являются методы, модели, алгоритмы и программные средства интеллектуального анализа данных цифровых информационных объектов, обеспечивающие интеграцию разнородных данных в единую модель, насыщение характеристик цифровых объектов и выявление неявных знаний для решения научно-технических и социально значимых задач.

Цель работы состоит в разработке и обосновании системы интеллектуального анализа информационных объектов, объединяющей модели представления данных, методы автоматизированного извлечения значений характеристик из разнородных ресурсов, методы насыщения данных, механизмы визуальной аналитики для выявления явных и неявных закономерностей и поддержки принятия решений в научно-технических и социально значимых задачах.

Для достижения указанной цели в работе решаются следующие **основные задачи:**

1. Обобщение модели комплексного цифрового информационного объекта, объединяющей разные виды характеристик (статические, динамические, вычисляемые) и отношения между объектами.

2. Разработка методов и алгоритмов автоматизированного извлечения информации из разнородных источников (научные статьи, веб-документы) и насыщения данных (выделение ключевых сущностей,

распознавание и нормализация физических величин, обработка изображений и таблиц, определение международных альянсов) для наполнения модели цифрового информационного объекта, обеспечивающих учёт специфики научно-технической информации.

3. Разработка методов идентификации целевых социальных объектов, в том числе алгоритмов анализа текстовых полей профилей цифровых объектов для обнаружения лингвистических индикаторов; методов анализа динамических характеристик активности цифровых объектов для выявления аномальных паттернов; метода расчета обобщённой вычисляемой характеристики, интегрирующей значения характеристик модели цифрового информационного объекта, и определения ее порогового значения для отнесения цифровых объектов к целевой группе.

4. Выбор и обоснование подходов к визуальному анализу научно-технической информации, позволяющих выявлять тенденции, ведущие организации, неявные связи, кластеры тематических направлений, международное сотрудничество на основе построения интерактивных аналитических панелей, обеспечивающих интеграцию различных цифровых объектов; построению графовых моделей, отражающих связи между объектами.

5. Проектирование и реализация программных средств, интегрирующих предложенные модели и методы в единый программно-аналитический комплекс, обеспечивающий функциональность поиска, выявление как явных, так и неявных связей между объектами, построение аналитических срезов научно-технологической и социальной сфер.

Научная новизна состоит в разработке и обосновании нового комплексного подхода к анализу разнородной информации, включающего оригинальные модели и методы. Основные научные результаты, полученные автором, заключаются в следующем:

1. **Предложена и обоснована обобщённая аналитическая модель комплексного цифрового информационного объекта,** объединяющая

статические, динамические и вычисляемые характеристики, а также систему связей между объектами. Разработанная модель обеспечивает единое представление разнородной научно-технической и социальной информации, что упрощает подготовку данных к анализу и повышает точность и воспроизводимость аналитических результатов за счёт стандартизации структуры данных.

2. **Разработан новый методический аппарат насыщения данных из неструктурированных разнородных источников**. Предложены оригинальные методы автоматизированного извлечения и насыщения данных, отличающиеся учётом специфики научно-технической информации, в том числе: метод распознавания физических величин и единиц измерения с их приведением к единому стандарту (СИ); метод обработки нетекстовых элементов научных документов – извлечение и структурирование содержимого таблиц и подписей к рисункам и др. Новизна указанных методов заключается в адаптации технологий обработки данных (NLP, OCR, геокодирование и др.) к задачам научно-технического контента, что позволяет существенно обогатить исходные данные и извлечь новые знания, недоступные при стандартной обработке информационных материалов.

3. **Предложена методика идентификации целевых объектов в социальной среде**, использующая методы семантического анализа текстовой информации профиля цифрового объекта (посты, комментарии, описания), обработки естественного языка и анализа тональности, количественной оценки динамических характеристик активности пользователя (частота и время публикаций, смена аудитории), ранжирования и нормирования множества характеристик профиля с учётом их значимости.

4. **Предложен метод построения интерактивных аналитических панелей**, позволяющий работать со слабоструктурированными массивами научной информации для проведения сравнительного анализа динамики развития различных областей науки, комплексного обзора научно-технологического ландшафта: выявление лидеров и динамики

публикационной активности, основных исследовательских трендов, картографирования международного сотрудничества. Впервые реализован междисциплинарный анализ публикационной активности, показавший эффективность при выявлении скрытых тематических акцентов и точек роста.

5. **Сформирована система интеллектуального анализа данных для решения широкого спектра актуальных научно-технических и социально значимых задач,** базирующаяся на обобщённой модели описания цифровых информационных объектов, методах преобразования информации из различных источников и визуальной аналитики больших объемов разнородных данных.

Теоретическая значимость работы заключается в развитии методологических основ интеллектуального анализа данных применительно к разнородным информационным объектам в научно-технической и социальной сферах. Предложенные автором модели формализуют новый подход к представлению знаний – через комплексный цифровой информационный объект с множеством разнотиповых (статических, динамических и вычисляемых) характеристик и связей. Разработанные модели и методы расширяют аппарат интеллектуального анализа данных: введены новые вычисляемые характеристики и метрики для анализа социальных и научных данных, предложены новые методы извлечения знаний из неструктурированных данных. Полученные в работе обобщения (например, категоризация характеристик социальных профилей, формулы нормировки признаков, концепция эволюции модели цифрового объекта) могут служить основой для дальнейших исследований по интеграции данных и развитию теории представления знаний. Полученные результаты закладывают методологический фундамент для развития гибких аналитических систем, объединяющих машинное обучение, базы знаний и механизмы визуализации. Теоретическая ценность работы состоит в междисциплинарном подходе, который объединяет концепции информатики (структуры данных, алгоритмы), искусственного интеллекта (машинное обучение, NLP),

социологии (психографический анализ) и наукометрического анализа в рамках единой научной парадигмы. Таким образом, диссертация вносит значимый вклад в развитие теории интеллектуального анализа данных, предлагая новые модели и методы, расширяющие границы применимости интеллектуального анализа данных.

Практическая значимость результатов исследования подтверждается их внедрением и опытным применением в ряде проектов. Разработанные модели и методы легли в основу программных средств, использованных для решения прикладных задач в интересах ведущих научных организаций. С 2018 года автор являлся руководителем или ответственным исполнителем 10 хоздоговорных работ в интересах Министерства образования и науки Российской Федерации, организаций контура Госкорпорации Росатом (НИИ «Графит», ВНИИА им. Н.Л. Духова, ФГУП «РФЯЦ-ВНИИТФ им. академ. Е.И. Забабахина»), Фонда перспективных исследований, Российского энергетического агентства. В рамках государственного задания Министерства образования и науки РФ №2.12915.2018.12.1 «Разработка и апробация информационной системы комплексной антисуицидальной интернет-профилактики» был реализован и прошёл тестирование метод идентификации целевых социальных профилей, показав работоспособность в реальных условиях мониторинга социальных сетей. По государственному заданию Министерства науки и высшего образования Российской Федерации №3466-22 «Создание учебно-методических материалов по финансовой безопасности для школьников и студентов, в том числе для передачи указанных учебно-методических материалов в зарубежные страны-партнеры Международного сетевого института в сфере противодействия отмыванию доходов, полученных преступным путем, и финансированию терроризма» апробированы методы построения интерактивных аналитических панелей и проведено исследование состояния научно-технических разработок по направлению «Финансовая безопасность». По договору №349ГС1ЦТС10-D5/80243 от 12.12.2022 «Разработка и тестирование прототипа мультиагентной системы обработки и

представления неструктурированных массивов данных» с Фондом содействия инновациям создан программный комплекс «СИА.Атташе», предназначенный для интеллектуального анализа научно-технической информации, прошедший апробацию по договорам № 2024-sia-dgk-1 от 15.04.2024 и № 2024-sia-dgk-2 от 15.05.2024. Практический эффект от использования комплекса выражается в существенном сокращении времени на сбор и обработку данных, например, по договору № 1707 от 29 августа 2022 г. по выполнению НИР «Разработка программы выборки данных по свойствам и структурам облученных реакторных материалов из мировых источников информации» с ВНИИА им. Н.Л. Духова за три месяца были выполнены работы по сбору и обработке более 40 тысяч научных публикаций по материалам реакторных исследований. В результате заказчику было передано ~8700 точек, описывающих требуемые свойства реакторных материалов. Оценочно, решение такой задачи в ручном режиме заняло бы более года. Таким образом, практическая ценность работы состоит в том, что её результаты и разработки доведены до прикладных решений, используемых для ускорения и улучшения аналитических процессов в науке и промышленности. Разработанные методы и технологии могут быть оперативно адаптированы под новые задачи – от мониторинга технологических направлений до систем поддержки принятия решений в социальной сфере – что подтверждает их практическую значимость и универсальность.

Методы исследования. В диссертации использован комплекс методов, соответствующий междисциплинарному характеру поставленных задач. Теоретической основой послужили методы системного анализа (для формализации модели информационного объекта и постановки требований к ней), теория графов и сетевой анализ (при разработке схем представления связей между объектами), методы машинного обучения и обработки естественного языка (NLP) (для извлечения сущностей из текстов, классификации и кластеризации данных), методы визуализации данных и человеко-машинного взаимодействия (для реализации интерактивных

аналитических панелей). Для реализации алгоритмов применялись современные технологии программирования и работы с данными: инструменты веб-скрапинга и парсинга HTML-документов для сбора информации из интернет-источников, библиотеки компьютерного зрения для извлечения содержимого из изображений и PDF-документов, платформы хранения и поиска по неструктурированным данным (NoSQL базы, Elasticsearch) и фреймворки визуализации данных (например, Kibana, D3.js).

На защиту выносятся следующие положения:

1. **Обобщённая модель комплексного цифрового информационного объекта,** обеспечивающая унифицированное представление различных по природе данных (научные статьи, социальные профили) за счёт выделения статических, динамических и вычисляемых характеристик объекта и использования графовой схемы для хранения взаимосвязей. Применение данной модели повышает точность и воспроизводимость анализа данных за счёт уменьшения количества ошибочных или неоднозначных трактовок и облегчения слияния данных из разных источников.

2. **Методы интеллектуальной обработки данных:** метод автоматического выделения информативных признаков (ключевых слов, физических параметров) из текстовых массивов научно-технической информации; метод распознавания структурированных элементов (таблиц, рисунков) в научных публикациях и их конвертации в цифровую форму; метод унификации геопространственной привязки организаций по месту работы авторов и выявления на этой основе научных сообществ и международного сотрудничества. Предложенные методы обеспечивают наполнение информационной модели достоверными и актуальными данными в автоматизированном режиме.

3. **Методика аналитического описания и идентификации социальных объектов,** включающая разбиение характеристик профиля на статические и динамические, введение системы весовых коэффициентов

значимости признаков и вычисление интегрального показателя соответствия профиля целевому образу. Пороговое правило по этому показателю позволяет выделять целевые группы пользователей. Предложенная методика продемонстрировала эффективность при решении задачи обнаружения групп риска в социальных сетях.

4. **Система интеллектуального анализа информационных объектов,** включающая концепцию комплексного цифрового информационного объекта с многоуровневой структурой характеристик и связей, совокупность методов извлечения и насыщения данных из разнородных источников и механизмы построения специализированных аналитических инструментов для анализа публикационной активности и обеспечивающие многомерное представление и исследование состояния науки в приоритетных направлениях.

5. **Специализированные программные средства интеллектуального анализа,** в частности: интерактивный инструмент построения графовых представлений, позволяющий выделять взаимосвязанные кластеры публикаций, авторов и организаций и обнаруживать в них неявные связи; инструмент автоматизированного построения научно-технологических ландшафтов, формирующий на основе массива публикаций карту исследовательского пространства с выделением ключевых тематик и динамики их развития.

6. **Программный комплекс интеллектуального анализа данных,** созданный в рамках работы, интегрирует процессы сбора, обработки, хранения и анализа информации из разнородных источников, обеспечивает высокую гибкость и масштабируемость системы при добавлении новых типов данных или методов анализа.

Апробация результатов работы. Основные результаты работы представлены на международных и всероссийских научных конференциях, семинарах и школах, в частности, на: Международной конференции GRID-2025 и GRID-2023 (Россия), Международной конференции по компьютерной

графике и зрению GraphiCon-2023 и GraphiCon-2022 (Россия), Научная сессия НИЯУ МИФИ-2024 (Россия), Международном симпозиуме по ядерной электронике и вычислительной технике NEC’2019 (Будва, Черногория), ежегодной конференции SPBPU IDE-2020 (Санкт-Петербург, 2020). Результаты работы в части анализа больших объемов научно-технической информации докладывались на научно-технических советах ГК «Ростех», ФГУП «РосРАО», в части решения социально значимых задач – на Межведомственной рабочей группе Министерства науки и высшего образования.

Тематика исследований соответствует паспорту специальности 2.3.1 – Системный анализ, управление и обработка информации, статистика, по разделам: П.2 – формализация и постановка задач системного анализа, оптимизации, управления, принятия решений, обработки информации и искусственного интеллекта; П.5 – разработка специального математического и алгоритмического обеспечения систем анализа, оптимизации, управления, принятия решений, обработки информации и искусственного интеллекта; П.12 – визуализация, трансформация и анализ информации на основе компьютерных методов обработки информации.

Личный вклад. Все основные результаты, изложенные в диссертации, включая постановки задач и их алгоритмические решения и созданное программное обеспечение, получены автором лично или выполнены под его научным руководством и при непосредственном участии.

Публикации. Основные положения диссертационной работы опубликованы в **47** печатных работах, из них **9** статей в изданиях, индексируемых в библиографических и реферативных базах данных Web of Science и/или Scopus, **4** статьи в изданиях, рекомендованных ВАК при Министерстве науки и высшего образования Российской Федерации для опубликования основных научных результатов диссертаций на соискание ученой степени доктора наук, **16** статей в материалах международных конференций, **2** учебно-методических пособия. По научно-техническим

разработкам в составе коллектива авторов получено **6** свидетельств о регистрации баз данных и **10** свидетельств о регистрации программ для ЭВМ в Федеральной службе по интеллектуальной собственности Российской Федерации.

Структура и объем диссертации. Диссертация содержит введение, шесть глав, заключение, перечень используемых сокращений, список используемой литературы, 3 приложения. Работа состоит из **287** страниц, из них основной текст **224** страницы, включая **76** рисунков, **12** таблиц и список литературы, содержащий **178** наименований.

ГЛАВА 1 СОВРЕМЕННЫЕ МЕТОДЫ ИНТЕЛЛЕКТУАЛЬНОГО АНАЛИЗА ДАННЫХ

1.1 Современные характеристики больших объемов данных

Современное развитие информационно-коммуникационных технологий и стремительный рост объёмов данных привели к пересмотру классических подходов к обработке и анализу научно-технической и социальной информации. Развитие технологий искусственного интеллекта (ИИ), особенно больших языковых моделей (БЯМ), обусловило рост не только информации, созданной человеком, но и появление огромного количества синтезированных данных. В условиях экспоненциального увеличения количества и объема информационных потоков актуализируется необходимость разработки методологических принципов и инструментов, которые обеспечивают эффективный анализ и интерпретацию перечисленных типов информации. Ключевой задачей является выявление новых данных из больших объемов текстовой информации, обеспечение высокой степени достоверности и релевантности для поддержки принятия управленческих решений.

Решение поставленной крупной задачи возможно с применением методов интеллектуального анализа данных в том числе посредством интеграции разнородной информации (социальные сети, научные публикации), что требует применения методов и технологий графового анализа, обработки естественного языка и мультимодального анализа больших массивов данных.

Современная парадигма работы с большими данными (Big Data) предполагает их непрерывный сбор, обработку, анализ и визуализацию из разнородных источников. Масштабный рост объёмов цифровой информации, а также её разнообразие и многоплановость требуют системного подхода к определению и структурированию данных. Основные характеристики, влияющие на работу с большими данными описаны в концепции 7V Volume, Velocity, Variety, Veracity, Variability, Visualization, Value, то есть объем,

скорость, разнообразие, достоверность, изменчивость, визуализация, ценность. Рассмотрим их детально:

Объём (Volume). Объём данных является первостепенным фактором, определяющим сложность их обработки и хранения [1]. Увеличение объёма данных ставит задачу обеспечения соответствующей инфраструктуры (серверных мощностей, распределённых систем хранения и эффективных алгоритмов обработки).

Разнообразие (Variety). Данные представлены в различных форматах, включая структурированные, полуструктурированные и неструктурированные [2]. Такой гетерогенный характер порождает трудности при их классификации и последующей аналитической обработке. Методы интеграции данных, онтологии и семантические модели играют важную роль в упрощении работы с разнородными источниками, позволяя выделять значимые взаимосвязи между элементами различных форматов.

Достоверность (Veracity). Под достоверностью понимается степень точности и надёжности информации [3]. В условиях, когда источники данных могут иметь неодинаковую степень валидности, крайне важно выявлять и корректировать потенциальные искажения, а также учитывать погрешности и неполноту сведений. Достоверность во многом определяет качество аналитических исследований, так как ошибки, допущенные на ранних этапах обработки, способны приводить к неверным выводам и стратегическим решениям.

Скорость (Velocity). Скорость формирования, поступления и обработки данных становится критически важной при работе с потоковой информацией, в том числе в режиме приближенного к реальному времени. Согласно работам в области крупных потоковых систем, возможность оперативной обработки непрерывного потока информации обеспечивает конкурентное преимущество при принятии решений.

Ценность (Value). Релевантность данных с точки зрения последующего анализа и принятия решений, или их ценность, занимает центральное место в

аналитических исследованиях. Сама по себе информация не всегда имеет непосредственную ценность; её значимость проявляется в том, насколько полученные выводы влияют на формирование эффективных стратегических и тактических мер. Методы оценки ценности включают различные системы показателей, позволяющие определить, какие данные представляют наибольшую пользу при решении конкретных исследовательских или прикладных задач.

Таким образом, **объём**, **разнообразие**, **достоверность**, **скорость** и **ценность** формируют базовые характеристики больших данных. Их учёт является необходимым условием для дальнейшей разработки методов анализа, хранения и обеспечения высокого качества выходных результатов.

Помимо характеристик, перечисленных выше, ряд дополнительных факторов существенно влияет на качество и применимость больших данных в аналитических исследованиях.

Избыточность (Redundancy). Избыточность возникает в случаях, когда в информационном массиве присутствуют дублирующие друг друга сведения. С одной стороны, это позволяет повысить надёжность и устойчивость к потерям данных (резервирование), однако избыточность усложняет процессы очистки и интеграции информации. Также под вопросом становится получение достоверных статистических данных. В ситуации высокого уровня избыточности важным становится применение алгоритмов дедупликации.

Связность (Connectivity). Под связностью данных подразумевается степень взаимосвязи между отдельными объектами в анализируемой системе [4]. Графовые структуры широко используются для формальной репрезентации подобных взаимосвязей, упрощая задачи выявления центральных узлов, кластеров и закономерностей в сетях (например, в социальных медиа или научно-технической коллаборации). Высокая связность, с одной стороны, усложняет аналитические вычисления, с другой – появляется возможности для глубинного анализа сетевых эффектов.

Контекстуальность (Contextuality). Контекстуальный аспект охватывает специфические условия и предметные области, в которых интерпретируются данные. В научной и промышленной практике адекватный учёт контекста существенно повышает точность и ценность получаемых результатов. Выявление и корректировка информации относительно конкретного контекста обеспечивают релевантность итоговых рекомендаций и прогнозов.

Степень влияния каждой из этих характеристик меняется в зависимости от конкретной предметной области, что подчеркивает необходимость разработки методологии интеллектуального анализа информационных объектов.

1.2 Обзор методов интеллектуального анализа данных

В последние годы наблюдается **стремительный рост объёмов и разнообразия данных** в научно-технической и социальной сферах. Развитие облачных технологий, Интернета вещей (IoT) и других информационно-коммуникационных систем привело к появлению *комплексных задач анализа данных*, имеющих непосредственное влияние на науку, экономику и общество [5, 6, 7]. Традиционные подходы к обработке информации уже не справляются с текущими реалиями – объёмы данных растут экспоненциально, данные поступают с высокой скоростью и в разнотипных форматах, что порождает новые вызовы по их хранению, обработке и интерпретации [8, 9]. В таких условиях возникает необходимость в развитии **методологических принципов и инструментов интеллектуального анализа данных (ИАД)**, способных эффективно выявлять знания в больших массивах разнородной информации и поддерживать принятие решений в сложных ситуациях [7, 10, 11, 12].

Одним из ключевых вопросов становится **интеграция и анализ данных из различных источников** – от научных публикаций и сенсорных измерений, до социальных сетей и мультимедийных потоков. Современные технологии искусственного интеллекта, включая методы машинного обучения и когнитивные вычисления, выдвигаются на первый план как основа для

решения подобных задач [7]. Благодаря им удаётся обрабатывать растущую сложность данных и формировать обоснованные стратегии развития и инноваций [7, 11, 12]. В результате *интеллектуальный анализ данных* превратился в междисциплинарное направление, объединяющее достижения в областях анализа социальных сетей, обработки естественного языка, гранулярных вычислений, машинного обучения и др. [7]. Это подтверждается широким участием исследователей различных профилей в создании методов ИАД, а также ростом числа работ, посвящённых данному направлению. Так, например, в сфере медицины доля исследований с использованием методов машинного обучения возросла многократно – к 2021 году **каждая четвертая научная статья** по диагностике посвящена применению ML-алгоритмов, причём ежегодный прирост таких публикаций (~39%) значительно опережает традиционные исследования [13, 14]. Эти тенденции подчёркивают актуальность развития методологии ИАД для решения актуальных научно-технических и социальных задач.

Интеллектуальный анализ данных (ИАД), часто отождествляемый с data mining, представляет собой совокупность методов и технологий для обнаружения в «сырых» данных ранее неизвестных, неочевидных, практически полезных и интерпретируемых знаний, значимых для принятия решений [15]. Иными словами, ИАД – это широкий *комплекс математических моделей, алгоритмов и программных средств*, предназначенных для автоматического выявления скрытых закономерностей и генерации новых знаний на основе эмпирических данных [15, 16]. В отличие от отдельных статистических методов, интеллектуальный анализ данных носит интегративный характер и охватывает многие направления, поэтому под данным термином понимается не один конкретный метод, а целый класс подходов к анализу информации [17, 18].

С точки зрения решаемых задач, методы ИАД охватывают несколько основных категорий.

– **Классификация** и **регрессия** нацелены на построение предсказательных моделей: в первом случае объекту присваивается один из заранее известных классов, во втором – предсказывается числовое значение целевой переменной. Эти методы относятся к *контролируемому обучению* и требуют наличия размеченных данных, но позволяют с высокой точностью проводить диагностику состояний, прогнозировать показатели и т.д.

– **Кластеризация**, напротив, относится к *неконтролируемым методам*: она группирует объекты по схожим признакам без априорных меток, выявляя скрытые структуры и сегменты в данных. Широко применяется и поиск ассоциаций – выявление устойчивых правил и связей между признаками, что особенно полезно в анализе потребительского поведения, биоинформатике и др.

– **Поиск аномалий** – выявление отклоняющихся наблюдений, потенциально указывающих на редкие события (например, сбои оборудования или мошенничество). Указанные задачи отражают разные аспекты анализа данных, и для их решения разработан богатый арсенал методов.

По происхождению и подходу можно условно выделить **две большие группы методов ИАД**. Первая группа – методы, происходящие из классической статистики и исчислений: сюда относятся регрессионный анализ, дискретный анализ, деревья решений, факторный анализ и др. Их достоинство – строгая теория, интерпретируемость результатов и относительная простота реализации. Однако применение таких методов затруднено при очень больших объёмах или высокой размерности данных; кроме того, они часто предполагают определённые гипотезы о распределении данных, которые не всегда выполняются на практике.

Вторая группа – методы **искусственного интеллекта и машинного обучения**, которые во многом революционизировали анализ данных за последние десятилетия. К ним относятся нейронные сети (в том числе глубокие), метод опорных векторов, случайные леса, градиентный бустинг, методы обучения без учителя (кластеризация, снижение размерности) и с

подкреплением, а также гибридные подходы. Эти алгоритмы способны *автоматически выявлять сложные нелинейные зависимости* и скрытые факторы в данных, минимизируя участие человека в настройке моделей. Например, современные **глубокие нейросети** превосходят классические подходы по точности и скорости при анализе больших многомерных массивов [9, 19, 20], что продемонстрировано во многих областях – от распознавания изображений до предсказания поведения сложных систем. Благодаря этому, машинное обучение стало мощным инструментом для работы с большими данными, позволяя эффективно обрабатывать огромные массивы и извлекать из них значимые инсайты, недоступные ранее [9, 19, 20].

В то же время **у** каждого класса методов имеются ограничения. Статистические модели зачастую плохо масштабируются и могут не учесть всю сложность реальных процессов. Методы машинного обучения требуют больших объемов обучающих данных, чувствительны к их качеству и могут быть трудоинтерпретируемыми («прозрачность» моделей снижается с ростом их сложности). Так, современные глубокие ансамблевые модели демонстрируют высокую точность и устойчивость к шуму, однако достигается это ценой существенных вычислительных затрат и сложной реализации [7]. **Интерпретируемость** результатов становится серьёзной проблемой: многие мощные алгоритмы функционируют как «черный ящик», что затрудняет их использование в ответственных областях (медицина, правовая сфера и пр.), требующих объяснимости выводов. Для преодоления этого барьера развивается направление *Explainable AI* (объяснимого ИИ), предлагающее методы раскрытия внутренней логики моделей. Ещё один вызов – **качество и безопасность данных**: при анализе реальных информационных потоков необходимо учитывать неполноту, погрешности и предубеждения в исходных данных. Особенно остро стоит вопрос о **конфиденциальности и приватности**: обработка больших пользовательских данных требует надёжных мер анонимизации и шифрования [21, 22]. Таким образом, методология ИАД постоянно эволюционирует, стремясь найти баланс между

мощностью моделей и контролем над их надежностью, прозрачностью и этичностью применения.

1.3 Применение ИАД в научно-технических и социально значимых задачах

Практическая ценность методов интеллектуального анализа данных подтверждается успешными кейсами их применения в различных сферах [23, 24]. Рассмотрим некоторые из них, наиболее значимые в научно-техническом и социальном контексте:

Медицина. Объем медицинских данных стремительно увеличивается (геномные последовательности, базы электронных медицинских карт, результаты исследований и пр.), что стимулирует внедрение ИАД для повышения эффективности здравоохранения. Современные ML-алгоритмы уже демонстрируют выдающиеся результаты в диагностике заболеваний (например, распознавание патологий по изображениям МРТ/КТ), прогнозировании течения болезней и подборе персонализированных схем лечения. От экспериментальных разработок индустрия перешла к реальным внедрениям: **системы на основе ИИ используются в клинической практике** и показывают высокую точность, порой сопоставимую с экспертами-врачами [25]. Яркий пример – алгоритмы глубинного обучения, позволяющие по рентгеновским снимкам выявлять онкологические заболевания на ранних стадиях, или решение AlphaFold, предсказывающее пространственную структуру белков, что открывает новые возможности для биомедицины [25]. Интеллектуальный анализ медицинских данных также помогает оптимизировать административные процессы: от управления потоками пациентов до анализа затрат, что особенно важно на фоне удвоения глобальных расходов на здравоохранение за последнее десятилетие [26]. В то же время сохраняются и проблемы – например, ограниченное количество данных в узкоспециализированных областях (как показал обзор по ИИ в нейрохирургии, таких исследований пока немного) и необходимость тщательной валидации алгоритмов перед клиническим применением [27]. Тем

не менее, социальная значимость этих направлений стимулирует дальнейшие исследования, призванные сделать ИАД неотъемлемым инструментом медицины.

Промышленность. В условиях концепции *Индустрии 4.0* интеллектуальный анализ данных стал ключевым фактором повышения эффективности производства. Сбор больших данных от датчиков на оборудовании и производственных линиях вместе с методами ML позволил реализовать *предиктивное обслуживание* – предсказание отказов оборудования до их возникновения. Это изменило подход к техническому обслуживанию: планирование ремонтов теперь основано на анализе реальных показателей в режиме близком к реальному времени, что снижает простои и издержки. **Методы машинного обучения значительно повлияли на производственный сектор**, обеспечивая оптимизацию процессов и качество продукции на основе анализа многомерных данных в реальном времени [28]. Например, алгоритмы аномалий используются для обнаружения дефектов на конвейере, а модели прогнозирования – для оптимизации цепочек поставок с учетом множества динамических факторов. Интеллектуальный анализ также находит применение в создании *«цифровых двойников»* – виртуальных моделей оборудования или процессов, обученных на данных сенсоров. Такие модели позволяют проводить эксперименты и прогнозировать развитие ситуаций без риска для реального производства. Отдельно стоит отметить влияние ИАД на **энергетику и ресурсосбережение**: анализ больших данных помогает более рационально использовать ресурсы, предсказывать пики нагрузки и предотвращать аварии. В промышленности методология ИАД сталкивается с такими вызовами, как интеграция разнородных данных (от IoT-устройств, систем ERP и др.), обеспечение кибербезопасности и требование интерпретируемости результатов для инженерного персонала. Достигнутые успехи – например, снижение затрат на обслуживание оборудования благодаря предиктивным моделям – свидетельствуют о революционном потенциале ИАД в этой сфере [29].

Астрономия и космические исследования. Астрономия одной из первых столкнулась с проблемой **больших данных**: современные телескопы, спутники и симуляции генерируют колоссальные объёмы высокоточного материала. Наступила новая эпоха, когда успех исследований напрямую зависит от способности обрабатывать и интерпретировать огромные массивы информации. Традиционные методы – визуальный анализ снимков, ручная классификация объектов – оказались **неэффективными и медленными** при таких масштабах [8, 9]. Интеллектуальный анализ данных открыл для астрономов принципиально новые возможности. С помощью алгоритмов машинного обучения автоматизируется классификация космических объектов (звёзд, галактик, сверхновых и т.д.), выявляются тонкие закономерности, ускользающие от глаза человека. Например, нейросетевые модели успешно применяются для распознавания экзопланет по незначительным колебаниям свечения звезды или для сортировки миллионов галактик по их морфологии. **Машинное обучение стало мощным инструментом** обработки астрономических данных – показано, что ML-алгоритмы способны эффективно работать с многомерными наборами и превышать точность традиционных подходов [9, 19, 20]. Это позволяет, в частности, автоматически определять свойства звёзд (массу, возраст, химический состав) по спектральным данным с точностью, ранее недостижимой вручную [9, 19, 20, 30, 31]. Важным направлением является и анализ *сети космических объектов* – так, графовые алгоритмы помогают изучать структуры космических скоплений, связи между галактиками, выявлять центральные объекты во взаимосвязанных системах. Без применения ИАД современные проекты вроде обзора неба LSST или миссии Gaia были бы просто невозможны: они требуют автоматизированного обнаружения событий (например, вспышек сверхновых) и быстрого извлечения знаний из постоянно поступающего потока данных. Основные сложности применения ИАД в астрономии связаны с необходимостью обработки **очень больших объемов данных в реальном времени**, а также с качеством данных (шумы, пропуски измерений). Тем не

менее, достижения последних лет демонстрируют, что интеллектуальный анализ стал неотъемлемой частью астрономических исследований и продолжает продвигать науку, помогая делать открытия на основе данных.

Социальные сети и социологический анализ. Глобальная цифровизация общественной жизни привела к тому, что социальные сети, блоги и другие онлайн-платформы генерируют огромные массивы текстовых, графических и сетевых данных об активах и предпочтениях людей. Интеллектуальный анализ этих данных предоставляет уникальные инструменты для понимания социальных процессов. Применение методов ИАД в анализе социальных сетей (Social Network Analysis, SNA) позволяет **выявлять скрытые структуры взаимоотношений, лидеров мнений, распространение информации и настроения масс**. Например, алгоритмы обработки естественного языка используются для анализа тональности (sentiment analysis) миллионов сообщений, чтобы отслеживать общественное мнение или реакцию на события в режиме реального времени. Графовые методы помогают обнаруживать сообщества пользователей по структуре их взаимодействий, идентифицировать «инфлюенсеров» и моделировать распространение новостей или слухов в сети. Современные исследования отмечают, что анализ социальных сетей переживает бурный рост, особенно с появлением технологий больших данных и машинного обучения [21, 22]. В то же время имеются и сложности: **приватность данных** становится острой проблемой, так как требуется анализировать информацию о поведении пользователей при соблюдении этических норм и законодательства [21, 22]. Кроме того, динамичность соцсетей затрудняет получение устойчивых выводов – модели должны регулярно адаптироваться к быстро меняющимся трендам и языковым мемам. Для более эффективного исследования социальных данных учёные все чаще интегрируют методы *ML* непосредственно в SNA, получая интеллектуальные системы мониторинга социальных процессов [21, 22, 32, 33]. Перспективным направлением является развитие *предиктивной аналитики* в социальных сетях – от прогнозирования

вирусного распространения контента до раннего обнаружения признаков кризисных ситуаций (например, социальных конфликтов или эпидемий) по аномалиям в пользовательской активности [34, 35]. Практическая ценность ИАД в социальной сфере уже доказана: так, анализ больших данных социальных медиа применяют государственные органы для мониторинга общественных настроений и противодействия дезинформации, маркетинговые компании – для таргетинга рекламы и прогнозирования поведения потребителей, а социологи – для изучения процессов самоорганизации сообществ. Можно ожидать, что дальнейшая интеграция машинного обучения и социального анализа (при одновременном решении вопросов приватности) станет одним из наиболее значимых драйверов развития данной области.

1.4 Роль визуализации данных в процессе ИАД

Визуализация данных играет критически важную роль на всех этапах интеллектуального анализа данных – от предварительного изучения наборов до представления полученных результатов. Человеческий мозг значительно лучше воспринимает информацию через зрительные образы, поэтому грамотное визуальное представление данных позволяет **быстро обнаруживать тренды, кластеры, аномалии и взаимосвязи**, которые сложно заметить при просмотре сырых цифр [36, 37, 38, 24, 24, 39]. Визуализация фактически превращает скрытые аспекты данных в наглядные образы, облегчая принятие решений на основе данных [40, 41, 42, 43]. Например, интерактивные **дашборды** позволяют аналитикам в реальном времени отслеживать ключевые показатели и реагировать на отклонения, а графики связей делают очевидными структуры социальных или научных сетей (коллаборации, цитирования и пр.). Не случайно визуализация выделяется как одна из ключевых характеристик концепции *Big Data* – наряду с объёмом, разнообразием, скоростью и достоверностью, **визуализация** включена в модель “7V” больших данных как необходимый компонент.

Однако эффективная визуализация больших данных связана с серьезными **технологическими и методологическими вызовами** [36].

Во-первых, объёмы данных могут быть столь велики, что их простое отображение затруднительно – возникают проблемы перегрузки графиков, снижения отзывчивости интерфейсов при потоковом обновлении и т.д.

Во-вторых, повышается **требование к интерактивности и времени отклика**: пользователям нужны инструменты, позволяющие исследовать данные практически в реальном времени, фильтровать, масштабировать и детализировать визуальные представления без задержек [36].

В-третьих, данные часто имеют высокую размерность (десятки и сотни признаков), и отразить многомерные взаимосвязи на плоскости или экране не тривиально. Поэтому традиционных статичных графиков может быть недостаточно – требуется применение специализированных методов: тепловые карты для матриц данных, проекции высоких размерностей (PCA, t-SNE) для кластеризации, геопространственные визуализации для данных с гео-привязкой и др.

Существующие инструменты визуализации (такие как библиотеки matplotlib, d3.js, платформы вроде Tableau и PowerBI) продолжают развиваться, но пользователям нередко требуется кастомизация под специфические задачи и умение правильно интерпретировать сложные графики. Критика современных инструментов связана также с тем, что неудачные визуализации способны ввести в заблуждение – например, некорректные масштабы на осях или перегруженные диаграммы затрудняют понимание и могут привести к неверным выводам.

В ответ на эти проблемы в последнее время активно разрабатываются новые методики визуализации, учитывающие особенности эпохи больших данных. Во-первых, набирают популярность **интерактивные визуализации** – системы, в которых пользователь может в режиме диалога с данными изменять параметры отображения, мгновенно получая обновленные графики. Это позволяет изучать данные гибко, проверяя гипотезы «на лету». Во-вторых, для задач мониторинга применяется **реал-тайм визуализация**: например, панели контроля, отображающие поступающие потоки данных (финансовых

транзакций, сетевого трафика, датчиков и т.д.) с минимальными задержками. В-третьих, на пороге широкого применения находятся **иммерсивные методы визуализации** – с использованием технологий виртуальной и дополненной реальности, которые обещают совершенно новый уровень восприятия больших данных [36]. Иммерсивные интерфейсы позволят буквально «погружаться» в данные, манипулировать ими в 3D-пространстве, что открывает путь к более глубокому пониманию сложных многомерных структур. Отдельно стоит отметить интеграцию технологий искусственного интеллекта в процесс визуального анализа: появляются средства автоматической генерации визуализаций на основе содержимого данных, *умные рекомендации* по выбору оптимального типа графика и адаптивные интерфейсы, подстраивающиеся под пользователя [36].

Такие решения помогают снизить роль человеческого фактора в этапе представления данных, избегать субъективных искажений и экономить время аналитиков. Наконец, в профессиональном сообществе отмечается необходимость **стандартизации и распространения культуры визуализации**: предлагаются единые подходы и библиотеки для построения воспроизводимых визуальных отчётов, обучение принципам визуального мышления включается в программы подготовки специалистов по данным [36]. Всё это направлено на то, чтобы визуализация окончательно превратилась из вспомогательного этапа в равноправный инструмент научного анализа, помогающий не только представлять результаты, но и находить новые знания.

Развитие методологии интеллектуального анализа информационных объектов является одним из краеугольных факторов успешной работы с **большими данными** в современных научно-технических и социально значимых проектах. Проведенный обзор показал, что методы ИАД уже доказали свою эффективность на практике – от медицины и промышленности, до астрономии и анализа поведения в Интернете.

Одновременно с решением прикладных задач, исследования в области ИАД способствуют развитию самой компьютерной науки и смежных

дисциплин, прокладывая путь к будущему с более высоким уровнем «данных грамотности» и технологических инноваций [7, 44]. Уже сейчас интеллектуальный анализ данных рассматривается как важный инструмент **интеллектуальной трансформации общества**, способный придать новый импульс развитию разных отраслей экономики и социальной сферы [7, 44].

Вместе с тем, необходимо отметить ряд **существенных пробелов и проблем**, которые предстоит решить научному сообществу. Во-первых, это **обеспечение качества и достоверности данных**: неполнота, шумы и предвзятость данных могут приводить к ошибочным выводам, а методы очистки и наполнения данных требуют дальнейшего совершенствования. Во-вторых, **проблема интерпретации и доверия** к моделям ИИ: необходимы исследования в области объяснимого ИИ, разработка понятных пользователю и эксперту методов верификации выводов сложных моделей. Без решения этого вопроса применение ИАД ограничено в высокоответственных областях. В-третьих, **конфиденциальность и этика** работы с данными: в условиях, когда анализ касается персональной или чувствительной информации, должны развиваться методы обезличивания, федеративного обучения и другие технологии, позволяющие получать инсайты без раскрытия частных данных [21, 22]. В-четвертых, **интеграция разнотипных источников и контекстуальных данных**: реальные задачи (например, в кризисных ситуациях) требуют объединения разнородной информации – текстов, изображений, геоданных, сенсорных показателей – в единую аналитическую модель; создание универсальных методов для таких случаев все еще находится в начальной стадии. Наконец, следует признать, что в ряде специализированных направлений (как упоминалось, например, в высокоточной хирургии) применение ИАД пока ограничено, и результаты исследований еще не внедрены в широкую практику [27]. Это свидетельствует о необходимости продолжать фундаментальные и прикладные изыскания, расширяя границы применимости интеллектуального анализа данных.

Подводя итог, **методология интеллектуального анализа данных** находится на этапе активного развития: она уже решает многие актуальные задачи, но в то же время ставит новые вопросы перед учёными. Заполнение

отмеченных пробелов – от повышения прозрачности алгоритмов до выработки стандартов работы с большими данными – станет приоритетом на ближайшие годы. Решение этих задач усилит доверие к системам ИАД и расширит сферу их применения. Таким образом, дальнейшее развитие методологии ИАД будет способствовать не только успешному решению научно-технических и социальных проблем, но и формированию новой интеллектуальной среды, где данные эффективно преобразуются в знания на благо общества [7, 44].

ВЫВОДЫ ПО ГЛАВЕ 1

1. Методы и технологии интеллектуального анализа данных являются одними из важнейших факторов успешной работы с большими данными для решения современных актуальных научно-технических и социально значимых задач.

2. Интегративный подход, объединяющий машинное обучение, статистический анализ, методы и алгоритмы обработки неструктурированных данных, обеспечивает получение новых знаний, ранее недоступных исследователю, что существенно повышает качество принимаемых решений и способствует развитию информационных технологий и смежных дисциплин.

3. Выделены проблемы, связанные с обеспечением качества и достоверности данных, интерпретации и доверия к моделям ИИ, конфиденциальности и этики работы с данными, а также интеграции разнотипных информационных ресурсов в едином хранилище.

4. Синтез методов искусственного интеллекта, визуальной аналитики и предметно-специфических знаний позволяет реализовывать интеллектуальные системы нового поколения, способные поддерживать принятие решений в условиях неопределённости, предсказывать и предотвращать кризисные ситуации, обеспечивать научно-технический прогресс.

ГЛАВА 2 МОДЕЛИ ФОРМАЛИЗОВАННОГО ОПИСАНИЯ ЦИФРОВОГО ИНФОРМАЦИОННОГО ОБЪЕКТА

2.1 Базовая информационная модель цифрового объекта

Информационно-коммуникационное пространство состоит из огромного количества слабоструктурированной информации. Рассматривая информационное пространство можно ввести следующую категоризацию представляемой информации:

- Текстовая информация,
- Аудиоинформация,
- Видеоинформация,
- Графическая информация,

В общем виде данная категоризация охватывает всю информацию, представленную в глобальной сети Интернет.

В свою очередь информацию, содержащуюся в глобальной сети, можно отнести к различным информационным объектам.

Под объектом будем понимать любую сущность, имеющую имя и границу (четкую или нечеткую) с окружающей средой. Необходимо сделать уточнение, что в работе рассматриваются информационные объекты, представленные в виртуальной среде Интернет – цифровой объект – и отдельно стоит задача соотнесения информационного объекта (образа физического объекта) с объектом реального мира. Данная задача не всегда может быть решена однозначно, что обусловлено самыми общими принципами организации сети Интернет и всемирной паутины.

Современные методы управления информацией и анализа данных базируются на формальных описательных моделях, позволяющих систематизировать информацию о различных сущностях и установить между ними логические связи. В контексте интеллектуального анализа научно-технических и социально значимых объектов особое значение имеет единообразие подходов к описанию свойств этих объектов и механизмов их трансформации.

Введем категоризацию двух основных групп информационных объектов глобальной сети Интернет, рассматриваемых в диссертационной работе:

- социальные объекты,
- научно-технические объекты.

Перечисленные группы тесно взаимосвязаны между собой, так как отдельные социальные объекты входят в состав научно-технической информации.

Введем категоризацию социальных объектов, рассматриваемых в работе:

- персона,
- организация,
- сообщение в социальной сети или мессенджере,
- государство.

Все остальные объекты будем считать частными случаями от перечисленных категорий.

В качестве научно-технических объектов рассмотрим:

- публикация,
- патент,
- научно-технический отчет.

В разделе решается задача разработки формального описания информационного объекта, способного:

1. Унифицировать представление данных разнородной природы (научные статьи, патенты, отчёты, социальные профили и т. д.) с учётом их особенностей.

2. Обеспечить связь статических (постоянных или редко изменяемых) и динамических (изменяющихся во времени) параметров, что особенно актуально для долгосрочных наблюдений и прогнозирования.

3. Формализованно определять вычисляемые характеристики, необходимые для анализа (например, рейтинги, индекс Хирша, суммарные метрики по публикациям и т. д.).

4. Гибко поддерживать реструктуризацию и расширение модели новыми характеристиками и связями по мере появления новых требований и задач.

5. Демонстрировать универсальность: модель должна быть применима в различных предметных областях, где необходимо хранить и интерпретировать комплексные объекты, а также синтезировать новую информацию на основе уже имеющихся данных.

Каждый из информационных цифровых объектов обладает некоторым, заранее неизвестным, набором характеристик. Приведем примеры характеристик для такого научно-технического объекта, как статья – авторы, аннотация, текст статьи, таблицы, библиография и др. Данный перечень характеристик не является конечным, так как для решения аналитических задач может потребоваться выделение специфичных характеристик, таких как формулы, рисунки и т.д.

Значения характеристик содержатся в глобальной сети как в агрегированном, так и в разрозненном виде.

Количество характеристик информационного объекта различное и может достигать несколько десятков. Причем объекты одного типа могут обладать всеми определенными характеристиками, а могут не обладать.

Введем понятие ***базового описания цифрового объекта*** в виде следующего набора компонент:

$$Obj = \langle ID, S, D \rangle, \tag{2.1}$$

где ID – уникальный идентификатор объекта,

S – совокупность статических характеристик,

D – совокупность динамических (изменяющихся во времени, либо под воздействием) характеристик.

Уникальный идентификатор (*ID*) служит для **однозначного определения** объектов при решении аналитических задач и для минимизации рисков дублирования или неоднозначной интерпретации. Характеристика является системной и задается при первой записи объекта в информационное хранилище.

Статические характеристики — это характеристики, значения которых либо **не меняется** в течение всего периода существования объекта, либо *меняется крайне редко*. Примеры:

1) ФИО автора научной публикации;
2) Основной государственный регистрационный номер организации;
3) Уникальный идентификатор (например, DOI статьи).

С формальной точки зрения, множество статических характеристик может быть задано как:

$$S = \{(s_1, \tau_1), (s_2, \tau_2) \dots, (s_k, \tau_k)\}, \quad (2.2)$$

где s_i – наименование i-*ой* статической характеристики, (например, автор, название организации),

τ_i – тип данной характеристики (строка, число и др.),

$i = 1,2, \dots, k$.

Разделение статических характеристик по типам данных позволяет корректно обрабатывать их на этапах агрегации и анализа (для текстовых полей и числовых полей применяются различные методы обработки).

Динамические характеристики — это характеристики, значения которых могут *изменяться во времени* или под воздействием внешних факторов.

В аналитических исследованиях иногда необходимо хранить историю изменений и иметь возможность возвращаться к состоянию объекта в определённый момент (или на конкретную дату). Примеры:

1) Количество просмотров сообщения;
2) Количество цитирований;
3) Количество сотрудников организации;

В рамках интеллектуального анализа такая изменчивость рассматривается как функция времени.

$$D = \{(d_1(t_1), t_1, \tau_1), (d_2(t_2), t_2, \tau_2), \ldots, (d_l(t_l), t_l, \tau_l)\}, \quad (2.3)$$

где $d_j(t_j)$ – значение j-*ой* динамической характеристики в момент времени t_j,

t_j – значение времени,

τ_j – тип данной характеристики (строка, число),

$j = 1,2, \ldots, l$.

Например, $(d_{cit}(t_1), t_1, \tau_1)$ – число цитирований статьи на момент времени t_1, τ_1 – числовое поле.

Такой подход позволяет анализировать состояния объекта в различное время *t* и сравнивать их; проектировать модели, учитывающие промежуточные стадии развития объекта.

Решение аналитических задач зачастую невозможно на основе только статических и динамических характеристик, что приводит к необходимости насыщения модели дополнительными характеристиками.

2.2 Аналитическая модель цифрового объекта

Проведение аналитических исследований предполагает не только структурирование информации об объекте, но и предоставление базиса для дальнейших процедур интеллектуального анализа. При этом важен баланс между уровнем детализации (чтобы модель реально отражала специфику данных) и общностью (чтобы она оставалась универсальной и расширяемой без избыточных сложностей).

Решение такой постановки задачи возможно при системном учёте не только статических и динамических характеристик, но и вычисляемых характеристик, а также механизма их связи с другими объектами.

Введем понятие аналитической модели цифрового объекта в виде следующего набора компонент:

$$AObj = \langle ID, S, D, F, Rel \rangle, \quad (2.4)$$

где *ID* – уникальный идентификатор объекта,

S – совокупность статических характеристик,

D – совокупность динамических характеристик,

F – совокупность вычисляемых характеристик,

Rel – множество связей (отношений) данного объекта с другими объектами.

Вычисляемые характеристики — это такие значения, которые не собираются непосредственно из информационных ресурсов, а рассчитываются исходя из значений статических и динамических характеристик.

$$F = \left\{ \left(f_1, \varphi_{f_1}\right), \left(f_2, \varphi_{f_2}\right), \dots, \left(f_p, \varphi_{f_p}\right) \right\}, \quad (2.5)$$

где f_j – наименование вычисляемой характеристики,

φ_{f_j} – формула или алгоритм, определяющий, как вычислять f_j на основе статических *(S)* и/или динамических полей *(D),*

$j = 1,2, \dots, p.$

Если f_j зависит от статического поля s_a и динамической характеристики $d_b(t)$, то указывается $f_j = \varphi_{f_j}(s_a, d_b(t))$. Такое представление вычисляемой характеристики делает процесс вычислений проверяемым на каждом шаге.

Множество связей (отношений) — это такие значения, которые позволяют описывать связи данного объекта с другими объектами в информационном пространстве.

$$Rel = \{(r_1, Obj_{j_1}), (r_2, Obj_{j_2}), \dots, (r_m, Obj_{j_m})\}, \quad (2.6)$$

где r_i — тип связи (например, «Автор», «Владелец», «Состоит в группе», «Цитирует»),

Obj_{j_i} — конкретный объект, с которым установлена данная связь,

$i = 1,2, \dots, m.$

Множество связей позволяет явно задавать иерархии, ассоциации, графы и структурные отношения, необходимые для интеллектуального анализа в социально и научно-технически значимых областях.

Однако, для аналитических задач нередко требуется группировать связи в семантические категории (или подклассы). Рассмотрим возможность введения «иерархии связей», где каждый тип r_i имеет собственную структуру, тогда каждая связь (*Rel*) может быть записана как:

$$Rel = (r_i, Obj_{j_i}) = ((Category_i, Subtype_i, Attributes_i), Obj_{j_i}), \quad (2.7)$$

где *Category* — общий класс отношения (например, «правообладание», «авторство», «расположение»);

Subtype — конкретизация в рамках категории (например, «владелец патента», «соавторство статьи», «территориальное расположение»);

Attributes — параметры или метаданные связи (срок действия авторства, тип доступа, условия использования).

В некоторых случаях *Attributes* могут содержать собственные динамические поля (например, дата начала или окончания действия лицензионных прав), а также вычисляемые показатели (например, «доля авторства в процентах»).

При необходимости более сложной организации можно иерархически вложить связи друг в друга:

$$Rel = \{(r_1, \{Obj_{j_1}, \dots, Obj_{j_n}\}), \dots, (r_m, \{Obj_{j_1}, \dots, Obj_{j_k}\})\} \quad (2.8)$$

Таким образом, каждый $r_i, i = 1, \dots, m,$ может связывать не два, а заранее неопределённое число объектов.

Например, «консорциум (r_i)» объединяет несколько организаций и авторов при работе над одним проектом. Это, в свою очередь, может быть отражено в виде графовых структур с узлами (объектами) и ребрами (связями), и дополнительными метками (семантическими категориями).

Детализировано аналитическая модель цифрового объекта выглядит следующим образом:

$$AObj = \langle \begin{matrix} ID, (s_1, \tau_1), \dots, (s_k, \tau_k), (d_1(t_1), t_1, \tau_1), \dots, (d_l(t_l), t_l, \tau_l), \\ (f_1, \varphi_{f_1}), \dots, \left(f_p, \varphi_{f_p}\right), \{(r_1, Obj_1), \dots, (r_m, Obj_m)\} \end{matrix} \rangle . \quad (2.9)$$

Предложенная аналитическая модель цифрового объекта соответствует принципу *иерархичности и модульности*, позволяет рассматривать каждый объект как (*ID* + статические характеристики, динамические характеристики) с «надстройками» (вычисляемые характеристики, связи между объектами).

Особенностями предложенной аналитической модели цифрового объекта являются следующие свойства:

1. **Ясная структура**: каждая компонента (ID, S, D, F, Rel) имеет собственное назначение и формальную интерпретацию (однозначное поле для идентификатора, чёткое разделение статических, динамических, вычисляемых характеристик и связей).

2. **Гибкость**: можно оперативно добавлять новые характеристики (как статические, так и динамические) или вычисляемые функции, не изменяя общую схему. Наличие связей *Rel* позволяет перейти к более сложным сценариям моделирования (графы, сети, онтологии), что актуально в междисциплинарных исследованиях.

3. **Удобство аналитических исследований**: вычисляемые характеристики *F* представлены формальными функциями/алгоритмами $\left(f_i, \varphi_{f_i}\right)$, что позволяет автоматически пересчитывать значения при изменении динамических характеристик (*D*).

4. **Модулярность**: наличие множества связей *Rel* даёт возможность интегрировать объект в сеть взаимосвязанных сущностей (авторы–публикации–патенты–организации), что особенно важно при решении комплексных научно-технических и социальных задач.

5. **Экономное использование вычислительных ресурсов**: разделение статических и динамических характеристик помогает системно обрабатывать обновления данных в режиме реального времени или с определённой периодичностью.

6. **Динамичность:** можно добавлять новые типы связей или дополнительные уровни иерархии (например, учитывать требования конкретной социально-экономической модели или узкоспециализированного научного проекта).

7. **Учет изменений состояний цифрового объекта:** контроль и анализ изменения характеристик (цитирования, рейтинги, патентная активность) с учётом изменения состояния объектов.

Введённая аналитическая модель цифрового объекта существенно расширяет возможности базовой информационной модели. Она позволяет отслеживать изменения, описывать разные типы отношений между исследуемыми объектами и использовать современные сетевые/графовые подходы к решению аналитических задач в реальном масштабе времени.

2.3 Аналитическая модель цифрового объекта для анализа комплексных цифрового информационных объектов

При проведении аналитических исследований встречаются ситуации, когда анализируемые объекты обладают вложенными объектами. Классический пример — статья, содержащая в себе список авторов (каждый из которых сам по себе объект), ссылки на литературу (каждая ссылка может быть объектом) и приложения (отдельные документы). При этом такие «вложения» могут рассматриваться как в качестве **отдельных** объектов, так и в качестве **подразделов** изучаемого объекта.

Введем понятие **комплексного цифрового информационного объекта** – объект, обладающий собственными характеристиками и набором вложенных объектов. Тогда **модель комплексного объекта** задается следующим образом:

$$CAObj = \langle ID, S, D, F, Rel_{complex} \rangle, \quad (2.10)$$

где $Rel_{complex} =$

$\{(r_1, \{AObj_{j_1}, \dots, AObj_{j_n}\}), \dots, (r_m, \{AObj_{j_1}, \dots, AObj_{j_k}\})\}$,

$r_i -$ тип связи,

$i = 1, \dots, m.$

Особенностью комплексной модели цифрового информационного объекта является то, что вычисляемые характеристики (*F*) рассчитываются не только на статических (*S*), динамических (*D)* и вычисляемых *(F)* характеристиках комплексного объекта $CAObj$, но и на аналогичных характеристиках вложенных объектов (S_{AObj}, D_{AObj}). Аналитическая цифровая модель является частным случаем комплексной информационной модели цифрового объекта.

Комплексная модель позволяет использовать *агрегирующие и комбинированные* функции для расчета сводных характеристик, таких как рейтинги и иные интегральные метрики, востребованные в задачах интеллектуального анализа научно-технических и социальных данных.

Потребность в агрегирующих функциях возникает при необходимости:

- Суммировать или подсчитывать числовые показатели (например, общее число публикаций всех подразделений организации).
- Вычислять усреднённые метрики (например, средний импакт-фактор статей, связанных с определённой лабораторией).
- Определять экстремальные значения (максимум или минимум; например, наибольшее количество заявок в рамках одной рабочей группы).

Пусть $LinkedSet\ (r_i, CAObj) = (r_i, \{AObj_{j_1}, \ldots, AObj_{j_n}\})$ – все объекты, связанные с $CAObj$ отношением r_i. В общем случае агрегирующая функция $A(x)$ для характеристики x может быть записана как:

$$A(x, CAObj, r_i) = agg(x(AObj_1), \ldots, x(AObj_n)), \quad (2.11)$$

где $x(AObj_{j_i})$ — значение интересующего признака (статического, динамического или вычисляемого) у объекта $AObj_{j_i}$,

agg — операция агрегирования (суммирование, среднее, минимум/максимум, конкатенация списков, более сложные операции),

$i = 1, \ldots, n$.

Модель комплексного цифрового информационного объекта $CAObj$ не ограничивается описанием одного объекта, а распространяется на иерархию или сеть взаимосвязанных объектов со своими уникальными идентификаторами и набором характеристик. Агрегирующие и вычисляемые функции позволяют консолидировать данные и получать интегральные показатели, упрощающие дальнейший интеллектуальный анализ данных. Особенностями предложенной модели являются:

1. Модульность: любой вложенный объект может быть использован для отдельных аналитических задач и при необходимости переиспользован.

2. Расширяемость: оперативное добавление новых связей (Rel) (организация → автор → патент → ... и т.д.).

3. Прозрачность вычислений: формулы агрегирования и зависимости между объектами однозначно записываются, что упрощает верификацию и воспроизводимость исследований.

Таким образом схема преобразований цифрового объекта представлена на рисунке (Рисунок 2.1), где:

φ – функции/алгоритмы расчета характеристик,

r – типы связей между объектами,

$A(x)$ – агрегирующие и комбинированные функции.

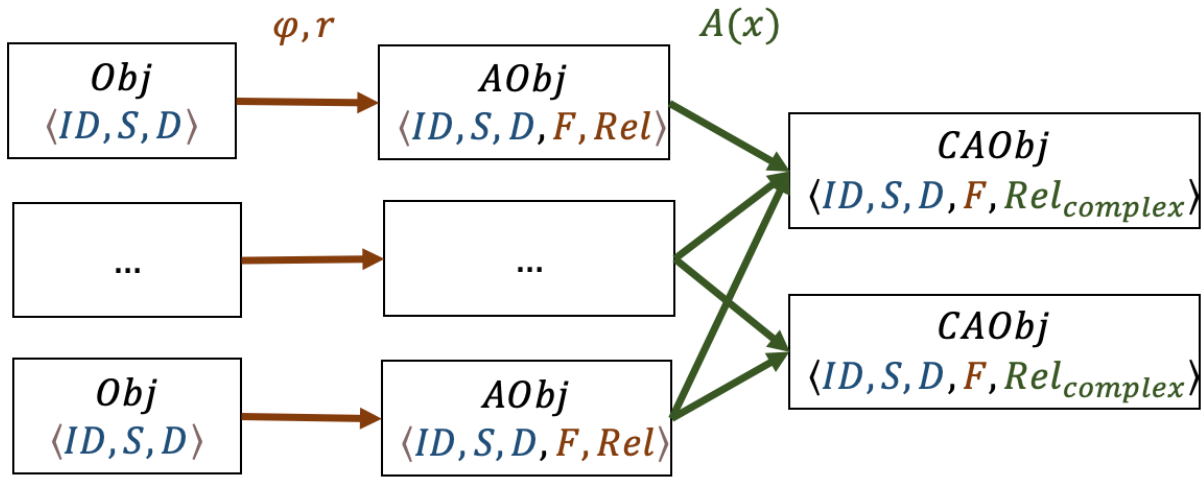


Рисунок 2.1 Схема преобразований цифрового объекта

Предложенный подход к построению комплексных цифровых информационных объектов и применению агрегирующих функций формирует

прочную концептуальную основу для реализации гибких и динамичных систем хранения и анализа данных в научно-технических и социально значимых областях.

2.4 Преимущества использования модели комплексного цифрового информационного объекта в проведении аналитических исследований

Современные технологии интеллектуального анализа данных (Data Mining, машинное обучение, статистические методы и т. д.) во многом полагаются на *корректно структурированную и контекстно богатую* информацию об исследуемых объектах. Предложенная модель комплексного цифрового информационного объекта с разделением на статические, динамические и вычисляемые характеристики, а также поддержкой сложных связей способна повысить **качество и эффективность** аналитических процедур. Рассмотрим некоторые ключевые аспекты применения комплексной модели в задачах подготовки, обработки и представления данных.

Процесс первичной подготовки данных (Data Preparation) включает в себя очистку, интеграцию и приведение их к единому формату. В условиях, когда разные источники информации (научные статьи, патенты, социальные сети, реестры организаций и т. д.) содержат объекты со схожими, но по-разному названными или хранящимися признаками, унифицированная структура *CAObj* позволяет:

1. **Сопоставлять** и **сливать** записи об одних и тех же реальных сущностях, опираясь на единый механизм идентификации *ID* или семантические связи *Rel*.
2. **Разделять** статические и динамические показатели, что упрощает обнаружение и коррекцию аномалий:
   - Статические поля (например, «ФИО автора») проверяются на согласованность (правильное форматирование, отсутствие дубликатов).
   - Динамические поля (например, «число цитирований») требуют дополнительной проверки временных аномалий (скачков).

3. **Сохранять** вычисляемые характеристики F в виде **алгоритмических описаний** φ_f, чтобы при изменении исходных данных не требовалось вручную пересчитывать производные поля.

Таким образом, комплексная модель является **инструментом** стандартизации и обогащения данных, повышающим надёжность исходной выборки для интеллектуального анализа.

В задачах обработки данных (классификация, регрессия, кластеризация и т. д.) **качество модели** во многом определяется **информативностью признаков**, благодаря чёткому разделению характеристик на статические, динамические и вычисляемые:

1. **Статические** характеристики обеспечивают базовый контекст (тип объекта, его постоянные свойства).

2. **Динамические** могут **трансформироваться** в новую форму (например, извлекаются статистики из временного ряда: среднее значение, дисперсия, тренд), дополняя модель признаками «скорости изменения», «сезонных колебаний» и т. д.

3. **Вычисляемые** характеристики позволяют генератору признаков автоматизировать получение комплексных метрик (индекс Хирша, рейтинги, специфические баллы и т. д.), которые являются основой для алгоритмов машинного обучения.

Важной частью любого анализа является представление итоговых данных в наглядном виде (аналитические панели, отчёты, диаграммы). Применение унифицированной структуры *CAObj* даёт следующие преимущества:

1. **Автоматическая генерация** визуализаций по типам связей (ролевых отношений). Например, панель для визуализации структуры организации: «Организация → Отделы → Сотрудники» (и метрик, связанных с ними).

2. **Динамическое обновление**: за счёт того, что данные о связях *Rel* и вычисляемые показатели *F* могут обновляться по событиям, аналитические панели не требуют ручной перестройки.

3. **Гибкая детализация**: благодаря вложенным объектам пользователь может «погружаться» на любой уровень иерархии (от организации к конкретному патенту/статье, далее к авторам и статистике).

Преимущества и перспективы применения предложенной комплексной цифровойы информационной модели:

1. **Универсальность**: формальная структура *CAObj* не привязана жёстко к одному типу данных, поддерживает как научно-технические (статьи, патенты), так и социальные (профили, сообщества) объекты.

2. **Модульность в анализе**: сложные эксперименты (например, сетевой анализ соавторства) легко запускать, опираясь на готовые связи *Rel*, а временные аспекты — на *D*.

3. **Упрощённая поддержка**: при добавлении новых типов объектов (например, технологические отчёты или финансовая отчётность) необходимо **расширить** модель (добавить новые виды признаков, связей, формул) без полной перестройки анализа.

4. **Интеграция с разными технологиями**: модель может быть реализована в реляционных, NoSQL или графовых системах, причём механизмы динамического расчёта позволяют встраиваться в архитектуры потоковой аналитики, гибридных хранилищ и больших данных (Big Data).

5. **Гибкость и адаптивность:** модель может «подстраиваться» под новые требования без тотальной перестройки существующей системы.

6. **Сохранение преемственности:** учет характеристики времени даёт возможность корректно работать с объектами, созданными в разное время и по разным правилам.

7. **Удобство для анализа:** аналитические алгоритмы могут получать более полную или современную структуру данных, не обрывая доступа к историческим записям.

Таким образом, предлагаемая модель комплексного цифрового информационного объекта **упрощает** процедуру подготовки данных к интеллектуальному анализу и **повышает** точность, надёжность и воспроизводимость результатов. Её использование актуально при построении **целенаправленных** (ad-hoc) аналитических решений и при создании **больших интегрированных** платформ (объединяющих научные репозитории, соцсети, государственные реестры), где требуются многогранные анализы, учитывающие неоднородные объекты, их временные изменения и множественные связи.

Таким образом, **механизм эволюции и реструктуризации** модели комплексного цифрового информационного объекта обеспечивает её актуальность, жизнеспособность и устойчивость к изменениям, которые неизбежно возникают при решении комплексных научно-технических и социальных задач на основе разнородных данных.

ВЫВОДЫ ПО ГЛАВЕ 2

1. Предложена базовая информационная модель цифрового объекта, представленная уникальным идентификатором объекта и множествами статических и динамических характеристик, формирующая первичное описание объекта и достаточная для отслеживания изменений его состояния во времени.

2. Предложена модель комплексного цифрового информационного объекта, включающая статические, динамические и вычисляемые характеристики, а также механизмы поддержки сложных связей. Модель упрощает процедуры подготовки данных к интеллектуальному анализу; повышает точность, надёжность и воспроизводимость результатов; обеспечивает повышение качества аналитических исследований при работе с разнородными данными.

3. Разработанная систематизация характеристик обеспечивает универсальность, стандартизацию, надёжность преобразования данных для интеллектуального анализа, за счет статических характеристик, фиксирующих

базовые свойства объекта; динамических характеристик, включающих параметры изменения объектов; вычисляемых характеристик, реализующих расчет комплексных метрик.

4. Модель комплексного цифрового информационного объекта обладает рядом существенных преимуществ, включая универсальность применения к разнородным данным; модульность представления данных, обеспечивающую поддержку сложных аналитических сценариев; масштабируемость для включения новых типов объектов; интеграционную совместимость с современными технологиями хранения и обработки данных, а также адаптивность к изменяющимся требованиям при сохранении преемственности данных.

ГЛАВА 3 МЕТОДЫ ПРЕОБРАЗОВАНИЯ ДАННЫХ ИЗ РАЗНОРОДНЫХ ИНФОРМАЦИОННЫХ РЕСУРСОВ

3.1 Методика наполнения моделей цифрового объекта

Направление автоматизированного извлечения и обогащения данных из неструктурированных источников активно развивается в последние годы, ключевыми направлениями являются: программное выделение данных, обработка естественного языка, геоанализ текстов, методы векторизации, а также извлечение химических элементов и физических величин из текста.

В области **автоматизированного выделения данных** из разнородных источников заметно внимание к интеграции больших языковых моделей и распределённых вычислений. Например, а работе [45] представлен обзор методов извлечения данных из неструктурированных документов с акцентом на применение гибридных систем, сочетающих правила и обучение. Отмечается, что комбинированные подходы позволяют добиться высокой точности на гетерогенных данных, таких как сканы документов и веб-страницы. В зарубежных работах 2023 года прослеживается тренд использования трансформерных моделей для парсинга сложных форматов (таблиц PDF, вложенных структур XML) – благодаря self-attention механизму эти модели могут учитывать контекст на уровне всего документа [46].

По направлению **обработки естественного языка** и **именованных сущностей** современные исследования сосредоточены на использовании предобученных больших языковых моделей (БЯМ) для задач информационного извлечения. Так, в работе [47] продемонстрировано, что модели типа GPT-3 и Llama2, дообученные на специализированных корпусах, способны извлекать структурированные факты из научных текстов без ручной разметки, генерируя сразу JSON записи. Другое актуальное направление – обеспечение интерпретируемости и контроль качества при применении БЯМ. Авторы работы [46] отмечают, что хотя крупные модели показывают высокие результаты на стандартных NLP задачах, они могут галлюцинировать несуществующие факты, поэтому для надёжного извлечения информации

предлагаются гибридные схемы, где БЯМ генерирует кандидатов, а затем более простой алгоритм их верифицирует [48]. Проводятся работы по выделению специфических сущностей, таких как химические наименования и физические величины.

Для задачи **геоанализа** текстов (выявление географических названий и привязка их к координатам) в последние года ведутся работы по улучшению геокодирования на основе нейросетевых моделей и обработке естественного языка [49, 50]. Например, в работе [51] описан один из вариантов реализации распознавания пространственных данных из текстов на естественном языке, основанный на лексико-синтаксическом анализе текстов, что требует использования специальных грамматик и словарей. Распознавание пространственных данных проводится для их последующего геокодирования и визуализации.

Продолжают развиваться методы **векторизации и эмбеддингов** текста. В 2020–2024 гг. появились новые модели языковых эмбеддингов, ориентированные на более короткие вектора при сохранении смысловой емкости, исследуются комбинирование текстовых и визуальных эмбеддингов для мультимодального анализа научных статей (например, одновременное представление содержания текста и связанных с ним графиков в общем векторном пространстве). В обзоре [52] отмечают, что встроенные представления текста находят применение в задачах детектирования тематики документов, поиске аномалий и даже генерации гипотез, особенно при наличии больших массивов литературы по определённой тематике.

Одной из ключевых задач при работе с текстовыми данными научных публикаций является **автоматическое распознавание и классификация именованных сущностей** (англ. Named Entity Recognition, NER), а также извлечение других семантически значимых элементов (терминов, фактов, взаимоотношений) [53]. Данная задача лежит в области *обработки естественного языка (Natural Language Processing, NLP)* – раздела

искусственного интеллекта и компьютерной лингвистики, изучающего компьютерный анализ и синтез текстов на человеческих языках.

Алгоритмы машинного обучения в базовом виде не могут напрямую работать с неструктурированным текстом, поэтому предварительная обработка текстовых данных крайне важна. NLP обеспечивает инструменты для приведения текста к виду, пригодному для анализа: токенизации (разбиения на слова), нормализации (приведения слов к базовой форме), удаления неинформативных элементов (стоп-слов, знаков), определения частей речи, синтаксического разбора и т.д. В совокупности эти шаги позволяют преобразовать «сырые» текстовые данные в форму, с которой могут эффективно работать алгоритмы анализа данных.

NLP находится на стыке дисциплин искусственного интеллекта и лингвистики и используется при решении следующих задач:

- машинный перевод;
- классификация текстов;
- извлечение именованных сущностей (NER);
- извлечения фактов и отношений (relation extraction);
- создание вопросно-ответных и диалоговых систем (чат-боты);
- саммаризация;
- машинное обучение.

Выделение именованных сущностей (NER). Цель NER – найти в тексте упоминания объектов реального мира и определить их тип (персона, организация, локация, дата, термин и т.д.). К примеру, в предложении «Google объявила о сотрудничестве с UNICEF в Африке в 2025 году» NER должна выделить «Google» (Organization), «UNICEF» (Organization), «Африке» (Location), «2025 год» (Date). Выделенные сущности служат отправной точкой для более сложного анализа: выявления отношений между ними, построения онтологий, графов знаний и др. [54, 55] В контексте научных публикаций NER позволяет выявлять авторов, организации, химические соединения, названия

приборов, географические объекты и другие сущности, имеющие значение для исследования связей и тенденций.

В последние годы появилось множество работ, посвящённых методам NER и их применению в различных предметных областях. Например, в медицине NLP используется для извлечения клинической информации из записей врачей. В работе [56] была показана возможность классификации неврологических исходов пациентов по данным медицинских записей с помощью NLP-модели, обученной извлекать результаты клинических обследований. В российском исследовании [57] предложены подходы к автоматическому извлечению структурированных данных (медицинских параметров) из неструктурированных медицинских текстов с использованием словарей и шаблонов. Другой пример – применение чатботов для ответов на вопросы абитуриентов, где используются инструменты морфологического анализа русского языка (NLTK, Pymorphy2, Mystem, Natasha) для понимания запросов пользователей [58]. Также внедряются системы автоматического извлечения атрибутов из юридических документов (договоров) с применением библиотек NLP (Natasha, spaCy) и специализированных инструментов для поиска телефонных номеров [59, 57]. Актуальные задачи включают и выявление пропагандистских приёмов в текстах – так в работе [60] разработали систему на основе моделей BERT и RoBERT для обнаружения признаков пропаганды в сообщениях, предварительно очистив тексты (удаление цифр, спецсимволов) и применив предобученные языковые модели.

На основе обзора научных трудов можно выделить основные этапы предобработки текста в задачах NLP: **нормализация**, **токенизация**, **удаление стоп-слов**, **стемминг/лемматизация**, **векторизация**.

Нормализация включает приведение текста к единому регистру, удаление пунктуации, чисел (или преобразование их к единому формату), удаление лишних пробелов. Цель – унифицировать написание, избавившись от элементов, не несущих полезной нагрузки, чтобы алгоритмы не считали, к примеру, слова с заглавной и строчной буквы разными.

Токенизация – разбиение текста на токены (слова, числовые значения, знаки). После этого удаляют **стоп-слова** – наиболее частотные и малоинформативные слова (предлоги, союзы, частицы типа «и», «в», «не» для русского языка). Это сокращает объём данных и фокусирует алгоритм на существенных терминах.

Стемминг – приведение слов к их основе (стемы) путём отбрасывания окончаний и суффиксов. Например, слова «научный», «научного», «науке» имеют общий стем «наук». Стемминг выполняется по заранее заданным правилам отсечения окончаний и суффиксов.

Лемматизация – альтернативный подход, где каждое слово приводится к нормативной словарной форме (лемме): для существительных – именительный падеж единственного числа, для глаголов – инфинитив и т.д. Например, «учёные» → «учёный», «выработанных» → «выработать». Лемматизация требует знания о части речи слова и часто использует словари или обученные модели. Стемминг проще и быстрее, но может давать неоднозначные основы (например, «стали» → «стал»), тогда как лемматизация точнее, но сложнее в реализации. В практике NLP часто применяются готовые лематизаторы (для русского языка – pymorphy2, Natasha и др.).

Векторизация текста. Поскольку большинство алгоритмов машинного обучения работают с числовыми признаками, необходимо представить текстовые данные в виде векторов чисел. Классический подход – **«мешок слов» (bag-of-words)**, при котором формируется словарь уникальных слов всего корпуса текстов, и каждому документу ставится в соответствие вектор размерности словаря, где на позиции слова стоит количество его вхождений в документ. Такие вектора получаются очень разреженными (в тексте используется лишь малая часть словаря) и высокой размерности. Более продвинутый вариант – TF-IDF-признаки, учитывающие не только частоту слова в документе (TF), но и обратную частоту по корпусу (IDF), что понижает вес общеупотребительных слов. Однако и bag-of-words, и TF-IDF не учитывают порядок слов и контекст использования.

Для учёта семантики слов разработаны методы **дистрибутивных векторных представлений слов** (*word embeddings*). Согласно гипотезе распределения Зеллига Харриса: «слова, встречающиеся в похожем контексте, имеют схожее значение» [61]. На этой идее основаны нейросетевые модели типа **Word2Vec** [62] и **GloVe** [63], которые обучаются на большом корпусе текста и представляют каждое слово в виде плотного числового вектора фиксированной размерности (например, 300-мерного). Близкие по смыслу слова оказывается близки и в векторном пространстве (например, векторы «наука» и «исследование» будут ближе друг к другу, чем к «магазин»). Word2Vec предлагает два подхода обучения: Skip-gram (предсказание контекстных слов по данному) и CBOW (предсказание слова по контексту). GloVe (Global Vectors) – альтернативная модель, оптимизирующая представления слов на основе статистики их совместной встречаемости в корпусе. Результирующие вектора улавливают разнообразные отношения между словами (например, операции над ними могут выявлять аналогии: **vec**(«король») - **vec**(«мужчина») + **vec**(«женщина») **vec**(«королева»)).

Современные языковые модели, такие как **BERT** (Bidirectional Encoder Representations from Transformers) [64], формируют контекстуальные представления слов, то есть один и тот же словоформ может иметь разные вектора в зависимости от окружения (модель учитывает весь текст слева и справа). BERT предложен в 2018 году командой Google и обучен на огромном корпусе (англоязычная Википедия и книги) на задаче восстановления пропущенных слов и предсказания следующего предложения. BERT установил новый стандарт качества для многих NLP-задач и с конца 2019 года используется Google для понимания поисковых запросов на естественном языке. Появились специализированные версии, например, **BioBERT** – модель BERT, дообученная на биомедицинских текстах, которая показала улучшение в задачах анализа биомедицинской литературы [65].

Векторные представления слов и предложений легли в основу множества прикладных решений: от поиска информации и кластеризации документов до

чат-ботов и машинного перевода. Они используются, в частности, для задач извлечения ключевых слов и тематической кластеризации текстов научных статей.

Для подготовки текстовых данных к автоматическому анализу необходимо:

1) извлечь именованные сущности (организации, географические названия, предметные термины и т.п.),

2) извлечь ключевые фразы, характеризующие содержание,

3) представить тексты в векторной форме для количественного анализа (например, для кластеризации по тематике).

Выделение ключевых слов из текста стало особенно актуальной задачей из-за лавинообразного роста объемов информации – автоматизация этого процесса позволяет оперативно обрабатывать большие корпуса документов. Существуют специализированные методы для автоматического *keyword extraction*. Одним из них является алгоритм **YAKE!** (Yet Another Keyword Extractor) – полностью автоматический метод извлечения ключевых слов, не требующий обучающих данных и не зависящий от языка или домена [66]. YAKE основан на статистических характеристиках текста: оценивает частоту слов, их позицию, распределение по документу и на основе локальных признаков ранжирует кандидаты на роль ключевых слов.

Методика наполнения моделей цифрового объекта состоит из трёх основных этапов

1. **Извлечение основных данных** из переданного объекта (документа) – получение статических и динамических характеристик. К базовым параметрам относятся: название публикации, авторы, аффилиации авторов, аннотация, ключевые слова (заданные авторами), основной текст статьи, информация о финансировании, список литературы, название журнала, год публикации, ссылки (DOI, URL).

2. **Насыщение данных** – вычисление и добавление новых сведений на основе извлечённой информации (ключевые слова текста, химические

упоминания, геоданные и т.п.). Используя извлечённые базовые данные (название, текст, список литературы и пр.), к ним добавляются новые сведения, получаемые из внешних источников. В работе рассматриваются несколько направлений насыщения: выделение из полного текста ключевых слов и фраз; распознавание в тексте физических элементов и упоминаемых единиц измерения с их нормализацией; обработка аффилиаций авторов для определения геолокации (координат) организаций и приведение названий стран к единому виду; на основе стран аффилиаций – определение, в какие международные научные альянсы входят соответствующие страны. Насыщение данных позволяет расширить возможности анализа: например, ключевые слова текста указывают на основные темы статьи, а определение координат организаций даёт возможность географического анализа распределения исследований.

3. **Организация хранения** результирующих данных о цифровом объекте. Собранные на предыдущих шагах сведения организуются в единую структуру (например, JSON документ определённого формата) и сохраняются в целевое хранилище. Структурирование предполагает приведение данных к заранее определённой схеме, включающей поля для всех типов информации (метаданные, текст, ключевые слова, таблицы, изображения, химические сущности, единицы и пр.). Подготовленный структурированный файл поступает в распределённое хранилище данных для построения визуализаций или дальнейшего анализа.

3.2 Методы извлечения данных из информационных ресурсов

Извлечение данных из различных информационных ресурсов является сложной нетривиальной задачей в связи с тем, что данные хранятся в разных форматах и имеют различные структуры. Перечисленные подходы и методы формируют отдельное направление в области компьютерных наук – программное выделение данных.

Программное выделение данных – это процесс автоматического извлечения структурированной информации из неструктурированных или

полуструктурированных источников (текстовых документов, веб-страниц, электронных таблиц и т.д.) [67]. Этот процесс реализуется с помощью алгоритмов и программных средств, анализирующих структуру исходных данных, выделяющих необходимые элементы и преобразующих их в структурированный формат, пригодный для дальнейшего анализа.

Программное выделение данных обладает рядом преимуществ. Во-первых, автоматизация сбора и обработки информации значительно сокращает время, затрачиваемое на эти процессы, что позволяет быстрее принимать решения и оперативно реагировать на изменения среды [68]. Во-вторых, применение алгоритмов уменьшает количество ошибок по сравнению с ручной обработкой данных. В-третьих, автоматические методы дают возможность анализировать большие объёмы данных, что затруднительно или невозможно вручную.

Существуют ограничения программного выделения данных. Например, алгоритмы могут испытывать трудности при обработке источников низкого качества (нечеткие изображения, отсканированные документы и т.п.). Также при недостаточной настройке программа может извлекать некорректные или нерелевантные данные. Кроме того, разные типы входных файлов требуют использования разных методов и инструментов программного выделения данных [69]. Так, для извлечения данных из веб-страниц обычно применяются методы разбора HTML-разметки, тогда как обработка электронных таблиц предполагает использование специализированных библиотек для чтения табличных форматов. Если формат или структура входного файла изменяются, алгоритмы извлечения необходимо адаптировать, чтобы они продолжали работать корректно.

Таким образом, программное выделение данных является важнейшим инструментом для автоматизации обработки информации из разнородных источников, которое способно существенно ускорить сбор и анализ данных, что в конечном счёте помогает принимать более обоснованные решения на основе актуальных данных [69, 70]. Для эффективного использования данных

методов следует учитывать специфику каждой задачи и подбирать соответствующие алгоритмы и инструменты автоматизированного извлечения данных [71, 72, 73, 74].

Рассмотрим типы данных и соответствующие им методы автоматизированной обработки. К числу основных видов данных относятся текстовые документы, веб-документы и изображения. Видеоматериалы и аудиоматериалы также являются одними из ключевых типов данных при анализе информации из информационных источников. Методы обработки таких типов данных в работе не рассматриваются, что обусловлено в том числе наличием существенного количества программных средств конвертации аудиозаписей, видеозаписей в текст и изображения соответственно. Кроме того, необходимо решать задачи идентификации сущностей в тексте, на основе технологий обработки естественного языка, выделение гео-данных, векторизацию текстов, а также выделение из текстов специальных сущностей – химических элементов и единиц измерения.

3.2.1 Текстовые документы

Текстовые файлы представляют данные в виде последовательности символов и могут иметь различные форматы. Ниже перечислены наиболее распространённые виды текстовых документов:

1. TXT – текстовые файлы с расширением .txt являются наиболее простым типом текстовых файлов. Они могут содержать текст в любом формате и часто используются для хранения документов, конфигурационных файлов, скриптов и других данных.

2. CSV (Comma Separated Values) – текстовые файлы, которые используются для хранения табличных данных. Они содержат данные, разделенные запятыми или другими символами, такими как точка с запятой или табуляция. CSV файлы часто используются для обмена данными между различными приложениями.

3. JSON (JavaScript Object Notation) – текстовый формат, используемый для хранения и обмена данными, файлы содержат данные в виде

пар «ключ-значение», разделенных запятыми и заключенных в фигурные скобки. JSON читается и генерируется множеством языков программирования и часто используется для конфигураций и обмена данными между системами.

4. XML (Extensible Markup Language) – это текстовый формат, используемый для представления данных в структурированном формате. Он используется в основном для обмена данными между различными системами и приложениями. XML файлы содержат данные в виде тегов и атрибутов, которые определяют структуру документа.

5. HTML (Hypertext Markup Language) – текстовый формат, используемый для создания веб-страниц. Он содержит разметку, которая определяет структуру и содержание страницы, а также ссылки на другие ресурсы, такие как изображения и стили CSS.

6. PDF документ (Portable Document Format) – это формат файла, который используется для представления документов в виде, не зависящем от программного и аппаратного обеспечения. Формат PDF позволяет включать в файл текст, растровую и векторную графику, формы, мультимедиа и другие элементы, обеспечивая единое визуальное представление документа независимо от платформы. PDF может содержать встроенные шрифты, сценарии (JavaScript), 3D-графику и др. Файлы PDF часто используются для финального представления документов, поскольку сохраняют оформление вне зависимости от среды просмотра.

7. DOC(X) документ – формат файла, используемый для хранения документов в Microsoft Word. В целом, формат .docx является одним из наиболее распространенных форматов файлов для хранения документов, и он широко используется в офисных приложениях.

8. XLSX – это формат файла, используемый для хранения электронных таблиц в Microsoft Excel. В целом, формат XLSX является одним из наиболее распространенных форматов файлов для хранения электронных таблиц, и широко используется в офисных приложениях, финансовом учете и других областях.

Перечисленные форматы охватывают большинство встречающихся типов текстовых данных. Каждый формат имеет свои особенности организации информации и инструменты для автоматизированной обработки. Для программной работы с распространёнными текстовыми форматами существуют готовые программные библиотеки. Для CSV существуют парсеры, например, модуль csv в Python. Форматы JSON и XML поддерживаются соответствующими модулями (json и xml в Python), упрощающими разбор и генерацию этих документов. Для HTML необходимо разрабатывать отдельные сборщики данных для обеспечения высоких показателей полноты и точности сбора.

В работе с научно-технической и социально значимой информацией основными источниками текстовой информации выступают текстовые файлы формата (PDF, DOCX) и различные веб-страницы.

3.2.2 Веб-документы

В настоящее время одним из ключевым источником информации являются веб-документы – HTML-страницы, размещенные в глобальной сети Интернет. HTML-документ представляет собой текстовый файл с разметкой на языке HyperText Markup Language, задающей структуру веб-страницы (заголовки, абзацы, списки, ссылки, изображения и т.д.) [75]. Кроме HTML-разметки, веб-страница может содержать встроенные или подключаемые **CSS**-стили (каскадные таблицы стилей), определяющие внешний вид элементов, а также **JavaScript**-скрипты, отвечающие за динамическое изменение содержимого и взаимодействие с пользователем.

Браузер, получая HTML-документ, строит на его основе **DOM-дерево** (Document Object Model) – иерархию объектов, отражающую структуру страницы. Узлами DOM являются элементы HTML. Вложенные теги образуют отношения родитель–потомок. Например, корневой элемент <html> содержит вложенные <head> и <body>, внутри <body> могут находиться теги абзацев <p> , заголовков <h1> и т.д. На рисунке (Рисунок 3.1) схематично показан пример DOM-дерева для простого HTML-документа.

```
<!DOCTYPE html>
<html lang="en">
<head>
    <meta charset="utf-8"/>
    <link rel="icon" href="%PUBLIC_URL%/imgs/favicon.svg"/>
    <meta name="viewport" content="width=device-width, initial-scale=1"/>
    <meta name="theme-color" content="#000000"/>
    <meta
            name="description"
            content="Пример HTML"
    />
    <title>Пример</title>
</head>
<body>
<div class="block" id="block-content">
    <div class="block-content">
        <article role="article"><h1><span>Структура HTML-кода</span></h1>
            <div class="info">
                <div><p>Если открыть любую веб-страницу, то она будет содержать в себе типичные элементы,
                    которые не меняются от вида и направленности сайта. В примере 4.1 показан
                    код простого документа, содержащего основные теги.</p>
                    <p class="exampleTitle2">Пример Исходный код веб-страницы</p>
                </div>
            </div>
        </article>
    </div>
</div>
</body>
</html>
```

Рисунок 3.1 Пример DOM-дерева

Веб-документы изначально являются слабоструктурированными: их основное назначение – представление информации пользователю, а не её машинная обработка. Тем не менее существуют методы программного сбора данных с веб-ресурсов, получившие названия *парсинг веб-страниц* или *веб-скрапинг (web scraping)* [17, 76, 77, 78]. Автоматизация сбора данных достигается с помощью **программных агентов** – специализированных программ, которые по расписанию или по запросу пользователя посещают веб-страницы и извлекают с них нужную информацию. Программный агент обычно осуществляет HTTP-запрос к URL-адресу страницы, получает HTML-код и затем анализирует его, извлекая определённые элементы [79, 80].

Структура URL-адреса в общем виде выглядит следующим образом:

`схема://[логин:пароль@]хост[:порт]/URL-путь[?параметры][#якорь]`

В данной записи используются следующие элементы:

1. Схема – это схема обращения к ресурсу; в большинстве случаев имеется в виду сетевой протокол (http или https).
2. Логин – это имя пользователя, используемое для доступа к ресурсу.
3. Пароль – это пароль указанного пользователя.

4. Хост – это доменное имя хоста в системе DNS или IP-адрес хоста в форме четырёх групп десятичных чисел, разделённых точками; числа – целые в интервале от 0 до 255.

5. Порт – это порт хоста для подключения.

6. URL-путь – это уточняющая информация о месте нахождения ресурса.

7. Параметры – это строка запроса с передаваемыми на сервер параметрами, следуют после символа «?». Параметр имеет структуру «название=значение». Если URL-адрес содержит несколько параметров, то они должны быть записаны через символ «&».

8. Якорь – это идентификатор элемента в HTML-разметке с предшествующим символом «#». Якорь может иметь значения как заголовка внутри документа, так и атрибут id элемента HTML-разметки. По такой ссылке браузер откроет страницу и переместит окно к указанному элементу.

Для извлечения данных из HTML-разметки веб-страницы часто используется язык запросов **XPath** (XML Path Language) [78].

XPath позволяет задать путь к элементу или набору элементов в XML/HTML-документе по именам тегов, атрибутам, позициям и другим характеристикам. В сочетании с библиотеками разбора HTML (такими как lxml для Python) XPath значительно облегчает извлечение информации.

Следует учитывать, что автоматизированный сбор данных с веб-сайтов может сталкиваться с ограничениями. Веб-серверы ориентированы на обслуживание интерактивных пользователей, а не массовые запросы от программных агентов. Интенсивный веб-скрапинг может создавать повышенную нагрузку на информационный ресурс, вызывая дополнительные затраты трафика, энергопотребления и даже сбои в работе серверов. Администраторы вебсайтов внедряют защиту от несанкционированного сбора данных: блокируют частые запросы, требуют решения капчи, используют динамическую подгрузку контента через JavaScript, что затрудняет работу простых скраперов. Кроме того, автоматический сбор данных может

затрагивать вопросы авторского права и конфиденциальности, особенно если извлекаются персональные данные пользователей.

Некоторые современные веб-сервисы имеют ограничения от веб-скрапинга и при работе агента по автоматическому сбору данных необходимо позаботиться о том, чтобы веб-сервис, на котором агент собирает данные, не распознал его работу [39, 81, 82, 83]. Также информационные источники часто используют JavaScript для генерации HTML-кода страницы, в данном случае невозможно получить доступ к контенту ресурса напрямую через запрос с URL-адресом, это необходимо делать через веб-браузер, где JavaScript выполняется при загрузке веб-страницы.

Существует сложность в сборе данных из-за наличия навигационных элементов (заполнение полей данными, скроллинг, клики мыши). Часть требуемой информации может содержаться в узлах-элементах, которые не отображают информацию без направленного обращения к ним.

Программное моделирование поведения человека помогает избежать блокировки пользователя, выявления программного агента и осуществление навигационных действий. Моделирование осуществляется с помощью программного управления браузером и может включать:

1. Добавление **случайных задержек** и эмулирование **навигационные действий** (прокрутку, клики), чтобы не выглядеть как скрипт, работающий слишком быстро и последовательно случайных и необходимых навигационных действий на веб-странице.

2. Отправление запросов через различные прокси-серверы. Можно создать набор IP-адресов и распределить между ними запросы к источнику.

3. Изменение данных пользовательских агентов (информация в заголовке User-Agent).

4. Подключение управляемого планировщика задач, который распределяет программные агенты в соответствии с выбранным расписанием на определенные задачи.

Для эмуляции полноценного поведения пользователя (включая выполнение JavaScript-кода, заполнение форм, нажатие кнопок) применяются инструменты автоматизации браузера, такие как **Selenium WebDriver** [84]. Инструмент имеет объектно-ориентированный API и набор клиентских библиотек. Selenium позволяет программно управлять браузером (Chrome, Firefox и др.): открывать страницы, вводить текст, нажимать на элементы. В сочетании с парсерами HTML (BeautifulSoup, lxml) это позволяет извлекать даже те данные, которые появляются на странице динамически (после выполнения скриптов или действий пользователя). Selenium часто используется, когда веб-страница защищена от простого доступа или когда данные загружаются постепенно (бесконечная прокрутка, ленивые загрузки).

Таким образом, технологии программного выделения данных из веб-документов включает понимание структуры HTML, использование языков запросов (XPath, CSS-селекторы), а также методов обхода ограничений. Предварительное знакомство с нормативными аспектами обеспечивает этичное использование технологий web scraping.

На основе, предложенной в подразделе 2.1 базовой модели цифрового объекта разработана стандартизованная структура данных, в которой объединяются все извлечённые и обогащённые сведения на примере научной публикации. Эта структура реализована в формате JSON и приведена в обобщённом виде в таблице (Таблица 3.1.). Выбор формата JSON обусловлен тем, что он поддерживает вложенность структур, легко читается человеком и обрабатывается программно [85].

Таблица 3.1 Структура данных научных публикаций

Ключ	**Содержание**
title	название публикации
published	тип публикации, журнал, дата публикации, язык, DOI, библиографическое описание
authors	список авторов (каждый автор: имя, уникальный идентификатор при наличии, электронная почта при наличии)

affiliations	список аффилиаций авторов: названия организаций, подразделений, страна, координаты организации (широта, долгота)
countries	список стран, к которым относятся организации авторов (основные страны аффилиаций)
abstract	аннотация
keywords	ключевые слова публикации (заданные авторами)
full_text	данное поле является составным и включает в себя другие: plain - полный текст публикации keywords - ключевые слова текста публикации (по наибольшей встречаемости), tables - выделенные таблицы, images – выделенные изображения, alliances - альянсы стран (в какие межгосударственные союзы входят страны организаций авторов), chemical_units - химические элементы, measurement_units - единицы измерения и их величины
acknowledgements	благодарности (если указаны в публикации)
funding	сведения о финансировании исследования (гранты, контракты, если указаны)
references	библиографический список источников, цитируемых в публикации
updated_at	дата и время последнего обновления информации о публикации
created_at	дата и время первоначального создания записи о публикации
labels	метки или категории, присвоенные публикации (например, принадлежность к определённому тематическому кластеру)

Таблица отражает данные, которые подлежат автоматизированному сбору, обработке и насыщению при работе с данными по научным публикациям.

Часть перечисленных полей (название, библиографические данные, авторы, аффилиации, аннотация, ключевые слова авторов, полный текст, список литературы) извлекается непосредственно из документа. Данные о

благодарностях и финансировании также извлекаются из текста при наличии. Поля updated_at и created_at позволяют отслеживать актуальность информации, а labels предназначены для хранения результатов дополнительной классификации записи (например, отнесение статьи к определённой тематической категории или кластеру).

Остальные поля (keywords внутри full_text, tables, images, alliances, chemical_units, measurement_units) заполняются на основе методов и программных средств насыщения данных.

Разработанная структура данных обеспечивает целостное представление всей информации о публикации, необходимой для последующего многостороннего анализа и пригодна для хранения в NoSQL-базах (документоориентированных), поскольку представляет собой единый документ JSON.

При извлечении данных из PDF-документов первым этапом происходит конвертация в HTML-формат. Это необходимо для упрощения процедур парсинга, так как HTML легче анализировать с помощью существующих инструментов (XPath, BeautifulSoup и пр.), чем разметку PDF. Для конвертации может быть использована библиотека PyMuPDF – загружается каждая страница PDF и сохраняется в HTML-представление. Переход к HTML позволяет сохранить пространственное расположение элементов, что важно для понимания структуры статьи (разделы, подписи к рисункам и таблицам и т.д.) [86].

Если в PDF встроен текст (а не только изображения страниц), PyMuPDF извлекает текст и формирует из него HTML-спанны с координатами; если страницы представлены как картинки (сканы), необходимо предварительно применить технологию оптического распознавания символов (OCR). Для повышения эффективности конвертация может выполняться параллельно на нескольких ядрах или GPU.

При тестировании на больших объемах данных выявлены значительные ошибки выделения необходимой информации, кроме выделения изображений.

Исходя из этого, может быть использована программная библиотека с открытым исходным кодом GROBID (GeneRation Of BIbliographic Data) [87].[1]. У программной библиотеки доступны следующие функциональные возможности: извлечение библиографической информации, полного текста с разбиением на абзацы и предложения, ссылок из статей.

На этом этапе из HTML-структуры извлекаются следующие характеристики публикации: заголовок статьи, список авторов, информация об издании (название журнала/конференции, год, номер, страницы), аннотация, авторские ключевые слова, список литературы. Для этого применяются регулярные выражения и XPath-запросы [88]. Например, название статьи обычно находится либо на первой странице в виде самого крупного текста (что можно определить по размеру шрифта, извлечённому в HTML-спане), либо явно отмечено тегом <h1> при конвертации. Авторы часто следуют сразу за названием – их можно распознать по типичным разделителям (запятые, значок “†” для примечаний) и по тому, что они находятся в области страницы под заголовком. Аффилиации авторов часто помечаются цифрами или символами, совпадающими со сносками у фамилий авторов; их извлечение требует сопоставления этих меток. Аннотация идентифицируется по ключевому слову “Abstract” или соответствующему переводу, за которым идёт один или несколько абзацев текста. Аналогично выделяются ключевые слова авторов – как правило, после аннотации следует пометка “Keywords:” и перечень терминов через запятую. Список литературы чаще всего начинается с заголовка "References" (или «Литература») в конце статьи; всё, что следует за этим заголовком, разбивается на отдельные источники либо по номерам, либо по разделителям (в зависимости от стиля оформления). Из каждого источника извлекаются стандартные поля: авторы, название, издание, год, страницы – для последующего сопоставления при обогащении данных.

---

[1] GROBID – библиотека машинного обучения для извлечения, разбора и реструктуризации исходных документов в структурированные документы с акцентом на технические и научные публикации

Помимо перечисленных характеристик необходимо получить основной текст статьи и встроенные объекты (рисунки, таблицы). При анализе HTML-структуры статьи выявлено, что основной текст часто разделён на колонки и параграфы, а также содержит внутри себя ссылки на рисунки и таблицы. Чтобы корректно сформировать текст по колонкам в один поток, применяется поиск последовательных текстовых элементов, текст считывается сверху вниз: сначала первая колонка, потом вторая, далее страницы объединяются по порядку, формируя цельный контент публикации.

Для проверки работоспособности была взята в качестве примера научная статья «A survey on multimodal large language models» – объемом 20 страниц (1,9 Мб), содержащая все требуемые элементы (3 рисунка, 5 таблиц, химические формулы и физические элементы, 154 источника литературы) [89]. Эта статья послужила тестовым полигоном для отладки программных модулей процесса обработки и насыщения данных. Обработанный по методике JSON-файл данной статьи имеет объём ~ 205 Кб (Приложение Б).

3.2.3 Изображения

Основное внимание в работе уделяется текстовым данным, однако, необходимо рассмотреть и методы работы с изображениями, поскольку при обработке цифровых объектов возникают задачи обработки графических объектов (рисунки, графики, диаграммы), из которых также требуется извлекать информацию. Современные разработки в области компьютерного зрения и машинного обучения предоставляют инструменты для распознавания содержимого изображений.

Среди современных нейросетевых алгоритмов можно выделить те, что используют сверточные нейронные сети. Такие нейронные сети хорошо приспособлены к работе с изображениями и могут выделять (детектировать) на них заданные графические элементы достаточно сложного представления.

Например, семейство моделей **YOLO (You Only Look Once)** – включая версии YOLOv3, YOLO9000 – позволяет детектировать и классифицировать объекты на изображении в реальном времени [90, 91]. Эти модели обучаются

на больших наборах размеченных изображений и способны выявлять сложные визуальные шаблоны. Однако обучение таких нейросетей требует существенных ресурсов: необходимо собрать и разметить сотни, а то и тысячи изображений для каждой категории объектов, после чего обучить модель. Кроме того, для разных типов объектов или условий съёмки могут понадобиться разные модели или дополнительное переобучение.

При разработке методов выделения данных из изображений необходимо обеспечить реализацию нескольких модулей:

1) модуль *фильтрации изображения* – обнаружение на изображении заданных типов графических элементов и получение их количественных характеристик;

2) модуль *фактографического поиска* – ответ на конкретный вопрос по изображению (например, подсчитать количество определённых объектов);

3) модуль *обучения* – дообучение модели на новых образцах;

4) модуль *ускоренной обработки* – использование упрощённых методов для быстрого предварительного анализа изображения.

Существенным ограничением массового использования нейронных сетей для анализа изображений является необходимость обучения под каждый тип объекта, для чего необходимы выборки изображений. Сам процесс обучения нейросетевого алгоритма может потребовать больших временных затрат. При этом заметим, что процесс обучения алгоритмов и настройка может выполняться параллельно.

Выполненные научно-технические исследования позволяют сделать вывод, что разработанные технологии анализа изображений могут быть использованы в рамках программной технической реализации программного обеспечения. Однако необходимо учитывать работы по донастройке обучающих выборок (разметки датасетов) и обслуживания аппаратного комплекса.

3.3 Методы насыщения данных

Насыщение или обогащение данных (*data enrichment*) – процесс извлечения новой информации или знаний из уже существующих наборов данных. В результате насыщения или обогащения данных получается новый, более полный набор данных, который может быть использован для принятия решений [92, 93]. Данный этап важен с точки зрения полноценного анализа и позволяет строить аналитические панели, с помощью которых заинтересованное лицо может получить больше информации об исследуемом объекте.

В рамках работы с научно-технической информацией реализовано несколько направлений обогащения:

- Выделение ключевых словосочетаний из полного текста статьи.
- Распознавание и нормализация физических величин (числовое значение + единица измерения) в тексте.
- Обработка изображений и таблиц научной публикации.
- Обработка аффилиаций авторов: унификация названий стран и организаций, определение координат организаций.
- Определение международных альянсов, к которым относятся страны авторов (по списку стран аффилиаций).

Насыщение данных позволяет расширить возможности анализа: например, ключевые слова текста указывают на основные темы статьи, а определение географических координат организаций даёт возможность географического анализа распределения исследований.

3.3.1 Выделение ключевых слов из полнотекстовых материалов

Для автоматического выделения ключевых слов и фраз из текста публикации автором разработан программный инструмент, в котором пользователь передает на вход параметры и текст для анализа. В качестве параметров задаются: максимальный размер словосочетания (в словах) и требуемое количество ключевых словосочетаний. Программа возвращает список ключевых фраз, упорядоченных по убыванию важности.

В основе алгоритма лежит использование улучшенной версии метода YAKE [66, 80]. Исходный текст разбивается на токены, для каждого токена вычисляется частота его встречаемости. Классический YAKE определяет кандидатов на ключевые фразы и ранжирует их по совокупности статистических критериев (позиция в тексте, частота, распространённость отдельных слов и др.). Предложен следующий алгоритм (Рисунок 3.2), дополненный рядом эвристик, учитывающих особенности научных текстов:

1) Исключение из рассмотрения словосочетаний, которые полностью входят в состав более длинных словосочетаний (например, если выделено «machine learning», то отдельное «learning» исключается из списка).

2) Игнорирование определённых слов или фраз, заданных пользователем или известных как нерелевантные для данной предметной области (например, общего слова «метод»).

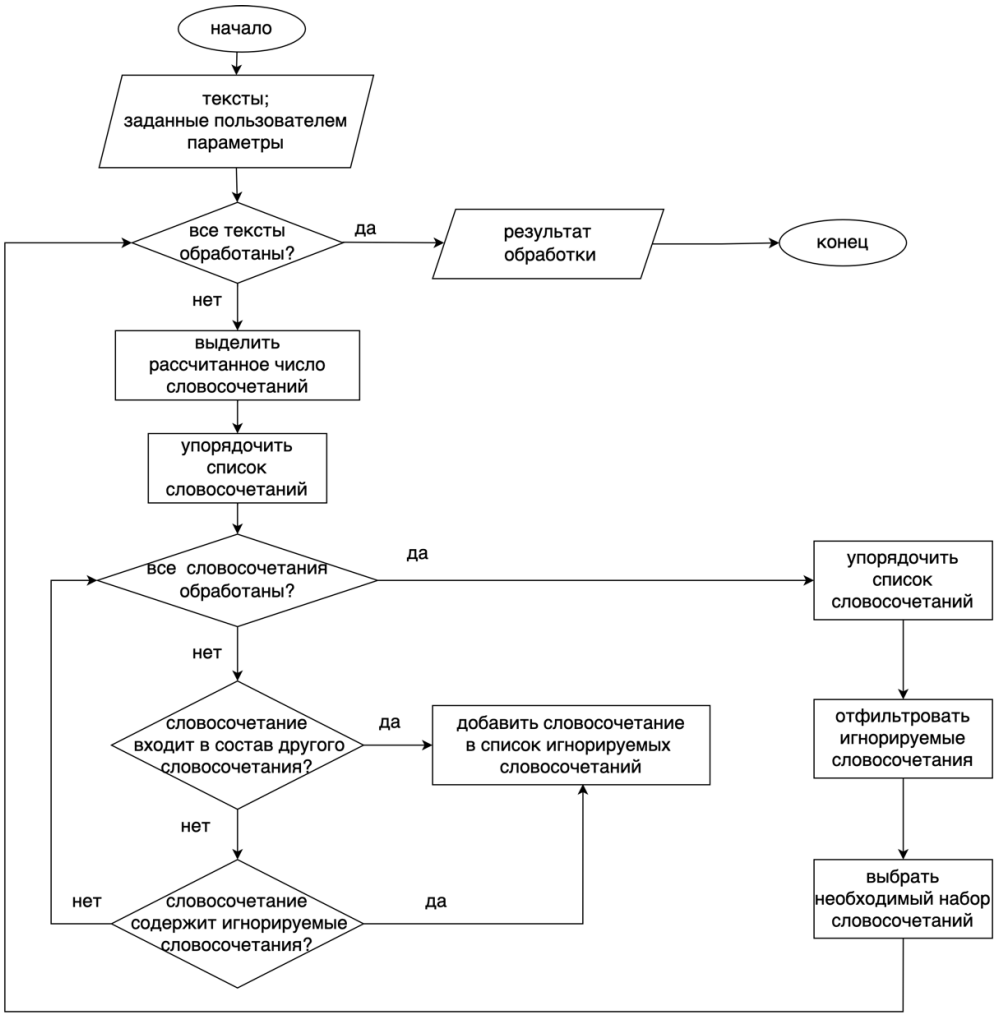


Рисунок 3.2. Блок-схема алгоритма выделения ключевых слов.

В качестве примера рассмотрим статью «A survey on multimodal large language models» [89]. Результат работы программного инструмента с параметрами: максимальный размер словосочетания = 2, требуемое число ключевых фраз = 10 на фрагменте аннотации приведен в таблице (Таблица 3.2). Результат обработки полного текста приведен в Приложении А (в поле keywords внутри full_text структуры JSON).

Таблица 3.2 Результат работы алгоритма по выделению ключевых слов

Данные	**Описание**
Параметры	Максимальный размер ключевого словосочетания = 2 Общее количество ключевых словосочетаний = 4
Текст	INTRODUCTION Recent years have seen remarkable progress in large language models (LLMs) [1 ,2]. By scaling up data size and model size, these LLMs raise extraordinary emergent abilities, typically including instruction following [3], in-context learning (ICL) [4] and chain of thought (CoT) [5]. … Interdisciplinary research: given the strong generalization capabilities and abundant pre-trained knowledge of MLLMs, a promising research direction could be utilizing MLLMs to boost research fields of natural sciences, e.g.", "leveraging MLLMs for analysis of medical images or remote sensing images. To achieve this goal, injecting domain-specific multimodal knowledge into MLLMs might be necessary.
Результат выполнения	["mllms", "models", "images", "training", "works", "instruction tuning", "multimodal instruction", "multimodal data", "caption data", "model performance"]

Из таблицы видно, что алгоритм корректно выделил ключевые термины, характеризующие содержание статьи: «mllms» и «models» отражают основную тематику (мультимодальные большие языковые модели), слова «images» и «caption data» указывают на работу с визуальной информацией, а «training» и «multimodal instruction» подчёркивают важность обучающих стратегий. Дополнительные фразы «works» и «model performance» демонстрируют ориентацию на практические применения и оценку качества моделей, а «instruction tuning» и «multimodal data» раскрывают детали методов тонкой настройки и используемых наборов данных. Результаты тестирования программного инструмента подтверждают корректность работы алгоритма на полном тексте статьи. Применение разработанного программного

инструмента обеспечивает возможность быстрого тематического поиска по коллекции статей [94].

3.3.2 Распознавание и нормализация физических величин в полнотекстовых материалах

Отдельным направлением насыщения данных является автоматическое распознавание физических величин (число + единица измерения) в тексте статьи с последующей унификацией единиц. Программное выделение единиц измерения и величин является важной задачей в научных областях, где много величин и единиц измерения используется в текстовом описании экспериментов и результатов. Особенно, когда необходимо собрать из публикаций количественные результаты экспериментов, параметры материалов и т.п., приведённые в различных системах единиц.

Процесс программного выделения единиц измерения можно представить с помощью следующего алгоритма:

1. Предварительная обработка текста, а именно удаление лишних символов и преобразование текста в стандартный формат, чтобы облегчить дальнейший анализ.

2. Разбиение текста на токены: текст должен быть разбит на отдельные слова и символы, чтобы можно было проанализировать каждый токен отдельно и выделить единицы измерения и величины.

3. Определение контекста: перед выделением единиц измерения и величин необходимо определить контекст, в котором они находятся. Например, если слово «метр» встречается в предложении «длина стержня составляет 5 метров», то очевидно, что «метр» является единицей измерения длины.

4. Выделение единиц измерения: после определения контекста можно выделить единицы измерения. Это может быть сделано на основе списка известных единиц измерения или алгоритмов машинного обучения, которые могут выделять единицы измерения на основе обучающих данных.

5. Выделение величин: после выделения единиц измерения можно выделить числовые значения, которые являются величинами, используя алгоритмы машинного обучения или правила, основанные на контексте.

6. Нормализация единиц измерения: различные единицы измерения могут использоваться для измерения одной и той же величины, например, метры и футы для измерения длины. Поэтому необходимо преобразовать значения величин в стандартные единицы измерения для удобства использования и анализа.

В рамках работы под руководством автора разработан алгоритм и программное средство выделения физических величин [95, 96]. Блок схема алгоритма приведена на рисунке (Рисунок 3.3).

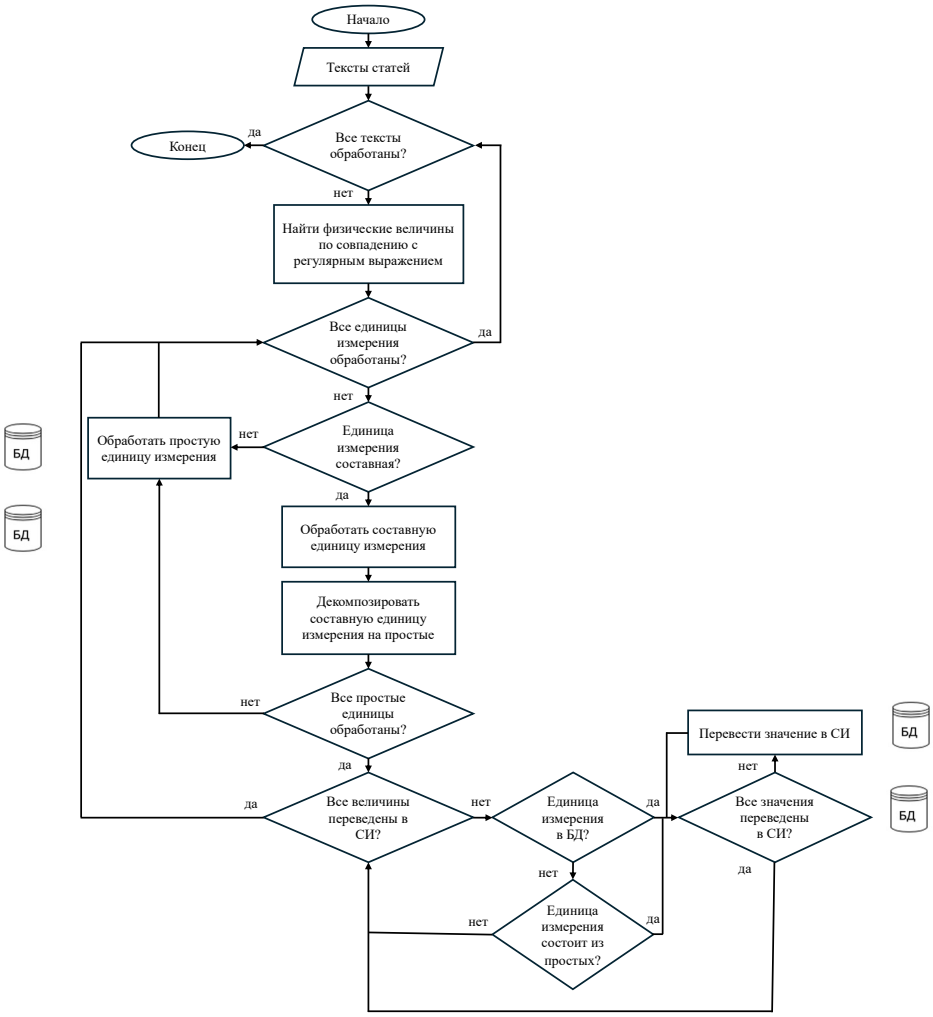


Рисунок 3.3 Блок-схема алгоритма выделения физических величин из текста

Вспомогательную функцию при переводе найденных физических величин в СИ выполняет справочник. Справочником является JSON файл, в

котором представлена информация о 74 объектах: всех приставках (24), базовых (7) и производных именованных (21) единицах СИ, а также добавленных единицах измерения (29). Для приставок записаны их название и числовое значение, тогда как для единиц измерения – название, категория, соотношение с базовыми СИ, возможность иметь степень.

Примеры записей приставок и единиц измерения отображены на рисунках (Рисунок 3.4, Рисунок 3.5).

```
'G':
        {'Name': 'giga',
        'Value': 1000000000
        }

'm':
        {'Name': 'milli',
        'Value': 0.001
        }
```

Рисунок 3.4 Примеры записей приставок в справочнике

```
'H': {
        'Name': 'henry',
        'Quantity': 'inductance',
        'SI': [1,'kg*m^2*s^-2*A^-2'],
        'Powers': True
      }

'J/kg': {
        'Name': 'joule per kilogram',
        'Quantity': 'specific energy',
        'SI': [1,'m^2*s^-2']
      }
```

Рисунок 3.5 Примеры записей единиц измерения в справочнике

В результате работы программного инструмента для каждой исходной физической величины формируется запись, включающая оригинальный символ единицы, исходное значение (или значения, если диапазон), и конвертированное значение с единицей СИ. Фрагмент списка результатов (в формате JSON) представлен в таблице (Таблица 3.3).

Таблица 3.3. Результат работы алгоритма по выделению физических величин

Данные	**Описание**
Текст	Such LMs can provide sufficient tritium breeding ratio and have high thermal conductivity (~101W/mK) and low viscosity (~10-7m2/s) that make them very favorable for heat removal. All LM blanket concepts have, however, feasibility issues associated with magnetohydrodynamic (MHD) interactions between the flowing high electrical conductivity LM (~106S/m) and a strong plasma-confining magnetic field. Only a weak flow is needed for a slow

	(0.1-1mm/s) circulation of the breeder toward the external ancillary system for tritium extraction and LM purification. In this concept, a high-temperature PbLi alloy flows slowly (velocity~10cm/s) in large poloidal rectangular ducts (duct size~20cm) to remove the volumetric heat and produce tritium, while the pressurized He (typically to 8MPa) is used to remove the surface heat flux and to cool the ferritic first wall and other blanket structures to <550°C.

Результат выполнения	[
	{
	"Symbol":"m2/s",
	"Value":[
	"1e-07"
	],
	"SI_converted":"1e-07, m**2/s"
	},
	{
	"Symbol":"W/mK",
	"Value":[
	"101"
	],
	"SI_converted":"101000.0, kg*m**2/(s**3*K)"
	},
	{
	"Symbol":"cm/s",
	"Value":[
	"10"
	],
	"SI_converted":"0.1, m/s"
	},
	{
	"Symbol":"°C",
	"Value":[
	"550"
	],
	"SI_converted":"550.0, °C"
	},
	{
	"Symbol":"S/m",
	"Value":[
	"106"
	],
	"SI_converted":"106.0, A^2*m^-1"
	},
	{
	"Symbol":"cm",
	"Value":[
	"20"
	],
	"SI_converted":"0.2, m"
	},
	{
	"Symbol":"mm/s",
	"Value":[
	"0.1",
	"1.0"
	],
	"SI_converted":"0.0001-0.001, m/s"
	},
	{
	"Symbol":"MPa",
	"Value":[
	"8"
	],
	"SI_converted":"8000000.0, kg/(m*s^2)"
	}
	]

Каждый объект содержит исходное обозначение единицы (Symbol), список значений (Value) – если было диапазон или несколько чисел, приводятся все, – и SI_converted строку: конвертированное числовое значение(я) и единица в СИ.

3.3.3 Обработка изображений и таблиц научной публикации

При работе с научно-технической информацией необходимо совместно с текстовыми данными обрабатывать изображения и таблицы, в которых содержится значимая информация.

Для каждого объекта (изображений или таблиц) происходит поиск его подписи, из которой выделяются основные параметры: тип объекта, номер, подпись. Выделенное изображений или таблица сохраняются в виде файла изображения (в формате jpeg, png и т.д.) или в HTML-формате соответственно. Название сохраняемого файла хранится в структуре JSON с добавлением идентификатора. Структура единичного объекта имеет следующий вид:

1) "name" – название объекта,
2) "number" – порядковый номер,
3) "caption" – подпись,
4) "src" – название файла с данными.

В рамках работы разработан алгоритм и программное средство для автоматизированного извлечения изображений из полнотекстовых публикаций. На рисунках представлено полученное из текста изображение (Рисунок 3.6) и запись данных об изображении (Рисунок 3.7).

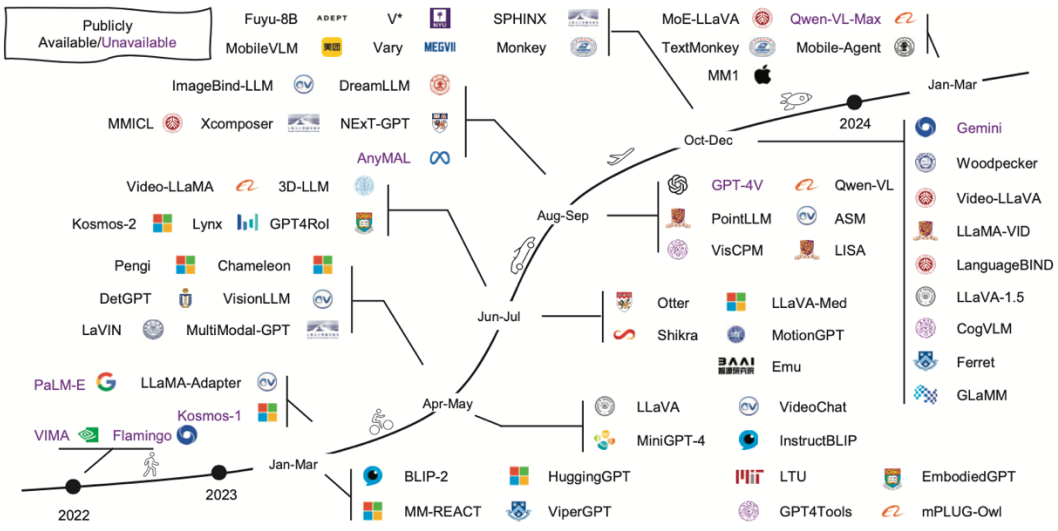


Figure 1. A timeline of representative MLLMs. We are witnessing rapid growth in this field. More works can be found on our released GitHub page, which is updated daily.

Рисунок 3.6 Извлеченное сообщение

```
   {
        "name": "Figure 1",
        "number": "1",
        "caption": "A time line of representative MLLMs. We are witnessing
rapid growth in this field. More works can be found on our released GitHub
page, which is updated daily",
        "src": "85a20e29-6e93-4a9e-a39f-0f47cc382a11/images/Figure 1
t51NVpc3.jpeg"
      },
```


Рисунок 3.7 Извлеченные данные по изображению

В части обработки таблиц имеются следующие особенности: в распознанном тексте необходимо идентифицировать три области (название, заголовки, значения) в отличие от одной (название) для рисунков, графиков, диаграмм и пр.; необходимо интерпретировать графические элементы на изображении (прямые линии), так как их расположение несет информацию о смысловых связях между текстовыми элементами внутри таблицы.

На рисунках (Рисунок 3.8, Рисунок 3.9) представлена извлеченная из текста таблица и пример файла с записью данных.

Table 2. A summary of commonly used open-sourced LLMs. En, Zh, Fr and De stand for English, Chinese, French and German, respectively.

Model	Release date	Pre-train data scale	Parameter size (B)	Language support	Architecture
Flan-T5-XL/XXL [44]	Oct. 2022	–	3/11	En, Fr, De	Encoder decoder
LLaMA [45]	Feb. 2023	1.4T tokens	7/13/33/65	En	Causal decoder
Vicuna [46]	March 2023	1.4T tokens	7/13/33	En	Causal decoder
LLaMA-2 [47]	July 2023	2T tokens	7/13/70	En	Causal decoder
Qwen [48]	Sept. 2023	3T tokens	1.8/7/14/72	En, Zh	Causal decoder
LLaMA-3 [49]	April 2024	15T tokens	8/70/405	En, Fr, De, etc.	Causal decoder

Рисунок 3.8 Исходная таблица


```
{
  "name": "Table 2",
  "number": "2",
  "caption": "A summary of commonly used open-sourced LLMs.En, Zh, Fr and De
stand for English, Chinese, French and German, respectively.",
  "src": "85a20e29-6e93-4a9e-a39f-0f47cc382a11/tables/Table 2 ezpkwxIC.html"
}
<table border="1">
   <tr>
      <td></td>
      <td>Release</td>
      <td>Pre-train</td>
      <td>Parameter size</td>
      <td>Language</td>
      <td></td>
   </tr>
```

```
    <tr>
       <td>Model</td>
       <td>date</td>
       <td>data scale</td>
       <td>(B)</td>
       <td>support</td>
       <td>Architecture</td>
    </tr>
    <tr>
       <td>Flan-T5-XL/XXL [44 ]</td>
       <td>Oct. 2022</td>
       <td>-</td>
       <td>3/11</td>
       <td>En, Fr, De</td>
       <td>Encoder decoder</td>
    </tr>
    <tr>
       <td>LLaMA [45 ]</td>
       <td>Feb. 2023</td>
       <td>1.4T tokens</td>
       <td>7/13/33/65</td>
       <td>En</td>
       <td>Causal decoder</td>
    </tr>
    <tr>
       <td>Vicuna [46 ]</td>
       <td>March 2023</td>
       <td>1.4T tokens</td>
       <td>7/13/33</td>
       <td>En</td>
       <td>Causal decoder</td>
    </tr>
    <tr>
       <td>LLaMA-2 [47 ]</td>
       <td>July 2023</td>
       <td>2T tokens</td>
       <td>7/13/70</td>
       <td>En</td>
       <td>Causal decoder</td>
    </tr>
    <tr>
       <td>Qwen [48 ]</td>
       <td>Sept. 2023</td>
       <td>3T tokens</td>
       <td>1.8/7/14/72</td>
       <td>En, Zh</td>
       <td>Causal decoder</td>
    </tr>
    <tr>
       <td>LLaMA-3 [49 ]</td>
       <td>April 2024</td>
       <td>15T tokens</td>
       <td>8/70/405</td>
       <td>En, Fr, De, etc.</td>
       <td>Causal decoder</td>
    </tr>
</table>
```

Рисунок 3.9 Файл с данными таблицы

Результат работы разработанной методики по выделению изображений и таблиц для научной публикации «A surveyon multimodal large language

models» [89] представлен в Приложение Б внутри поля «images» и поля «tables».

3.3.4 Обработка аффилиаций авторов: унификация названий стран и организаций, определение координат организаций

Одной из важных задач обработки данных является их унификация (приведение данных к единому виду/формату). Полученные в результате унификации данные позволяют проводить более точный анализ, что является критическим для процесса анализа информации. В рамках работы разработан подход унификации аффилиаций организаций для получения их официальных названий и географических координат (широта и долгота).

Добавление географических координат расширяет возможности анализа: построение карт в различных масштабах, оценка экономики регионов и т.д. При этом официальные названия организаций позволяют более точно определять геолокацию по сравнению с аффилиациями.

Необходимость определения официальных названий связана с тем, что не существует жесткого стандарта написания аффилиаций, поэтому один и тот же объект (организация) может быть записан по-разному (например Объединенный институт ядерных исследований в реферативной базе данных Scopus записан 105 разными способами, в том числе – «JINR, Dubna, Moscow reg., 141980, 6 Joliot-Curie, Russian Federation» и «Joint Institute for Nuclear Research, 141980 Dubna, Moscow Region, Russian Federation» определяют одну и ту же организацию,). Это приводит к ошибочным расчетам статистических значений встречаемости цифрового объекта.

Для представления алгоритма работы рассмотрим решаемые задачи. Пусть аффилиация представлена одним из следующих видов:

- <Организация>, <Страна>;
- <Организация>, <Город>, <Страна>;
- <Подразделение>, <Организация>, <Город>, <Страна>;
- <Подразделение>, <Организация>, <Аббревиатура>, <Город>, <Страна>;

- <Организация>, <Город/Индекс>, <Субъект>, <Страна>;
- <Подразделение>, <Организация>, <Адрес>, <Почтовый индекс>, <Город>, <Страна>;
- и т.д.

Разработан алгоритм определения страны по аффилиации (Рисунок 3.10).

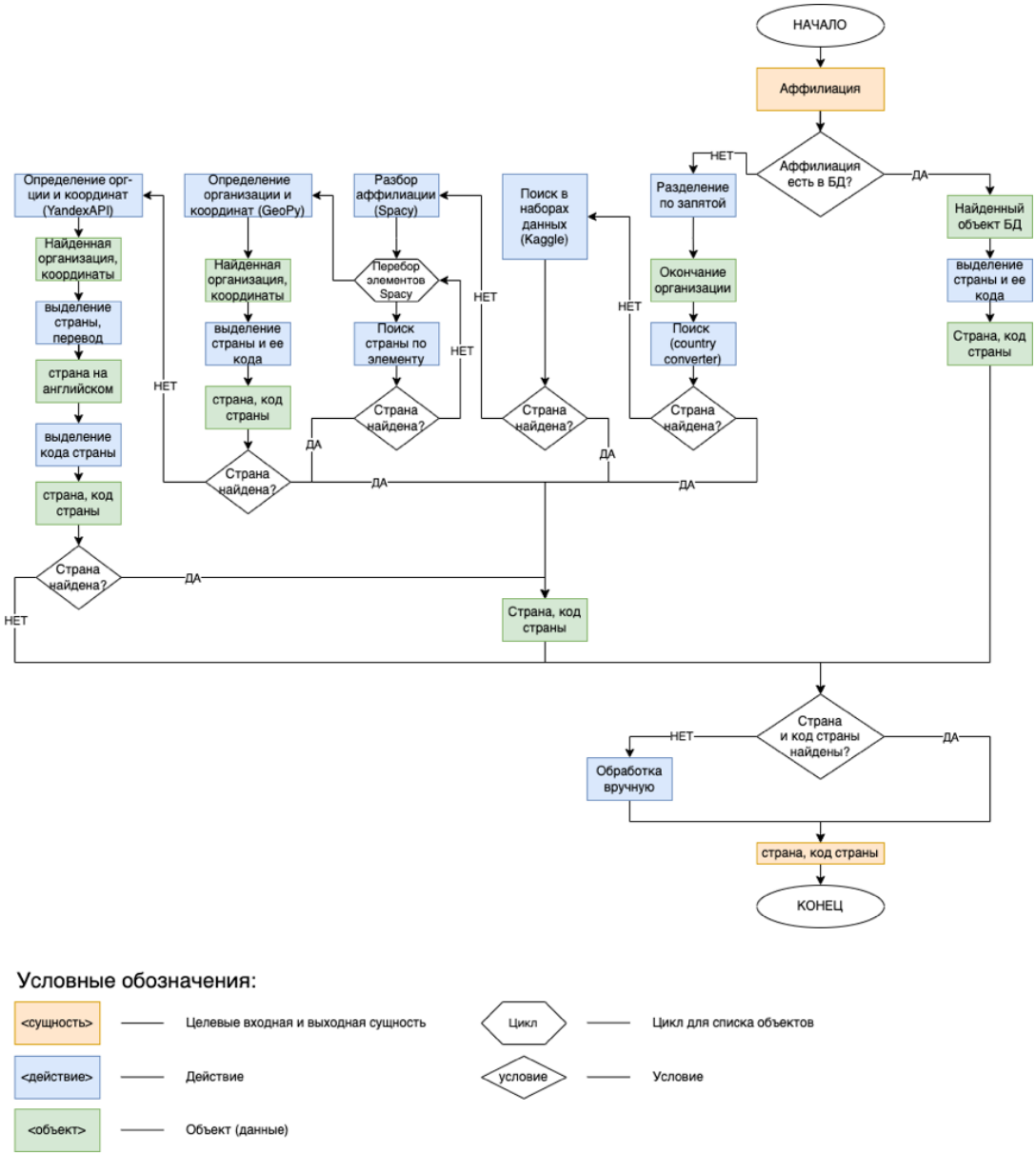


Рисунок 3.10 Блок схема алгоритма определения страны по аффилиации

Сначала проверяется наличие аффилиации в локальной базе (сопоставление полным строкам). Если находится, извлекаются сохранённые страна и код страны. Если нет, пытается выделить страну по последней части строки (после последней запятой) – часто там указывается страна или её

сокращение. Если и это не даёт результата, выполняется поиск подстроки аффилиации в заранее загруженных данных (например, список университетов мира); либо, как альтернативный путь – разбирается аффилиация с помощью NLP (spaCy) на отдельные сущности ORG и GPE. Затем проверяется каждая найденная сущность: если она соответствует стране или содержит известное название страны, выбирается она.

Если указанные способы не находят страну, запускается запрос к внешним геокодерам – **GeoPy** (с сервисом Nominatim OpenStreetMap) [97] и **Yandex Geocoder API** [98]. Эти сервисы по строке адреса возвращают предполагаемые координаты и объект, к которому они относятся (обычно город или страна). Если страна определена (тем или иным методом), берётся её двухбуквенный код (ISO 3166-1 alpha-2) – это необходимо для стандартизации. Также, если определена организация (например, из разбора spaCy ORG), запускается процедура поиска её координат: берётся найденная организация, выполняется запрос к геокодеру, результатом становятся широта и долгота организации (если найдена точная локация). Выявленная пара «официальное название организации + координаты + код страны» сохраняется в базу данных, чтобы при следующем появлении этой же аффилиации или организации сразу взять готовый ответ. Если ни один способ автоматически не дал результата, аффилиация помечается как требующая ручной обработки.

Аналогично рассмотрим процесс определения официальных названий организаций и их координат (Рисунок 3.11). Поскольку целевым объектом является организация, стоящая в шаблонах на первом или втором месте, ее можно выделить при помощи разделителя (запятая). Однако такой способ работает некорректно, если в подразделении стоят дополнительные запятые или входящая строка имеет другой шаблон с другим числом разделителей.

В связи с этим выделение организации из аффилиации осуществляется при помощи обработки естественного языка, а именно выделения именованных сущностей при помощи spaCy. В процессе выделения именованных сущностей выделяются сущности типа «организации» (ORG,

organisation) и «географические объекты» (GPE, geopolitical entity); последние служат для более точного определения организаций.

После выделения именованных сущностей с помощью модуля spaCy имеется список организаций, причем более чем в 90% случаев сначала идут подразделения, после чего идет организация. Поэтому дальнейший поиск географических координат проводится для выделенного списка организаций, записанного в обратном порядке.

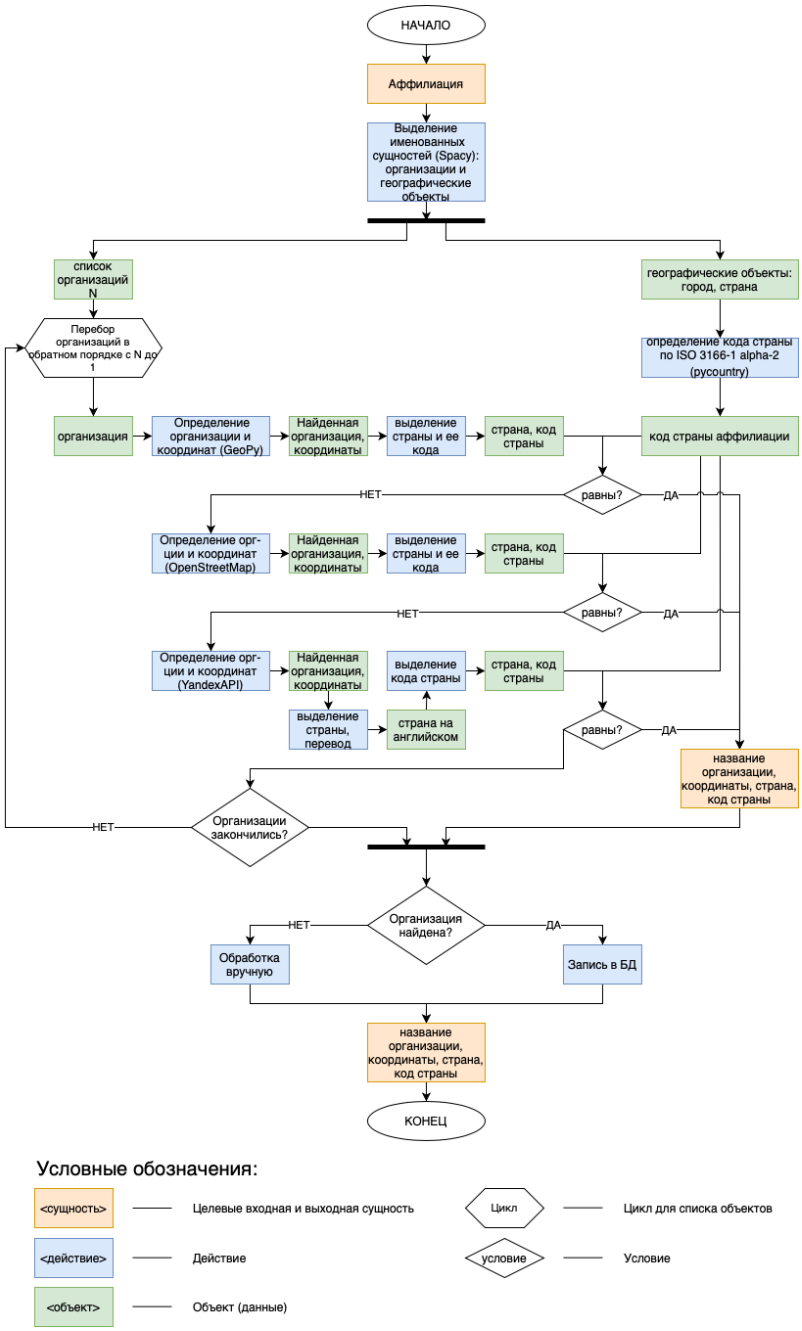


Рисунок 3.11 Блок-схема алгоритма определения официальных названий организаций и их координат

Проведена апробация данного сервиса на ~26 000 аффилиациях из базы Scopus. По результатам: около 4 000 организаций определены корректно через GeoPy, ~3 000 – через YandexAPI, ~6 500 – через алгоритмическое объединение результатов GeoPy/Яндекс (устранение ошибок), оставшиеся ~12 500 потребовали ручной проверки (либо не определились автоматически). Причины сложностей – опечатки, разные языки, неполнота данных OSM для некоторых цифровых объектов.

В итоге, для каждой аффилиации, если она разобрана успешно, в структуре данных формируются: официальное название организации, страна, код страны, широта и долгота. Эти данные сохраняются в поле affiliations. Кроме того, список стран (уже определённых кодов) сохраняется в поле countries, а их двухбуквенные коды – в countries_code. Например, organizations – Tel Aviv University (Israel, координаты…), ETH Zurich (Switzerland, координаты…), и т.д., countries – Israel, Switzerland, USA; countries_code – IL, CH, US.

3.3.5 Выделение международных объединений по аффилиации авторов статьи

Одно из направлений насыщения данных связано с геополитическим анализом: по списку стран, к которым относятся организации (аффилиации) авторов статьи, определяется, входят ли эти страны в определённые международные объединения (альянсы). Эта информация полезна для анализа международного сотрудничества – какие международные блоки (ЕС, СНГ, НАТО, G7, БРИКС и т.д.) представлены в авторах публикации.

Для решения задачи разработана база данных, включающая основные международные альянсы: ООН, СНГ, БРИКС, ШОС, Лига арабских государств, МЕРКОСУР, НАТО, G20, G7, AUKUS. В базе данных для каждого альянса хранится список стран-участниц (обычно по официальному списку на текущий момент). На рисунке (Рисунок 3.12) представлена схема базы данных международных альянсов.

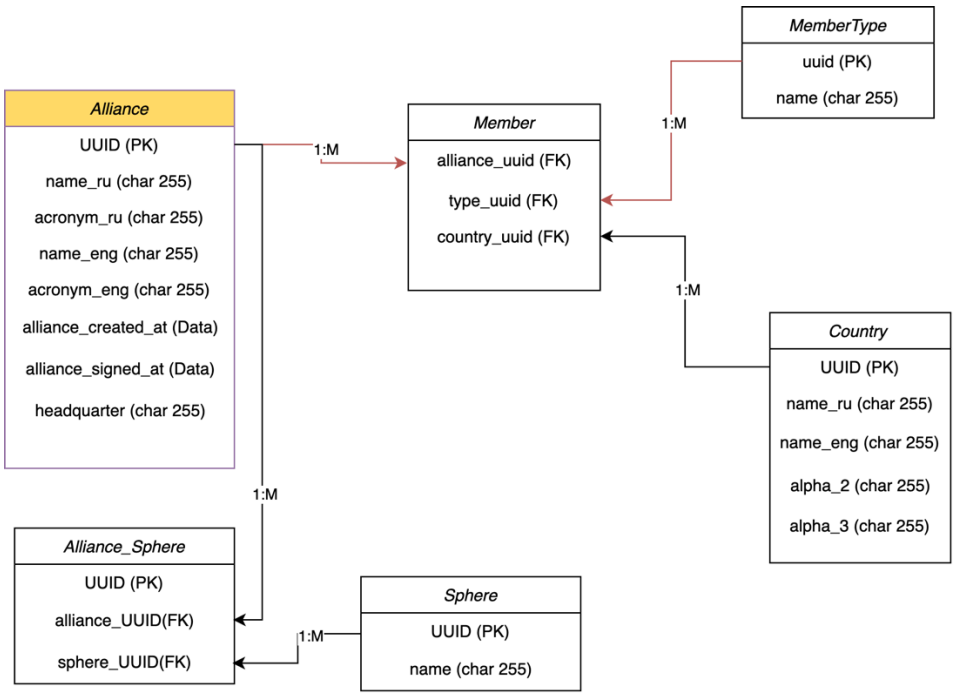


Рисунок 3.12 Схема базы данных международных альянсов

Разработанный программный инструмент принимает на вход список стран, каждое название страны приводится к стандартному виду – используется библиотека pycountry для получения стандартного наименования и кода страны [99]. Если какая-то страна отсутствует в базе, она добавляется (что позволяет базе расширяться по мере обработки новых данных). По каждой стране проходит проверка по записям альянсов и проверяет наличие кода страны в списке участников. Результатом работы является список названий альянсов, представленных среди стран авторов данной статьи.

Например, если у статьи авторы из России, Китая и ЮАР, то страны: Russia, China, South Africa. Россия и Китай – входят в ШОС, БРИКС, ООН, также Россия – в СНГ, Китай и ЮАР – в G20 и ООН, все три – в ООН, и т.д. Итоговый набор альянсов будет: ООН, БРИКС, ШОС, G20, СНГ. Эти данные сохраняются в поле full_text.alliances структуры данных.

В случае статьи-примера (A survey on multimodal large language models) авторы принадлежат организации из Китайской Народной Республики (China). На основе выделенной и унифицированной информации о стране, программный инструмент извлекает список соответствующих международных объединений. Таким образом, alliances для текущего примера

включает: ООН, БРИКС, G20, ШОС. Эти данные включены в структуру (Приложение Б, поле alliances).

3.4 Хранение данных о цифровом объекте

После выполнения этапов извлечения и насыщения данных необходимо обеспечить хранение и доступность полученной информации для анализа и визуализации. Рассмотрим используемые технологии баз данных и общее устройство хранилища, в котором аккумулируются сведения о цифровых объектах.

В современных информационных системах применяются две основные парадигмы управления данными: реляционные (SQL) базы данных и нереляционные (NoSQL) хранилища [100]. Реляционные СУБД оперируют таблицами фиксированной структуры (с колонками определённого типа) и отношениями между таблицами. Данные извлекаются с помощью языка SQL, а целостность поддерживается ограничениями и транзакциями. Нереляционные базы, напротив, допускают гибкие динамические схемы или вовсе их отсутствие, храня данные в виде документов, графов, пар ключ-значение или других моделей. Ниже приведено сравнение этих подходов (Таблица 3.4).

Таблица 3.4 Описание отличий SQL и NoSQL баз данных

Критерий	**SQL БД**	**NoSQL БД**
Представление хранения данных	Данные хранятся в таблицах, структура строк фиксирована и строго типизирована.	Данные могут храниться в виде документов (JSON, XML), графов, пар «ключ–значение» или колонок. Структура записи может быть разной для разных элементов.
Изменение структуры и типов данных	Требует изменения схемы БД (что может затронуть все существующие данные.	Структура записи может меняться «на лету»: новые поля могут добавляться без затрагивания других записей, отсутствующие поля просто не хранятся.

Доступ к данным	Используется язык SQL: данные создаются, обновляются, удаляются и извлекаются с помощью SQL-запросов.	Данные манипулируются через API конкретной БД; некоторые NoSQL поддерживают SQL-подобные запросы, но не стандартный SQL.
Целостность данных	Схема и ограничения (типы, NULL/NOT NULL, FOREIGN KEY и пр.) гарантируют целостность: некорректные данные не попадут в БД.	Строгих ограничений на структуру нет, целостность данных контролируется на уровне приложения или вообще не обеспечивается (в традиционном смысле).
Масштабируемость	Вертикальная: увеличение ресурсов сервера (CPU, RAM) для повышения производительности на одном узле.	Горизонтальная: масштабирование за счёт добавления узлов, распределение данных и нагрузки между несколькими серверами.

Для хранения данных и построения визуализаций автором используется набор программных инструментов ELK (Elasticsearch, Logstash, Kibana).

В настоящее время Elasticsearch является одной из наиболее используемой платформой для хранения текстовых данных [91]. Кроме того, Elasticsearch является распределенной программной поисковой системой с открытым исходным кодом, которая имеет JSON REST API для взаимодействия с загруженными данными. Данные в Elasticsearch хранятся в Apache Lucene как инвертированный индекс. Инвертированный индекс представляет собой структуру данных, в котором каждому слову соответствует список из названий документов и позиции, где это слово встречается. Единичную структуру записи можно представить в виде «слово документ: позиция». С помощью данного представления и собранной информации можно найти все документы, в которых встречалось отдельное слово. С добавлением новых документов индекс каждого слова при необходимости дополняется.

Еще одной особенностью является то, что при загрузке данных используется морфологический метод, который заключается в сокращения слова до его корня, то есть в ходе индексирования документа сохраняется

корневое слово вместо фактического. Данный метод способствует нахождению всех слов в документах в независимости от их падежей. Поисковая система использует словарь синонимов для поиска слов с одинаковым смыслом, при этом уменьшается размер индекса, и устраняются различия между похожими словами. За счет сохранения позиций слова в документе появляется возможность поиска не только отдельных слов, но и фраз в документах. Это достигается путем сравнения позиций в поисковой строке и загруженных документах: если в документе есть слова в том же порядке, что и в запросе, то фраза является подходящей.

При добавлении документа JSON в Elasticsearch он сохраняется как исходный JSON (source) и параллельно индексируется. Впоследствии, при поиске, Lucene позволяет находить документы, содержащие определённые термины, а за счёт хранения позиций – поддерживается и поиск фраз. Например, если искать «металлических биоматериалов», система найдёт документ, где эти слова стоят рядом (как фраза). Благодаря индексации данные извлекаются очень быстро даже на коллекциях сотен тысяч документов.

Визуализация данных может быть организована,через плагин Kibana (инструмент, входящий в ElasticStack) либо через специально написанное веб-приложение. Kibana позволяет строить аналитические панели, графики, тематические облака ключевых слов, распределения по годам на основе данных из Elasticsearch. Также плагин Kibana предоставляет интерфейс для полнотекстового поиска по индексу.

Нереляционная база данных обеспечивает оперативный поиск и удобную визуализацию обработанных данных научных публикаций. Она масштабируема и может быть развёрнута на нескольких узлах, что позволит системе работать эффективно даже при существенно возросшем количестве данных и пользователей.

ВЫВОДЫ ПО ГЛАВЕ 3

1. Разработаны алгоритмы и программные средства автоматизированного извлечения и насыщения, структурирования и хранения данных для наполнения моделей цифрового объекта, при которых разнородные входные данные (тексты документов, веб-страницы, изображения, таблицы и др.) конвертируются в унифицированный формат.

2. Разработаны методы и программные средства интеллектуального анализа данных, в том числе выделения ключевых словосочетаний из полного текста статьи, физических величин и их нормализации; унификации названий стран и организаций; определения координат организаций; определения международных альянсов, к которым относятся страны авторов.

3. Разработан программный инструмент для автоматического выделения иллюстраций (графиков, схем) и таблиц из разнородных документов. Предложенный алгоритм корректно определяет границы таблиц и выделяет структуры строк/столбцов, позволяет извлекать табличные данные для последующего насыщения базовой модели цифрового объекта не только текстовыми свойствами, но и графическими и числовыми данными, что повышает информативность дальнейшего анализа.

4. Разработана стандартизованная структура данных для представления интегральной информации о цифровом объекте на примере научной публикации, поля которой соответствуют статическим, динамическим характеристикам модели. Такая структура гибко поддерживает вложенность (например, список авторов с их аффилиациями и странами, список цитирований, список разделов текста и т.п.) и позволяет хранить цифровой объект в нереляционной документно-ориентированной базе данных (NoSQL).

ГЛАВА 4 АНАЛИТИЧЕСКИЕ ИССЛЕДОВАНИЯ ЦИФРОВЫХ ОБЪЕКТОВ В СОЦИАЛЬНОЙ СРЕДЕ

4.1 Аналитическая модель цифрового объекта в социальной сфере

Проведение исследований в социальном сегменте неразрывно связано с исследованием социальных сетей различной направленности. В социальных сетях можно выделить два крупных информационных объекта – это профиль персоны и группы/сообщества. В общем виде группа – набор персон, привязанных к некоторой странице социальной сети. Каждый объект в социальной сети является многомерным объектом. На практике он описывается двумя - тремя десятками характеристик [101, 102].

В работе с социальной информацией исследователи сталкиваются с несколькими крупными проблемами:

- Сведения о пользователе в большинстве случаев неполны (ряд полей заполняется по выбору, часть данных скрывается настройками приватности).
- Высокий информационный шум за счёт рекламных, анонимных или роботизированных аккаунтов.
- Динамические характеристики изменяются с неодинаковой частотой: текстовый статус может обновляться ежедневно, тогда как список групп существенно сдвигается лишь при смене интересов.

При рассмотрении цифрового профиля персоны в социальной сети можно определить как статические, так и динамические характеристики. Проведенные автором исследования ведущих социальных сетей выявили, что цифровой профиль (аккаунт) устойчиво описывается 41 статическими характеристиками.

Статические характеристики для удобства восприятия можно разделить на 5 категорий: основное, контакты, деятельность, интересы, жизненная позиция (Таблица 3.1). Выделенные критерии можно отнести к категории статичных характеристик (S), так как они заполняются пользователем при первичной регистрации и не обновляются регулярно.

Таблица 3.1 Категоризация статических характеристик

№	Категории	Характеристики
1	Основное	ссылка на страницу, имя, фамилия, пол, семейное положение, партнёры, дата рождения, родной город, языки, дедушки, бабушки, родители, братья/сёстры, дети, внуки
2	Контакты	город, дом, мобильный телефон, дополнительный телефон, skype, личный сайт
3	Деятельность	образование, место работы, номер школы, военная служба
4	Интересы	деятельность, интересы, любимая музыка, любимые фильмы, любимые телешоу, любимые книги, любимые игры, любимые цитаты, о себе
5	Жизненная позиция	политические предпочтения, мировоззрение, главное в жизни, главное в людях, отношение к курению, отношению к алкоголю, вдохновляют

К характеристикам, содержащим основную информацию о профиле персоны, может быть добавлена такая информация как: аудио и видео записи, фотография профиля (аватар) и текстовый статус. Все эти данные могут изменяться с течением времени существования цифрового профиля, поэтому выделим их в качестве динамических характеристик (D).

Существенное значение на анализ цифрового профиля оказывают связи с другими объектами (Rel) – связь с такими объектами как друзья, подписчики, вхождение в группы (r =«Друг», «Подписчик», «Состоит в группе»).

Аналитическое описание цифрового профиля персоны в социальной среде, согласно введенной формуле (2.4.) обладает следующей размерностью:

$$|S| = 41, |D| = 4, |Rel| = 3.$$

В решении некоторых аналитических задач анализа социальных объектов используются не все характеристики цифрового объекта. Например, при решении аналитической задачи отнесения цифрового профиля персоны к

школьнику достаточно вычисляемой характеристики возраст (f_{age}) на основе данных статической характеристики – даты рождения, значение которой должно попадать в диапазон от 6 до 18 лет.

Значения перечисленных характеристик могут относится к различным типам данным – текстовые, числовые, временные и т.д., что *обуславливает необходимость учета типа данных при разработке методов обработки*.

4.2 Методика построения аналитической модели цифрового объекта

Для решения аналитических задач необходимо разработать метод перехода от разнотиповых значений характеристик к единой системе оценки. Список фактических характеристик в социальной сети ранжируется по степени важности (в зависимости от решаемой задачи). Отранжированный ряд с помощью формальной процедуры преобразуется в ряд весовых коэффициентов, нормированных от 0 до 1. Имея таблицу взвешенных характеристик, можно минимизировать количество характеристик, пренебрегая наименее значимыми. По каждой характеристике аналитической цифровой модели необходимо построить собственную методику и алгоритм измерений сопоставления разнотиповых значений.

Каждой характеристике присваивается детерминированная оценка: численная, балльная или по тезаурусу с указанием диапазона возможных значений. Полученные детерминированные значения оценки имеют разную размерность, то есть являются неаддитивными величинами, поэтому их нельзя складывать в интегральном критерии [103].

Вследствие чего по каждому критерию вводятся относительные значения характеристики, которые преобразуются в безразмерные. В качестве интегральной оценки объекта используется сумма произведений относительной значимости критерия на относительное значение его характеристики.

Устанавливая некоторое значение критерия как нормативное, мы приходим к простому правилу маркировки объектов. Все объекты со

значением критерия больше нормативного, маркируются, как целевые объекты.

Аналитическая модель цифрового объекта, предложенная выше, предоставляет формальную основу, однако для практического применения её необходимо «привязать» к конкретным задачам идентификации. Решение аналитических задач базируется на синтезе методов обработки исходных данных и разработке методов расчета вычисляемых характеристик (F) для каждого цифрового информационного объекта. Рассмотрим некоторые примеры расчета вычисляемых характеристик в реальных аналитических задачах.

Одной из базовых задач, решаемых в социальной среде, является соотнесение цифрового профиля с различными социальными группами, например, школьник, студент и др.

Идентификация школьника по данным в социальной среде возможна по значениям двух характеристик:

- наличие школы (φ_{boolen});
- целевой возраст (φ_{in_range}).

В данном случае примем, что перечисленные характеристики являются равнозначными, так как значение возраста информационного цифрового профиля должно попасть в заданный диапазон (от 6 до 18 лет) и в профиле должна быть указана школа; тогда их вес в методике соотнесения цифрового профиля к социальной группе школьника одинаков $v_{birth} = v_{nsch} = 0{,}5$.

$$f_{sch} = \varphi_{sch}(s_{birth}, s_{nsch}) = \tag{4.1}$$

$$v_{birth} \cdot \varphi_{in\_range}(s_{birth}) + v_{nsch} \cdot \varphi_{boolen}(s_{nsch}), f_{sch} \in [0,1],$$

где s_{birth} – дата рождения (временное),

s_{nsch} – поле номера школы (текстовое),

v_{birth}, v_{nsch} – весовые значения вычисляемых характеристик.

На рисунке (Рисунок 4.1) приведен фрагмент файла в формате JSON, в котором указано каким методом обрабатывать то или иное поле профиля. В дальнейшем будем называть такие файлы «Диагностической картой».

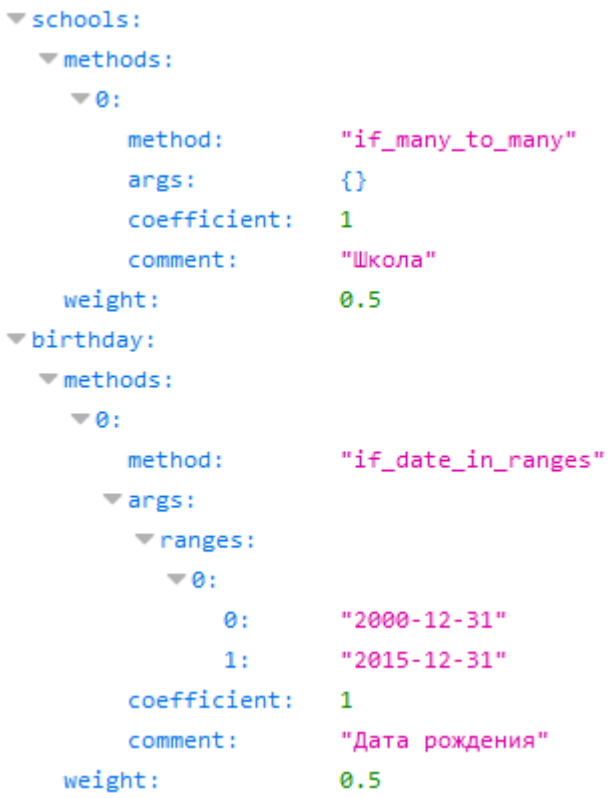


Рисунок 4.1 Диагностическая карта маркирования школьника

Для анализа возраста указывается диапазон дат, которому должно соответствовать значение возраста у исследуемого объекта. Данный диапазон дат динамически изменяется ежегодно. При решении практических аналитических задач $\varphi_{boolen}(s_{nsch})$ может учитывать заведомо неверные значения, например, номер школы свыше 3000, номер 1, 420 и др.

Работа с цифровыми профилями в социальной среде предполагает решение задачи определения заведомо ложных цифровых профилей (роботов), занимающихся автоматизированным сбором информации либо технических/ коммерческих аккаунтов, создаваемых для искусственного наращивания аудитории. Такие профили характеризуется большим количеством связей с группами r =«Состоит в группе»> 300) и количеством друзей или подписчиков таких профилей (r =«Друг», «Подписчик»> 300).

Так же в данном случае используется метод «от обратного», то есть если аккаунт был подтвержден в социальной среде, то данный цифровой профиль не может являться роботом [104].

Отдельно учитывается нижний порог: при $r =$«Друг»< 120 и $r =$«Подписчик»< 50 профиль получает дополнительный штрафной коэффициент, так как низкое социальное окружение коррелирует с девиантным поведением.

Интегральная формула расчета вычисляемой характеристики:

$f_{bot} = \varphi_{boolen}(d_{friends}) + \varphi_{boolen}(d_{subs}) + \varphi_{boolen}(d_{groups}),$ $f_{bot} \in [0,1].$	(4.2)

При $f_{bot} \geq 0{,}5$ цифровой профиль помечается как робот. Пороговое значение подбиралось эмпирически: меньше 0,5 порождает слишком много ложных срабатываний, выше — упускает «умных» ботов.

4.3 Решение сложных идентификационных задач в социальной среде

4.3.1 Формирование обучающей выборки целевых объектов

Рассмотрим более сложные примеры проведения аналитических исследований в цифровой социальной среде. Одной из существенных проблем является девиантное поведение различных социальных групп.

Подростковый возраст характеризуется повышенным риском девиантного поведения и суицидальных проявлений. В последние годы наблюдается тревожный рост числа попыток суицида среди подростков: согласно анализу, «суицидальные попытки подростков находятся на подъеме, представляя существенную проблему общественного здравоохранения». Оценочно примерно 17% подростков в развивающихся странах хотя бы раз предпринимали попытку суицида [105, 106, 107]. В таких условиях актуальной является задача ранней идентификации групп риска. Одним из перспективных подходов к идентификации групп риска является психографический анализ — выявление индивидуальных психологических характеристик (например, черт личности, ценностных ориентаций, мотивов).

Важной частью психографического анализа являются стандартные психологические методики, измеряющие черты личности, ценности, мотивацию и стратегии преодоления. Классическим инструментом являются опросники по модели «Большой пятерки» (Big Five): NEO-PI, BFI и др [108, 109]. По результатам мета-анализа, проведённого на подростках, установлено, что повышенный нейротизм ассоциируется с суицидальным поведением и является его фактором риска [110]. С развитием машинного обучения и цифровых технологий возрастает интерес к разработке моделей, которые на основе опросных данных и цифровых «индикаторов» могут выявлять учащихся с высоким суицидальным риском и осуществлять раннее вмешательство [111, 112, 107, 113].

С появлением цифровых технологий резко расширился набор индикаторов для психографического анализа. Например, концепция «цифрового фенотипирования» (digital phenotyping) предполагает непрерывный сбор данных о поведении человека через смартфоны и носимые устройства [114, 115]. Активные данные (опросы, self-report) дополняются пассивными: GPS-трекеры, данные акселерометра, журналы вызовов и текстовых сообщений, активность в приложениях, публикации в социальных сетях и т. д. [114, 112, 116]. Это позволяет создавать «цифровой портрет» личности и поведения подростка.

Социальные сети и мессенджеры подростков содержат важные сигналы суицидального риска. Так, в исследовании анализа письменных описаний саморазрушительных переживаний подростков были выделены три основные группы слов: связанные с «болью или отчаянием», «отношениями/связью» и «способностью совершить попытку» [117]. Из лексических кластеров выявлялись подтемы: психологическая боль, безнадежность, поиск помощи, описания методов попыток и т.п. Кроме того, были выделены специфические сленговые термины [117, 118].

Помимо контента сообщений, анализируются взаимосвязи между субъектами и феноменами. Например, социально-сетевой анализ изучает

структуру дружеских групп подростков: наличие в окружении склонных к риску сверстников повышает вероятность подражания. Методы графовых нейросетей позволяют учитывать информацию о социальных связях наряду с индивидуальными признаками (например, если у цифрового профиля низкий статус в сети друзей и его посты содержат депрессивные выражения, что комплексно повышает риск).

Таким образом, цифровые индикаторы включают широкий спектр данных: от текстовых сообщений и записей речи до изображений лица и социальных связей. Обобщение таких данных многомерными методами позволяет моделям точнее определять «цифровой фенотип» каждого подростка и предсказывать риск на основе неожиданных закономерностей.

Под руководством автора в 2018 году по поручению Межведомственной рабочей группы по информационной безопасности Совета Безопасности Российской Федерации и в рамках выполнения государственного заданий Министерства образования и науки был проведен комплекс исследований и разработана платформа комплексной антисуицидальной интернет профилактики среди подростков. В рамках выполнения работ была проведена апробация предложенных автором методов.

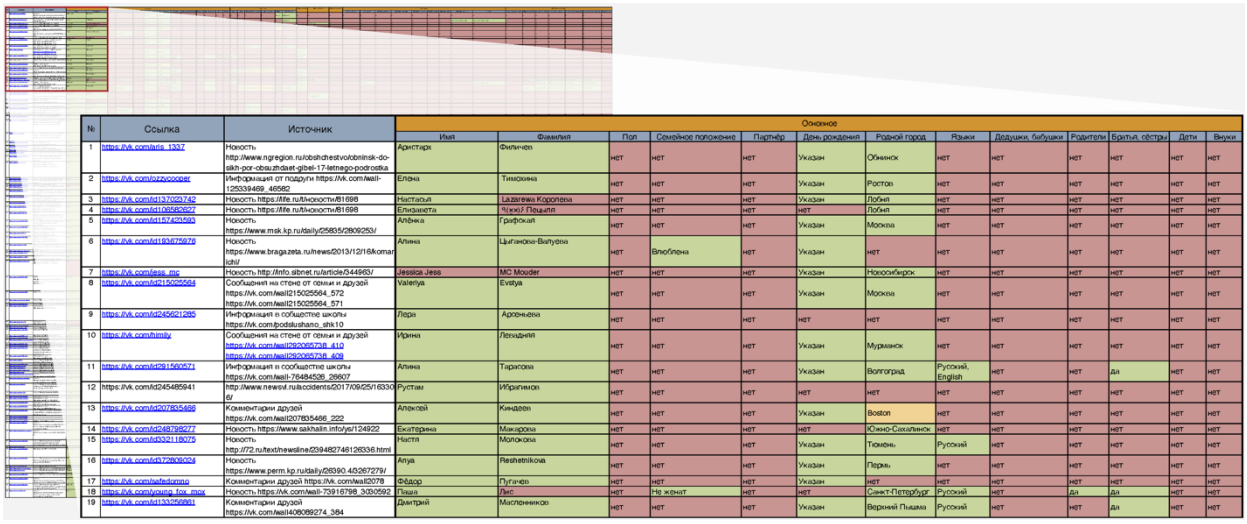
В исследовании решалась задача поиска индивидуумов с целью последующего отнесения их к «группе риска», проведен комплекс работ по разработке аналитических моделей цифровых объектов «персона» и «группа» в социальной сети и методов обработки отдельных характеристик (здесь и далее «группа риска» – представители возрастной группы от 10 до 22 лет, наиболее уязвимые к совершению девиантных действий, проявлению агрессии по отношению к себе и окружающим).

Первым этапом стало формирование эмпирической базы, на которой строится последующий расчёт весов и пороговых значений. В качестве исходного корпуса были выявлены 100 цифровых профилей объектов, характерных для девиантного поведения. Для достижения высокого уровня адекватности и актуальности информации верификация профилей

проводилась на основе таких открытых источников информации, как новостные ресурсы, группы памяти, сообщения в группах учебных заведений, комментариев и сообщений родственников в социальной сети (верификация проходила по минимум двум информационным источникам).

Со страниц цифровых профилей были извлечены все доступные (с учётом настроек приватности) данные анкеты по выделенным полям и представлены в виде таблицы (Рисунок 4.2). Дополнительно, добавлены два функциональных столбца:

1. № — последовательная нумерация всех аккаунтов и источников, используемых при поиске и верификации пользователя.
2. Источник (ссылка) — гиперссылка на наиболее информативный внешний материал, подтверждающий принадлежность страницы к выборке.



№	Ссылка	Источник	Основное											
			Имя	Фамилия	Пол	Семейное положение	Партнёр	День рождения	Родной город	Языки	Дедушки, бабушки	Родители	Братья, сёстры	Дети	Внуки
1	https://vk.com/aris_1337	Новость http://www.ngregion.ru/obshchestvo/obninsk-do-sikh-por-obsuzhdaet-gibel-17-letnego-podrostka	Аристарх	Филичев	нет	нет	нет	Указан	Обнинск	нет	нет	нет	нет	нет	нет
2	https://vk.com/izzycooper	Информация от подруги https://vk.com/wall-125339469_46582	Елена	Тимохина	нет	нет	нет	Указан	Ростов	нет	нет	нет	нет	нет	нет
3	https://vk.com/id137023742	Новость https://life.ru/t/новости/81698	Настасья	Lazarewa Koroleva	нет	нет	нет	Указан	Лобня	нет	нет	нет	нет	нет	нет
4	https://vk.com/id108582627	Новость https://life.ru/t/новости/81698	Елизавета	Ч(ж)Ц Пец4ля	нет	нет	нет	нет	Лобня	нет	нет	нет	нет	нет	нет
5	https://vk.com/id157425593	Новость https://www.msk.kp.ru/daily/26835/2809253/	Алёнка	Грифская	нет	нет	нет	Указан	Москва	нет	нет	нет	нет	нет	нет
6	https://vk.com/id193675979	Новость https://www.bragazeta.ru/news/2013/12/16/komarichi/	Алина	Цыганова-Вятухна	нет	Влюблена	нет	Указан	нет	нет	нет	нет	нет	нет	нет
7	https://vk.com/jess_mc	Новость http://info.sibnet.ru/article/344963/	Jessica Jess	MC Mouder	нет	нет	нет	Указан	Новосибирск	нет	нет	нет	нет	нет	нет
8	https://vk.com/id215025564	Сообщения на стене от семьи и друзей https://vk.com/wall215025564_572 https://vk.com/wall215025564_571	Valeriya	Evstya	нет	нет	нет	Указан	Москва	нет	нет	нет	нет	нет	нет
9	https://vk.com/id245621285	Информация о сообществе школы https://vk.com/podslushano_shk10	Лера	Арсеньева	нет	нет	нет	нет	нет	нет	нет	нет	нет	нет	нет
10	https://vk.com/himily	Сообщения на стене от семьи и друзей https://vk.com/wall292095738_410 https://vk.com/wall292095738_409	Ирина	Леонидная	нет	нет	нет	Указан	Мурманск	нет	нет	нет	нет	нет	нет
11	https://vk.com/id291560571	Информация о сообществе школы https://vk.com/wall-76484526_26607	Алина	Тарасова	нет	нет	нет	Указан	Волгоград	Русский, English	нет	нет	да	нет	нет
12	https://vk.com/id245485941	http://www.newsvl.ru/accidents/2017/09/25/163306/	Рустам	Ибрагимов	нет	нет	нет	нет	нет	нет	нет	нет	нет	нет	нет
13	https://vk.com/id207853486	Комментарии друзей https://vk.com/wall207853486_222	Алексей	Кондеев	нет	нет	нет	Указан	Boston	нет	нет	нет	нет	нет	нет
14	https://vk.com/id248798277	Новость https://www.sakhalin.info/ys/124922	Екатерина	Макарова	нет	нет	нет	нет	Южно-Сахалинск	нет	нет	нет	нет	нет	нет
15	https://vk.com/id332118075	Новость http://72.ru/text/newsline/2394827461265361.html	Настя	Молокова	нет	нет	нет	Указан	Тюмень	Русский	нет	нет	нет	нет	нет
16	https://vk.com/id372809024	Новость https://www.perm.kp.ru/daily/26590.4/3267279/	Anya	Reshetnikova	нет	нет	нет	Указан	Пермь	нет	нет	нет	нет	нет	нет
17	https://vk.com/iefedomno	Комментарии друзей https://vk.com/wall2078	Фёдор	Пугачев	нет	нет	нет	Указан	нет	нет	нет	нет	нет	нет	нет
18	https://vk.com/young_fox_mox	Новость https://vk.com/wall-73916798_3030592	Пэша	Лис	нет	Не женат	нет	нет	Санкт-Петербург	Русский	нет	да	да	нет	нет
19	https://vk.com/id133556861	Комментарии друзей https://vk.com/wall408089274_384	Дмитрий	Масленников	нет	нет	нет	Указан	Верхний Пышма	Русский	нет	нет	да	нет	нет

Рисунок 4.2 Представление данных о цифровых профилях

Для визуального контроля заполненности применена цветовая градация: заполненные и типовые значения — зелёная заливка, незаполненные — красная, нетипичные варианты (например, абсурдный родной город) — жёлтая.

К типовому заполнению относится информация, являющаяся вероятно правдивой для соответствующего поля: номер телефона, содержащий необходимое количество символов, верные данные об имени и фамилии и так

далее. К нетиповому заполнению относится заведомо ложная информация соответствующая или несоответствующая наименованию поля: ложный или несуществующий город проживания, текстовая информация в поле «номер телефона» и т.д.

Такой подход позволяет оперативно отслеживать распределение пропусков и аномалий. Данные, полученные по каждому полю, анализируются на предмет содержащейся в них информации.

Выявлено, что аккаунты имеют сходства в содержании и наполненности полей, что позволяет на их основе составить базовое описание цифрового объекта, относящегося к «группе риска» — а также выделить критерии для их дальнейшего поиска.

Анализ полей таблицы (Рисунок 4.2) на предмет наполненности и содержательности выявил, что 20 из 41 полей являются малоинформативными и не имеют ярко выраженные особенности. Оставшееся 21 поле выделены в качестве характеристик для наполнения базового описания цифрового объекта.

Базовое описание исследуемого цифрового объекта обладает следующей размерностью:

$$|S| = 21, |D| = 4, |Rel| = 3.$$

Анализ собранных из социальной сети профилей пользователей заключается в обработке полей заранее указанными методами и выставление весовых характеристик [119]. Математическое описание данного процесса можно представить в универсальной форме аппроксиматоров:

$$y(x) = \sum_{i=1}^{N} \varphi_i(x) \cdot \theta_i, \tag{4.3}$$

где x – это профиль пользователя,

i – это индекс поля в профиле,

N – количество всех полей в профиле,

$\varphi_i(x)$ – это метод (или список методов) для обработки i-го поля в профиле x,

θ_i – это весовой коэффициент i-го поля в профиле.

Для определения f_{destr} разработана 28-критериальная балльная модель.

Для перехода от таблицы (Рисунок 4.2) к количественной модели была сформирована рабочая группа из пяти специалистов — психологов, аналитиков данных и системных инженеров. Каждый эксперт ранжировал 21 статическую характеристику по шкале 1...21, где «1» обозначало минимальную значимость, а «21» — максимальную.

Таблица 4.2 Экспертная оценка значимости критериев

№	Критерий	Эксперт 1		Эксперт 2		Эксперт 3		Эксперт 4		Эксперт 5	
1	Имя	21	20,5	21	20,5	20	19	21	20,5	21	16,5
2	Фамилия	21	20,5	21	20,5	20	19	21	20,5	21	16,5
3	Семейное положение	18	17	19	18	3	8	20	18,5	19	2
4	Родной город	7	7,5	1	1	3	8	20	18,5	20	7,5
5	Японский язык	19	18	20	19	21	21	19	17	21	16,5
6	Братья/сестры	15	15	15	16	9	10	18	16	21	16,5
7	Мобильный телефон	12	12	4	3,5	11	11,5	16	13,5	20	7,5
8	Skype	4	5,5	4	3,5	2	4	16	13,5	20	7,5
9	Доп. Телефон	4	5,5	4	3,5	2	4	16	13,5	20	7,5
10	Личный сайт	7	7,5	5	7,5	11	11,5	16	13,5	20	7,5
11	Instagram	2	3	5	7,5	2	4	15	10	20	7,5
12	Twitter	2	3	5	7,5	2	4	15	10	20	7,5
13	Facebook	1	1	5	7,5	2	4	15	10	20	7,5
14	Образование	2	3	4	3,5	3	8	17	4,5	19	2
15	Военная служба	16	16	18	17	1	1	17	4,5	19	2
16	Место работы	20	19	14	15	20	19	17	4,5	21	16,5
17	Деятельность	12	12	13	12,5	19	16,5	17	4,5	21	16,5
18	Интересы	12	12	13	12,5	19	16,5	17	4,5	21	16,5
19	Любимые цитаты	12	12	13	12,5	15	15	17	4,5	21	16,5
20	О себе	12	12	13	12,5	13	13	17	4,5	21	16,5
21	Вдохновляют	8	9	12	10	14	14	17	4,5	21	16,5

Для определения веса каждой характеристики были подсчитаны такие показатели как сумма рангов, полученных каждой характеристикой; среднее арифметическое сумм рангов; отклонение суммы рангов каждой характеристикой от среднего арифметического сумм рангов. Расчёты представлены в таблице (Таблица 4.3).

Таблица 4.3 Расчёт веса каждого критерия

Эксперт (m)	Критерий (n)																				
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	20,5	20,5	17	7,5	18	15	12	5,5	5,5	7,5	3	3	1	3	16	19	12	12	12	12	9
2	20,5	20,5	18	1	19	16	3,5	3,5	3,5	7,5	7,5	7,5	7,5	3,5	17	15	12,5	12,5	12,5	12,5	10
3	19	19	8	8	21	10	11,5	4	4	11,5	4	4	4	8	1	19	16,5	16,5	15	13	14
4	20,5	20,5	18,5	18,5	17	16	13,5	13,5	13,5	13,5	10	10	10	4,5	4,5	4,5	4,5	4,5	4,5	4,5	4,5
5	16,5	16,5	2	7,5	16,5	16,5	7,5	7,5	7,5	7,5	7,5	7,5	7,5	2	2	16,5	16,5	16,5	16,5	16,5	16,5
Сумма рангов, полученная каждым критерием ($\sum X_i$)	97	97	63,5	42,5	91,5	73,5	48	34	34	47,5	32	32	30	21	40,5	74	62	62	60,5	58,5	54
Отклонение от средней суммы ранко ($x-\bar{x}$)	42	42	8,5	-12,5	36,5	18,5	-7	-21	-21	-7,5	-23	-23	-25	-34	-14,5	19	7	7	5,5	3,5	-1
Квадрат отклонения сцммы рангов $(x-\bar{x})^2$	1764	1764	72,25	156,3	1332	342	49	441	441	56,25	529	529	625	1156	210,3	361	49	49	30,25	12,25	1
Вес критерия	0,084	0,084	0,055	0,037	0,079	0,064	0,042	0,029	0,029	0,041	0,028	0,028	0,026	0,018	0,035	0,064	0,054	0,054	0,052	0,051	0,047

Для проверки статистической достоверности полученных весов подсчитан коэффициент конкордации (согласованности) экспертов (W):

$$W = \frac{12S}{m^2(n^3-n)} = \frac{12 \cdot 9970}{25(9261-21)} = \frac{119\,640}{231\,000} \approx 0{,}517, \qquad (4.4)$$

где S −сумма квадратов отклонений от среднего,

m – количество экспертов,

n – количество критериев.

Полученное значение коэффициента конкордации ($W = 0{,}517$) отражает высокую степень согласованности мнений экспертов.

В ходе работ по выделению характеристик допускалось, что в процессе использования количество характеристик, отнесенных к ключевым, а также вес выделенных характеристик могут меняться.

Динамическим характеристикам и множеству связей были присвоены собственные весовые значения (Таблица 4.4).

Таблица 4.4 Веса динамических характеристик и множества связей

№	Динамическая характеристика (D)	Вес
1	Аудиозаписи	0,20
2	Статус	0,18
3	Аватар	0,17
4	Видеозаписи	0,13
	Множество связей (Rel)	

5	Кол-во друзей	0,12
6	Кол-во групп	0,10
7	Кол-во подписчиков	0,10

Наибольшую долю в интегральной оценке получили поля, отражающие невербальные признаки самопрезентации. Такой результат эмпирически подтверждает вывод психолингвистов о том, что подростки с деструктивными настроениями склонны выражать состояние преимущественно в неформализованном контенте.

Вторую группу по вкладу в интегральную оценку образуют контактные данные (мобильный телефон, дополнительный номер, сайт) в сочетании с параметрами социальной окружённости. Их вес — следствие двух факторов: во-первых, значительная часть «группы риска» оставляет эти поля пустыми, что математически усиливает отличие от среднестатистического профиля; во-вторых, при наличии номера телефона появляются возможности для быстрой офлайн-верификации, что ценно в практических профилактических программах.

4.3.2 Метод анализа текстовых полей

Для каждой характеристики необходимо разрабатывать свои методы обработки. Кроме методов анализа наличия и отсутствия информации в полях разработан метод анализа текстовых полей на основе заданных словарей ключевых слов.

Например, текстовый статус, динамично изменяемый в цифровом профиле, является уникальной лакмусовой полоской (краткий эмоциональный выпад часто заметно опережает по динамике все остальные изменения), анализировался на основе составленного корпуса ключевых слов, включающего в себя 142 наименования (в их числе лозунги и аббревиатуры).

Было выявлено, что наиболее часто встречаются русские и иностранные ключевые слова. Весь корпус ключевых слов был разбит на 11 разделов. Каждый раздел дополнен значениями, которые могут быть использованы подростками, например, добавлены недостающие числа, буквы и синонимы

слов. Всего было определено 583 ключевых слов (значений). На основе полученных данных был составлен «Базовый тезаурус». К каждому значению были составлены соответствующие поисковые образы для обеспечения точности поиска информации. В таблице (Таблица 4.5) приведены статистические данные по количеству значений в каждом из разделов, а также их примеры.

Таблица 4.5 Количество и примеры значений базового тезауруса

Тип	**Количество значений**	**Пример**
Негативные слова (русские)	129	Гроб Вены
Негативные слова (иностранные)	99	Dark Fuck
Иероглифические символы	86	グ ف
Иконки	79	╪ •
Инверсные буквы	58	ɯ ɐ
Знаки препинания	31	± *
Тематический	29	Кит Паленкова
Латинские буквы	26	G P
Смайлики	19	☹ /ˊ
Греческий алфавит	18	ψ Σ
Числа	9	1 2

Совместно с базовым тезаурусом формируются частные тезаурусы для поиска слов в каждом отдельном из следующих полей цифрового профиля:

- любимая музыка (76 ключевых слов),
- интересы (46 слов и словосочетаний),
- любимые книги (9 наименований),
- языки (12 наименований),
- любимые фильмы (13 ключевых слов),
- любимые игры (16 ключевых слов).

На основании статистического анализа для каждого типа значения базового тезауруса было присвоено свое весовое значение в зависимости от частоты использования и степени возможного эмоционального окраса, что

позволяет гибко реагировать на изменения сленга без полной переоценки модели. Распределение по весам представлено в таблице (Таблица 4.6).

Таблица 4.6 Распределение значений базового тезауруса по весам

Тип	Вес
Негативные слова (русские)	1
Негативные слова (иностранные)	1
Тематический	1
Иероглифические символы	0,75
Иконки	0,75
Инверсные буквы	0,75
Греческий алфавит	0,75
Латинские буквы	0,5
Знаки препинания	0,25
Смайлики	0,25
Числа	0,25

Данный тезаурус является динамичным и возможно его изменение (пополнение/удаление) в соответствии с изменениями анализируемого информационного поля (графическое представление тезауруса – Рисунок 4.3).

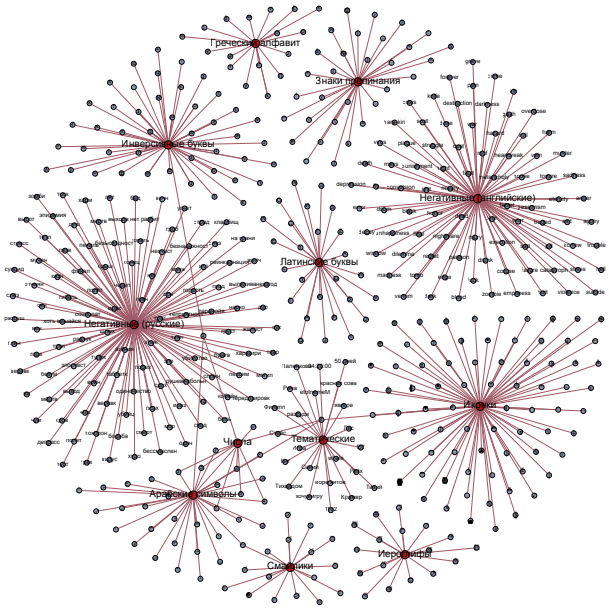


Рисунок 4.3 Графическое представление базового тезауруса

В рамках программно-технической реализации был сформирован соответствующий файл в формате JSON со всеми значениями базового тезауруса. Фрагмент файла приведен на рисунке (Рисунок 4.4).

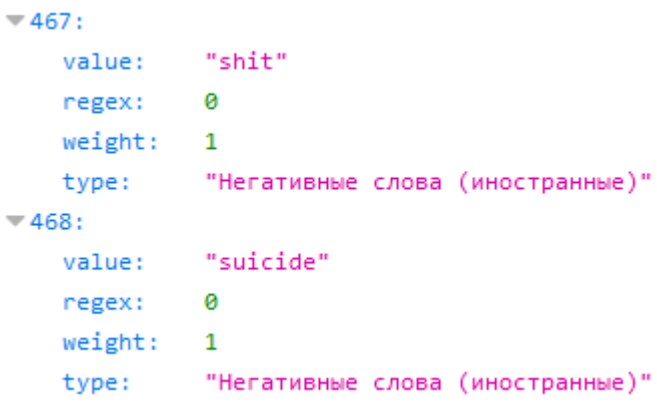


Рисунок 4.4 Фрагмент файла JSON Базового тематического тезауруса

Анализ динамической характеристики «статус» происходит также посредством использования приведенного тезауруса.

Для каждого критерия были разработаны различные методы анализа. Так, например, поля «имя» и «фамилия» анализируются по принципу наличия в поле термина из базового тезауруса (Рисунок 4.5).

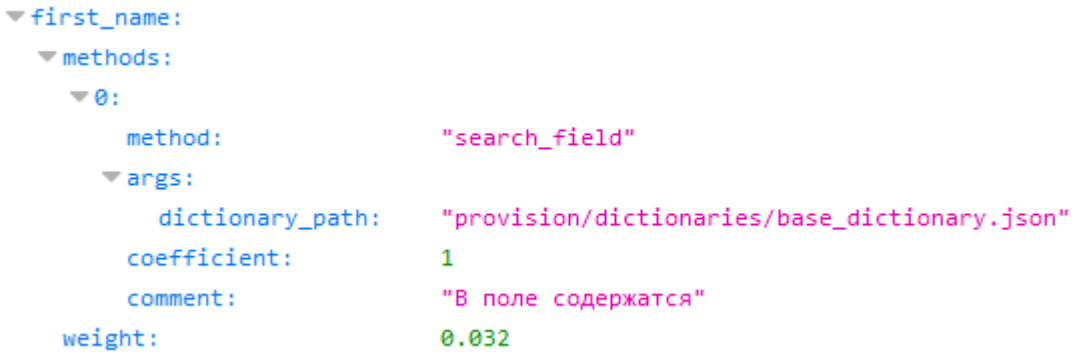


Рисунок 4.5 Фрагмент методики анализа данных по базовому тезаурусу

4.3.3 Метод анализа динамических характеристик

Динамические характеристики цифрового профиля — это живое «эхо» психоэмоционального состояния пользователя в текущем моменте. В этом подразделе приводятся методы обработки аудио- и видеоконтента, фотографии профиля (аватар).

Из профилей обучающей выборки извлечены 26 928 аудиозаписей; после устранения дубликатов осталось 12 839 уникальных трека. Популярная музыка (топ-чарты конкретного года) исключалась вручную, поскольку она не несёт смысловой нагрузки. В результате сформирован подкорпус из 184 композиций, тематически связанных с подростковым протестом, саморазрушением,

наркотическим опытом и смертью. Большинство из них содержат ненормативную лексику. Схема формирования аудиотезауруса приведена на рисунке (Рисунок 4.6).

Композиции распределены по четырём жанровым категориям:

1) Русский рэп депрессивно-агрессивного содержания;
2) Агрессивно-депрессивная (метал, пост-хардкор);
3) Инди-треки меланхолической направленности;
4) Тематическая.

Рисунок 4.6 Схема формирования аудиотезауруса

Каждый тип дополнен наименованиями, которые также могут быть использованы подростками, например, схожие по смыслу композиции или композиции того же автора.

Каждая аудиозапись определяется следующими характеристиками: ID, исполнитель, название, категория и снабжается весовым коэффициентом, зависящим от частоты встречаемости внутри референтной группы.

Фрагмент аудиотезауруса приведен на рисунке (Рисунок 4.7).

```
55:
    audio_id:    456239198
    author:      "ЛСП"
    song:        "Тело (prod. MatoOof)"
    comment:     "Русский рэп"
56:
    audio_id:    456239156
    author:      "ФАРАОН"
    song:        " На твоем теле "
    comment:     "Русский рэп"
```


Рисунок 4.7 Фрагмент аудиотезауруса

Полученный тезаурус был масштабирован, на основе 184 «ядерных» треков, выполнен автоматический сбор свыше 85 тысяч аудиозаписей разных исполнителей, родственников по тематике и лингвистическому ландшафту. Полученный аудиотезаурус может пополняться в онлайн режиме [120].

Графическое представление аудиотезауруса представлено на рисунке (Рисунок 4.8).

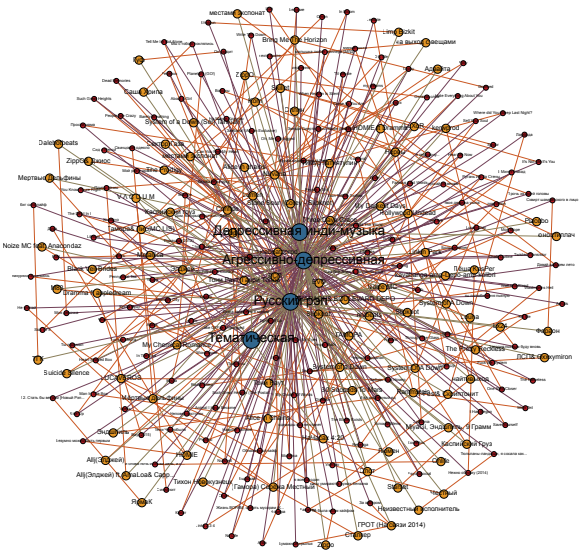


Рисунок 4.8 Графическое представление аудиотезауруса

Поле аудиозаписи анализируется по принципу нахождения в поле маркированных аудиозаписей из «аудио» тезауруса. Методологический файл составлен в формате JSON. Пример методики по анализу аудиозаписей представлен на рисунке (Рисунок 4.9).

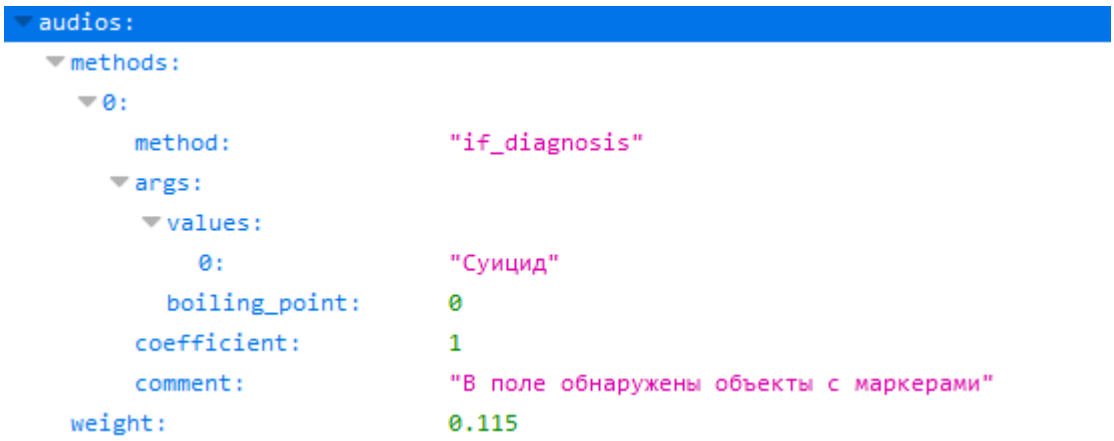


Рисунок 4.9 Фрагмент методики анализа данных по аудиозаписям

Видеораздел у подавляющего большинства профилей обучающей выборки либо пуст, либо скрыт. Тем не менее, обнаруженные ролики сегментируются нейросетевым классификатором, обученным на специальном датасете девиантного характера.

На фотографиях многие подростки намерено скрывают свои лица. Также в качестве фотографий профилей используются изображения с персонажами японских анимационных фильмов «аниме» и животными (в частности, китами).

Поле «аватар» анализируется при помощи нейросетевых технологий, которые фиксируют:

1) наличие/отсутствие человеческого лица;
2) закрытые глаза, маски, анонимизирующие элементы;
3) специфические символы (киты, аниме-персонажи, оружие).

Еще одним из методов анализа динамических характеристик является анализ по количественным значениям. Так, например, целевое значение количества друзей в случае девиантного поведения должно быть меньше 120.

4.3.4 Определение порогового значения функции соотнесения с маркером девиантного поведения

Каждый профиль обучающей выборки описывается вектором нормализованных признаков (статических и динамических характеристик):

$$x = (x_1, \dots, x_{28}), \tag{4.5}$$

где x_i — относительное значение i-го признака после базовой нормализации, $x_i \in [0; 1]$, $i = 1, \dots, 28$.

Финальное интегральное значение вычисляемой характеристики отнесения к «группе риска» вычисляется по формуле

$$f_{destr} = \alpha f_S + \beta f_D, \alpha + \beta = 1, \quad (4.6)$$

где $f_S = \sum_{i=1}^{21} w_i x_i$ — взвешенная сумма статических характеристик,

$f_D = \sum_{j=1}^{7} w_j x_j$— взвешенная сумма динамических характеристик,

β по умолчанию установлена в 0,45 — экспериментально выявленный баланс, при котором точность и полнота достигают оптимального компромисса.

Интегральная оценка определяется по формуле (4.6). Для определения порогового значения f_{destr} проведены расчеты значений характеристики $f_{d\_trainset}$ по каждому цифровому профилю эмпирической базы, собранной на первом этапе исследования. Значение $f_{d\_trainset}$ лежит в диапазоне [0; 1]. Распределение $f_{d\_trainset}$ по обучающей выборке оказалось одномодальным с минимальным зафиксированным значением $f_{d\_trainset_{min}} = 0{,}176$ и пиком плотности в области 0,32.

Критерий классификации устанавливается простым пороговым правилом:

если $f_{destr} \geq f_{d\_trainset_{min}}$, то профиль маркируется как объект «группы риска».

Такой подход гарантирует, что все цифровые профили, использованные при обучении, будут корректно отнесены к целевому множеству, а пороговая величина допускает последующее уточнение на этапе контрольных испытаний.

Практика показала, что единичный маркер решает строго локальную задачу, тогда как социальная реальность многофакторна. Определим

финальное интегральное значение вычисляемой характеристики отнесения школьника к «группе риска» (f_{comb}):

$$f_{comb} = \lambda_1 f_{destr} + \lambda_2(1 - f_{bot}) + \lambda_3 f_{sch}, \sum_{i=1}^{3} \lambda_i = 1, \quad (4.7)$$

где f_{destr} – вычисляемая характеристика девиантного поведения,

f_{bot} – вычисляемая характеристика «робота» в цифровой среде,

f_{sch} – вычисляемая характеристика «школьника» в цифровой среде.

По умолчанию $\lambda_1 = 0{,}6, \lambda_2 = 0{,}3,\ \lambda_3 = 0{,}1$. Таким образом, приоритет отдаётся рисковой составляющей, затем — проверке на техническую достоверность страницы, и затем — возрастному фактору.

4.3.5 Апробация метода идентификации социального объекта

Предложенная методика была реализована в качестве цифровой платформы по сбору, анализу и идентификации целевых профилей, склонных к девиантному поведению. Цифровая платформа реализована на языке Python 3.5. В качестве веб-фреймворка используется Django 2.0.2, для хранения и управления данными задействована база данных PostgreSQL 10. Система написана как кроссплатформенная, разработка ядра системы проходила на операционной системе Windows 10, а испытания проводились на операционной системе Ubuntu Server 16.04.

Для решения задачи сбора информации по пользователям был использован специализированный алгоритм – «Волновой алгоритм сканирования сети» [121, 122]. Принцип алгоритма заключается в том, чтобы не сканировать всю социальную сеть, а распространятся с «очагов», а именно точкой отправления для агентов является выявленный массив цифровых профилей девиантного поведения и выявленных групп депрессивной и суицидальной направленности.

По каждому из цифровых профилей осуществляется анализ по выявленным критериям в соответствии с разработанной диагностической

картой. Итог анализа – отнесение или не отнесение пользователя к целевой группе.

Апробация проводилась на экспериментальном стенде со следующими характеристиками: сервер операционная система Linux, Intel Xeon E5-2650 v4, 2,2 ГГц (2 шт.); Жесткий диск 300 Гбайт, SAS 12 Гбит/с, 10 000 об/мин, (2 шт.); 1.8 Тбайт, 10 об/мин SAS 12 Гбит/с (6 шт.); Модуль памяти RDIMM 16 Гбайт, 2 400 MT/c(8 шт.); NVIDIA Tesla M60 GPU (1 шт.).

Проведены тестовые испытания методики идентификации целевых объектов на выборке из 800 цифровых профилей. В результате поисковых испытаний определено целевое значение вычисляемой характеристики $f_{d\_trainset} = 0{,}3$, при которой однозначно идентифицируется цифровой объект принадлежащий к девиантному поведению, а также подтверждены методики расчета вычисляемых характеристик «школьник» по параметрам – возраст, наличие школы и ее номера в аккаунте, и «робот».

Предложенный автором подход к решению задач идентификации целевых объектов использовался в разработке информационно-аналитической платформы агентного поиска целевых объектов в социальных сетях в соответствии с государственным заданием Минобрнауки России №2.12915.2018/12.1 «Разработка и апробация информационной системы комплексной антисуицидальной интернет-профилактики», результаты тестирования зафиксированы в акте тестовых испытаний информационной системы комплексной антисуицидальной интернет-профилактики от 25 мая 2018 г. 1/ДСП .

ВЫВОДЫ ПО ГЛАВЕ 4

1. Разработана аналитическая модель профиля цифрового объекта в социальной среде, на основе предложенной автором комплексной модели цифрового объекта, характеризующаяся существенным значением связей с другими объектами (Rel). Определено, что цифровой профиль социального объекта типа персона устойчиво описывается 48 характеристиками.

2. Разработаны и апробированы методы анализа текстовой и аудиовизуальной информации для наполнения комплексной модели цифрового информационного объекта, предложены методы расчета вычисляемых характеристик для выявления различных групп, в том числе «школьник», «робот».

3. Разработана методика решения задач идентификации целевых социальных объектов на основе предложенных методов преобразования разнородной информации цифрового профиля, включая введение системы весовых коэффициентов значимости характеристик и вычисление интегрального показателя соответствия профиля целевому образу. Пороговое правило по этому показателю позволяет надёжно выделять целевые группы пользователей. Разработанная автором методика продемонстрировала эффективность при решении задачи обнаружения групп риска в социальных сетях.

ГЛАВА 5 ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ПУБЛИКАЦИОННОЙ АКТИВНОСТИ И СТРУКТУРИРОВАНИЕ НАУЧНЫХ ДАННЫХ

5.1 Построение интерактивных аналитических панелей по тематическому направлению

Визуализация данных – это процесс представления информации в графической форме, которая может быть оперативно воспринята и понята человеком. Визуализация является критически важным инструментом для любой организации, которая хочет увеличить свою эффективность и принимать обоснованные решения на основе анализа данных. Одним из преимуществ визуализации данных является то, что она позволяет увидеть связи и тренды в больших объемах данных, которые могут быть трудно обнаружены при просмотре таблиц или списков чисел. При грамотном использовании графических средств можно принимать обоснованные решения на основе фактов, а не просто на основе интуиции или субъективного мнения. Для всестороннего анализа данных могут быть использованы динамические аналитические панели (dashboard). Существуют готовые инструменты, с помощью которых появляется возможность построение аналитических панелей на основе данных, хранящихся в базе данных, а также их внедрения в итоговую систему [123, 124].

Elasticsearch используется как распределённое хранилище, поддерживающее полнотекстовый поиск и аналитику по индексированным данным. Для корректного хранения сложных JSON документов настраивается специальная схема типов данных (mapping) в Elasticsearch, учитывающая тип каждого поля (целое число, дата, текст, геокоординаты и т.д.) [125]. Например, числовые поля (год публикации, количество цитирований) хранятся как числовые значения, названия стран или идентификаторы – как строковый Keyword – для возможности фильтрации по точному значению, большие тексты – как Text с поддержкой полнотекстового поиска. Такой mapping

позволяет избегать потери информации при загрузке и обеспечить возможность агрегирования по нужным полям.

Визуализация данных осуществляется с помощью Kibana – веб-интерфейса для построения аналитических панелей поверх Elasticsearch. Kibana обеспечивает интерактивный анализ: выборка данных, построение графиков, фильтрацию, поиск по индексам. Преимущество этого инструмента – обновление данных в реальном времени: новые данные, загружаемые в Elasticsearch, автоматически отражаются на панелях. Таким образом, после выполнения этапов сбора и обработки, специалисты получают доступ к интерактивным аналитическим панелям по каждой тематике, содержащим актуальную информацию для дальнейшего исследования.

Графический интерфейс Kibana содержит три основных раздела: «Discover» – просмотр доступных данных, «Visualize» – визуализация собранных данных, «Dashboard» – основная информационная панель.

Раздел «Discover» позволяет исследовать данные с помощью стандартных методов обнаружения Kibana, представлен на рисунке (Рисунок 5.1). В разделе появляется доступ к каждому документу в каждом индексе, который соответствует выбранному шаблону индекса на панели. Можно отправлять поисковые запросы, фильтровать результаты поиска и просматривать данные документа. Также можно просмотреть количество документов, соответствующих поисковому запросу, и получить статистику значений полей. Если для выбранного шаблона индекса настроено поле времени, то распределение документов по времени отображается в виде гистограммы в верхней части страницы.

Раздел «Visualize» позволяет создавать визуализации данных в индексах Elasticsearch. Визуализация Kibana основана на запросах Elasticsearch. Kibana предлагает несколько видов визуализации: вертикальные и секторные диаграммы, отображение данных на карте, таблицы данных. Визуализацией можно поделиться с другими пользователями, которые имеют доступ к Kibana.

Также существует возможность создавать визуализации из поиска, сохраненного в «Discover», или начинать с нового поискового запроса.

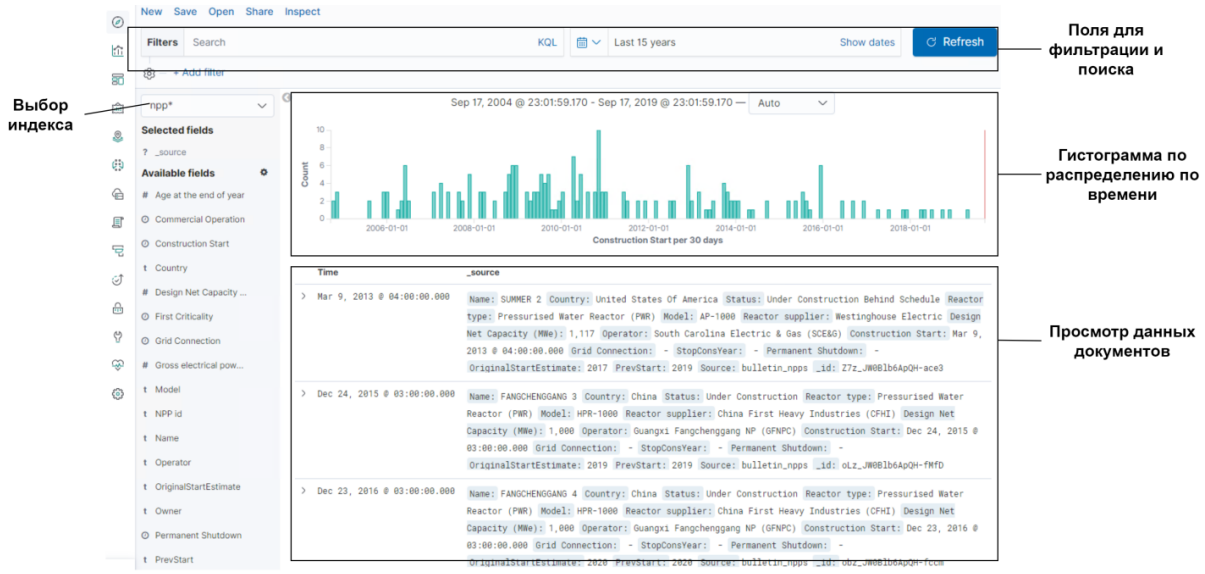


Рисунок 5.1 Раздел «Discover» в плагине Kibana

На основе визуализации, созданной в разделе «Visualize», можно создавать информационные панели в разделе «Dashboard», которые объединяют несколько единичных визуализаций в одну страницу и могут фильтровать их. Использование раздела «Dashboard» помогает получить полный обзор по выбранным данным и сравнить несколько визуализаций, пример представлен на рисунке (Рисунок 5.2). Появляется возможность фильтровать данные в информационной панели, изменив фильтр времени или выбирая элементы визуализации. К примеру, выбрав один из сегментов диаграммы, можно получить подробные данные о нём. Все фильтры данного раздела работают так же, как в разделе «Discover», но в данном случае они применяются только к определённому набору данных.

Таким образом, использование поискового сервиса Elasticsearch вместе с плагином Kibana способствует поиску целевых объектов и помогает «понять» данные с помощью таблиц, диаграмм и различных визуализаций. Одно из самых главных преимуществ использования данного инструмента – обновление данных в реальном времени, то есть если в поисковую систему Elasticsearch будут загружены новые данные, то они автоматически отобразятся и в разделах Kibana.

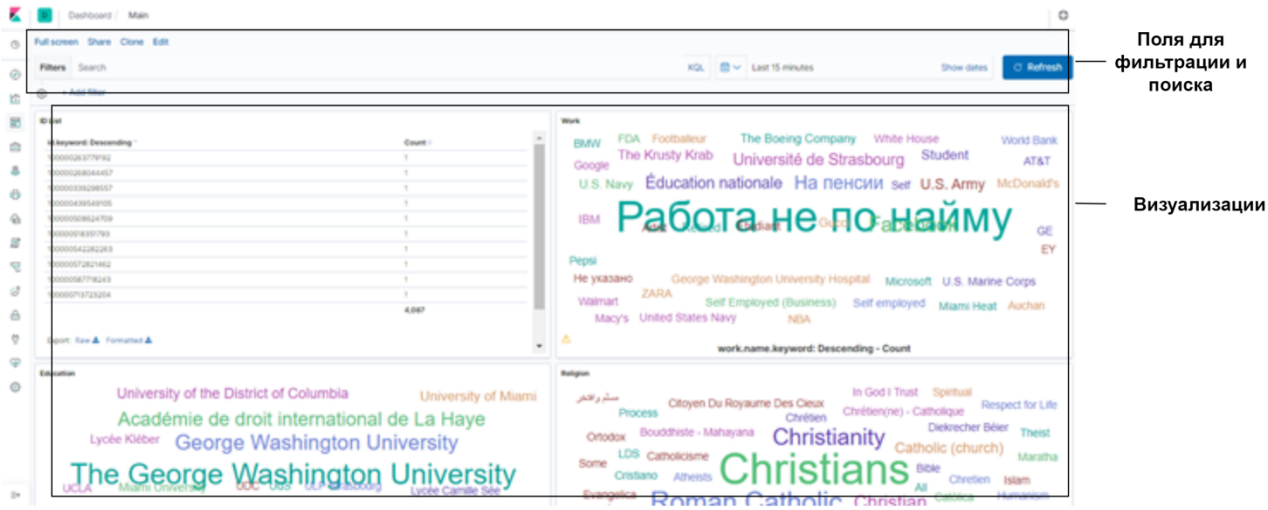


Рисунок 5.2 Раздел «Dashboard» в плагине Kibana

В разделе 2.3 предложена аналитическая модель цифрового объекта, разработанная автором, описывающая структуру и свойства научной публикации. На основе этой модели в работе решаются следующие задачи:

1) **Интеграция разнородных данных.** Проводится сбор и структурирование информации из различных источников по трём тематическим направлениям: технологиям больших данных, медицине и финансовой безопасности. Для каждого направления используются профили научных данных: публикации в научных журналах, данные о проектах, а также отраслевые источники (например, клинические рекомендации в медицине, специализированные журналы по финансовой безопасности).

2) **Автоматизированная обработка и насыщение данных.** Применяются разработанные алгоритмы извлечения и насыщения данных для обогащения цифрового профиля каждой публикации метаданными: выделение ключевых слов, идентификация аффилиаций авторов и их геолокаций, определение тематических классификаций, объединений стран и т.д. Особое внимание уделяется слабоструктурированным данным – текстам научных статей, таблицам, изображениям – их преобразованию в унифицированный JSON формат.

3) **Разработка интерактивных аналитических панелей.** Выполняется визуализация обработанных данных с помощью интерактивных

панелей, позволяющих анализировать публикационную активность по выбранным фильтрам (время, страна, организация, ключевое слово и др.) и отслеживать динамику развития тематик. Для каждой области создаются панели с набором графических элементов: временные ряды, диаграммы, географические карты, таблицы и облака тегов, обеспечивающие мультиаспектный анализ.

4) **Анализ и выводы по направлениям.** На базе разработанных панелей проводится интеллектуальный анализ публикационной активности в выбранных областях. Формируются статистические метрики и выявляются закономерности: лидеры по числу публикаций (страны, организации, авторы), динамика публикаций по годам, основные тематические ключевые слова и их эволюция во времени, степень международной кооперации и др. По каждому направлению делаются выводы о текущем состоянии и тенденциях развития, сравниваются особенности трех областей.

Актуальность такой работы обусловлена необходимостью получать целостное представление о состоянии наук в приоритетных сферах (большие данные, медицина, финансовая безопасность) и оперативно выявлять новые тренды. Интеграция разнородных источников – от научных статей, патентов до государственных реестров и отчетов – и их унификация в рамках одной системы аналитики позволяет исследователям самостоятельно (без участия программиста) извлекать нужную информацию и проводить анализ больших массивов данных. В отличие от существующих решений, ориентированных на строго структурированные базы данных, предлагаемый автором подход нацелен на слабоструктурированные данные научно-технической информации. Это обеспечивает более глубокое покрытие информационного поля и гибкость в работе с разноплановыми данными [126, 127].

Рассмотрим примеры таких аналитических панелей в решении задач анализа публикационной активности в трёх выбранных областях – большие данные, медицина и финансовая безопасность. Каждая область имеет свои особенности с точки зрения доступности данных и их структуры. Тем не

менее, предложенный автором в работе единый подход (модель данных + архитектура) обеспечивает универсальный каркас для интеллектуального анализа данных.

5.1.1 Организация сбора и обработки данных для озера данных по «Финансовой безопасности»

Рассмотрим организацию построения аналитической панели на примере тематического направления «Финансовая безопасность». Эта тематика относится одновременно к экономической и правовой сфере и включает исследования по противодействию отмыванию денег, финансовым мошенничествам, рискам в банковском секторе и т. д. Особенностью является отсутствие его явной идентификации в классических научных рубрикаторах (например, в базах Web of Science или Scopus нет отдельной категории “Financial Security”). Поэтому сбор данных требовал предварительной разработки тематического рубрикатора и поиска профильных источников.

В рамках выполнения задачи формирования озера данных публикаций было выбрано научное издание Journal of Money Laundering Control в качестве стартовой точки. На декабрь 2022 года в данном журнале имеется 100 выпусков в 25 томах с общим количеством публикаций 1008. Статьи данного научного журнала публикуются с 1997 года. На основе собранных данных был проведен обзор тематического направления «Финансовая безопасность», в результате которого выявлены основные ключевые слова, описывающие область исследований, а также информационные источники. Таким образом, в результате обзора тематического направления и анализа собранных данных были выбраны новые источники из списка используемой литературы, а также ключевые слова статей, по которым производился следующий этап отбора. На основе предложенных автором методов и разработанных средств, представленных в разделе 3 были проведены процедуры сбора, обработки и насыщения данных для загрузки в ElasticSearch.

После загрузки данных всех собранных научных публикаций в ElasticSearch была сформирована информационная панель для анализа

информации по тематическому направлению «Финансовая безопасность» с различными типами визуализаций в Kibana. Сформированная аналитическая панель по финансовой безопасности стала эффективным инструментом для визуальной аналитики. Она предоставляет как сводные показатели, так и детальные разрезы данных.

5.1.2 Описание работы с озером данных публикаций по «Финансовой безопасности»

В подразделе представлено описание возможностей панели визуализаций. Данная панель была построена в программной утилите для визуализации данных Kibana, реализованная на основе нереляционной базы данных ElasticSearch.

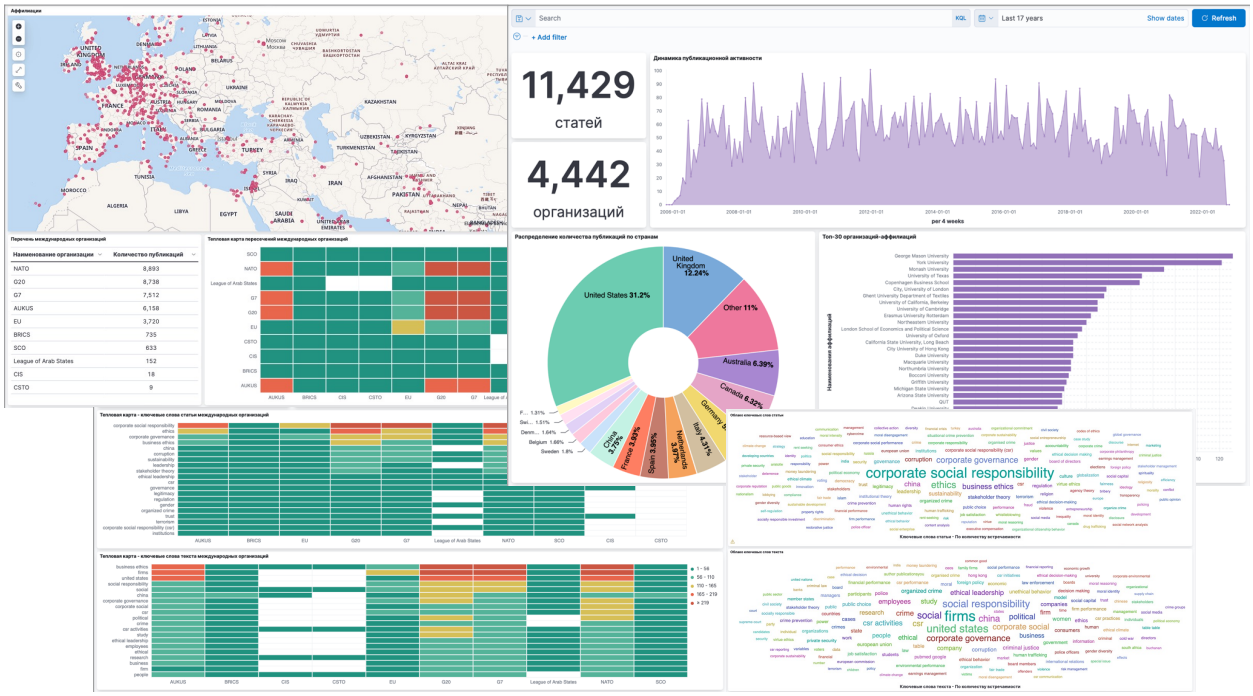


Рисунок 5.3 Визуализации данных по направлению «Финансовая безопасность»

На панели представлено 14 визуализаций разного типа (Рисунок 5.3): от круговой диаграммы до карты. Для проведения анализа научно-технической информации по тематике «Финансовая безопасность» было собрано 11383 научных публикаций. Данное количество представлено на первой визуализации, расположенной в левом верхнем углу (Рисунок 5.4). Число

отражает количество публикаций, данные которых визуализированы в рамках данной панели.

Рисунок 5.4 Метрика, отражающая количество публикаций, представленных на панели

На второй визуализации, расположенной под первой, представлена другая метрика, отражающая количество организаций-аффилиаций авторов публикаций данной выборки (Рисунок 5.5).

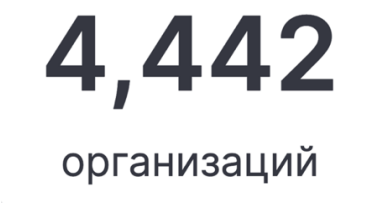


Рисунок 5.5 Метрика, отражающая количество организаций-аффилиаций авторов публикаций, представленных на панели

Сверху справа от предыдущих визуализаций расположена временная шкала, отражающая динамику публикационной активности (Рисунок 5.6). По оси *x* расположена временная шкала с делением по годам, по оси *y* – числовая шкала, отражающая количество публикаций.

График является интерактивным. Выбрав интересующий временной отрезок на графике, можно задать фильтр, в следствии которого панель визуализаций перестроится таким образом, что останутся только те статьи, которые были опубликованы в выбранный отрезок времени.

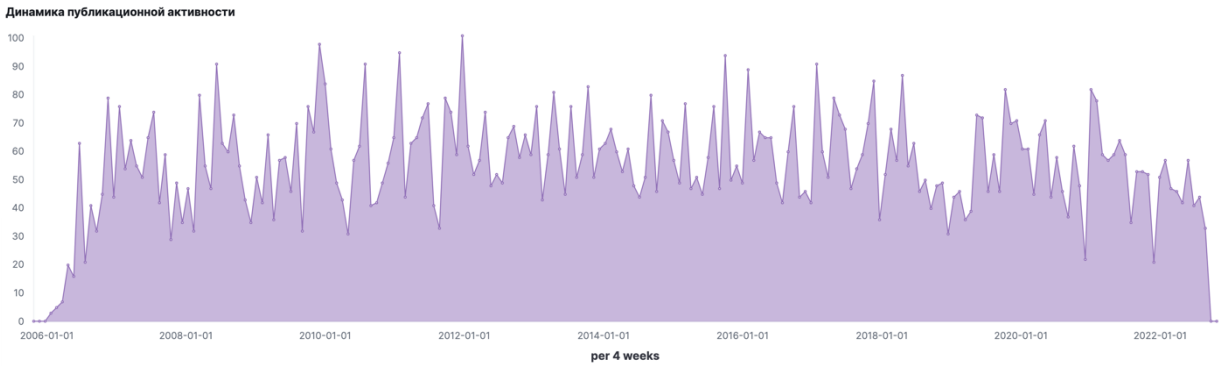


Рисунок 5.6 Временная шкала, отражающая динамику публикационной активности

Следующий график, расположенный ниже предыдущих визуализаций, представляет собой круговую диаграмму, отражающую вклад стран авторов публикаций в исследованиях по данной выборки в процентах (Рисунок 5.7).

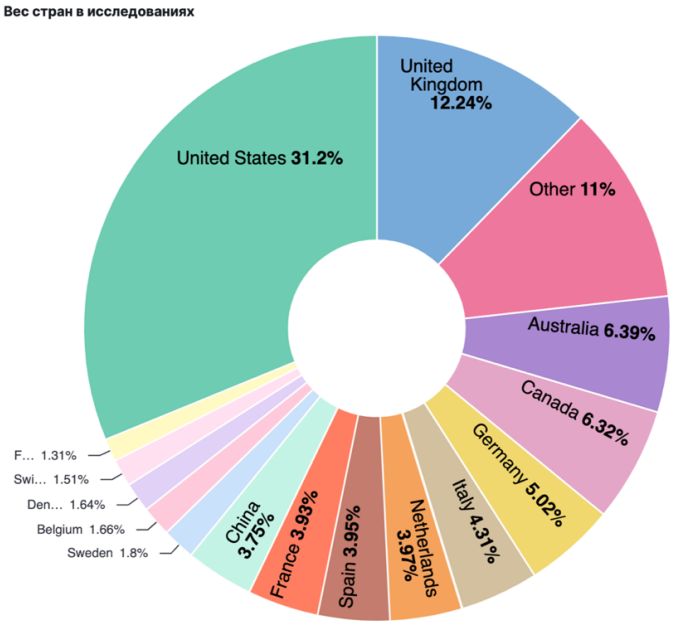


Рисунок 5.7 Круговая диаграмма, отражающая вклад стран авторов публикаций в исследованиях по данной выборке

График также является интерактивным. Нажав на интересующую страну, задается фильтр, который перестраивает панель, оставляя публикации только выбранной страны. Географический анализ представлен круговой диаграммой стран, показывающей процентный вклад различных государств в совокупный массив статей. Лидерство удерживают экономически развитые страны: на первом месте США (~36% публикаций), далее Великобритания,

Китай, Германия и Австралия. Интерактивность диаграммы позволяет, например, кликнуть на сегмент “United States” – и панель перестроится, отобразив только ~4093 публикации из США (Рисунок 5.8.). Одновременно все другие графики обновятся: так можно мгновенно получить сведения, какие организации, темы и журналы доминируют именно в американском сегменте исследований по финансовой безопасности.

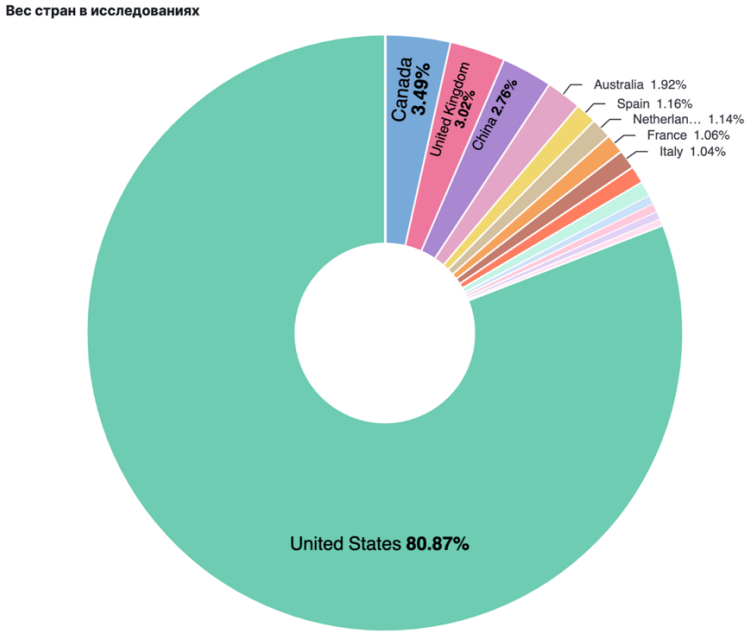


Рисунок 5.8 Измененный вид круговой диаграммы в соответствии с заданным фильтром по «United States»

Для анализа институциональной структуры исследований на панели имеется горизонтальная гистограмма топ-30 организаций по числу публикаций (Рисунок 5.9). Здесь видны ведущие игроки: например, на первых позициях Университеты США (Harvard, MIT и др.), крупные британские центры (Oxford, London School of Economics) и международные организации (МВФ, Всемирный банк). График позволяет фильтровать данные по конкретной организации, что дает возможность изучить профиль ее исследований.

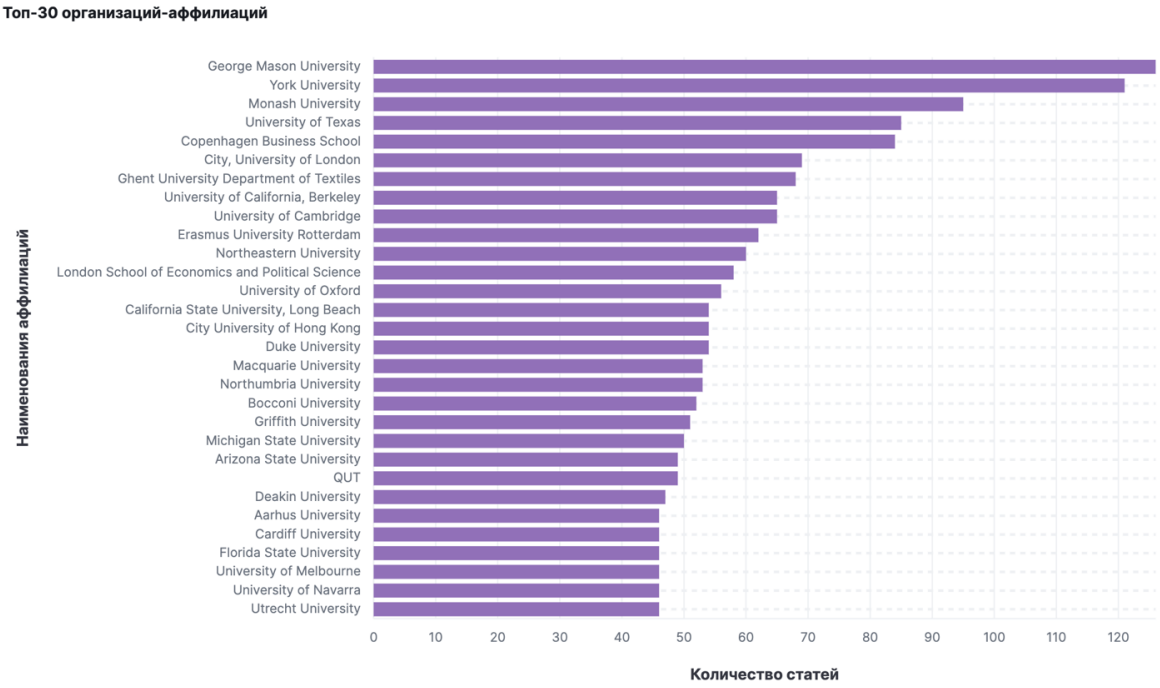


Рисунок 5.9 Горизонтальная гистограмма с 30 организациями-аффилиациями с наибольшим количеством публикаций из выборки

Также на панели расположена карта, на которой в качестве розовых кругов отображены местоположения организаций-аффилиаций (Рисунок 5.10). Карта является интерактивной. Панель позволяет приближать или отдалять объекты: чем ближе объект, тем точнее его местоположение отображено на карте.

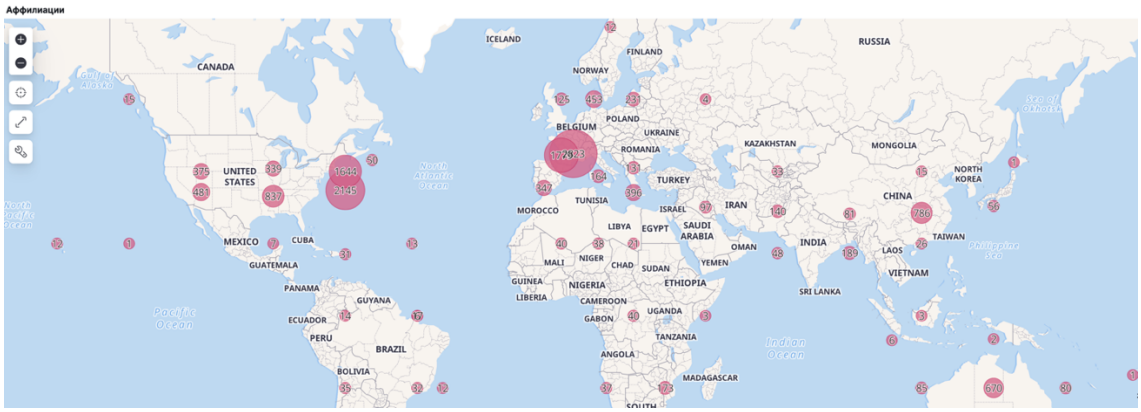


Рисунок 5.10 Карта с местоположениями организаций-аффилиаций авторов публикаций выборки

Для работы с картой слева расположена панель с инструментами. Иконки со знаками + и – позволяют менять масштаб карты. С помощью

третьей иконки сверху возможно вручную задавать значения широты и долготы объекта и значение зума (Рисунок 5.11).

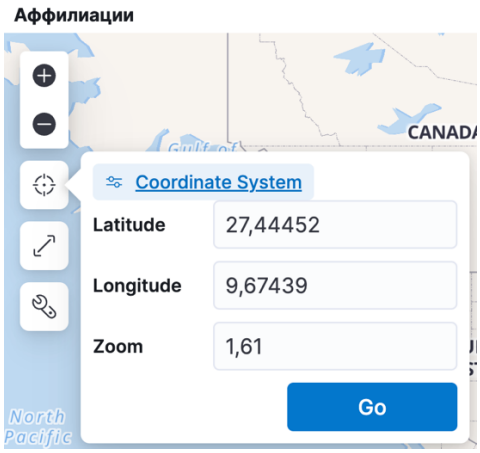


Рисунок 5.11 Панель управления картой

Ниже карты расположены две визуализации. Слева расположена таблица с перечнем международных организаций и количеством публикаций стран-участников (Рисунок 5.12).

Публикации международных организаций

Наименование организации	Количество публикаций
NATO	8,893
G20	8,738
G7	7,512
AUKUS	6,158
EU	3,720
BRICS	735
SCO	633
League of Arab States	152
CIS	18
CSTO	9

Рисунок 5.12 Таблица с перечнем международных организаций и количеством публикаций стран-участников

Таблица также является интерактивной. При наведении на любое значение появляются три иконки (Рисунок 5.13).

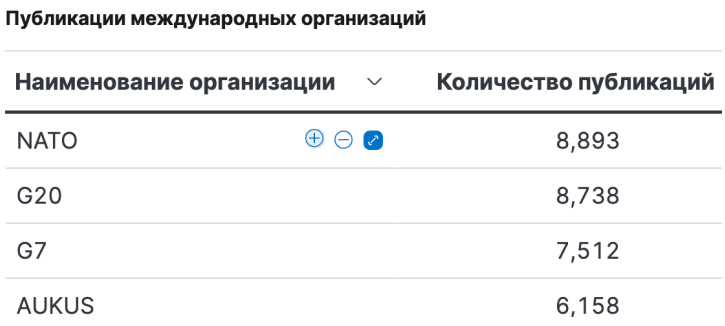


Рисунок 5.13 Инструменты фильтрации графика табличного вида

При нажатии на иконку с «+» задается фильтр, который оставляет лишь те статьи, которые были опубликованы авторами организаций-аффилиаций, расположенных в странах, которые является участниками выбранной международной организации. Соответственно, если нажать на кнопку с «–», напротив, задается фильтр, который убирает из данной таблицы выбранную организацию, а для других визуализаций оставляет только те статьи, которые были опубликованы авторами не из стран-участников выбранной международной организации. Иконка с двойной стрелкой выводит панель, на которой можно выбрать те же фильтры с иконками с «+» и «–» (Рисунок 5.14).

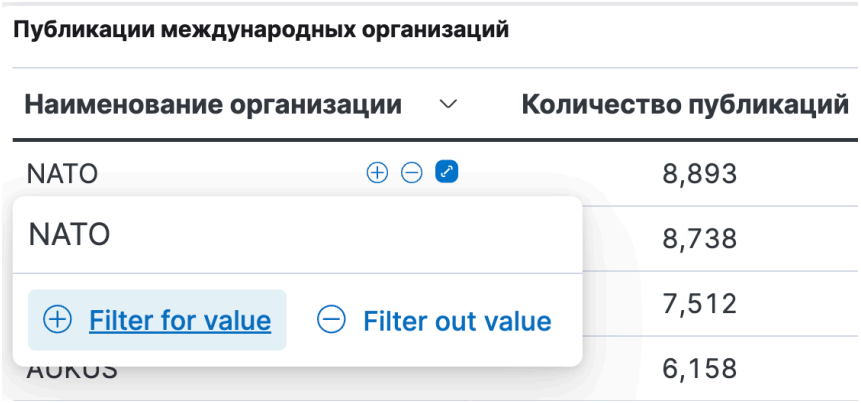


Рисунок 5.14 Создание фильтра в таблице с помощью иконки с двойной стрелкой

График, расположенный справа от предыдущей таблицы (Рисунок 5.3), является тепловой картой. В качестве строк и столбцов тепловой карты заданы международные организации, а значения в ячейках – это количество публикаций, написанных авторами стран-участников организаций. В таблице (Рисунок 5.15) перечислены ключевые международные организации (ООН, G7, G20, НАТО), а напротив каждой указано количество публикаций, авторы

которых аффилированы со странами-членами данной организации. Это позволяет оценить вклад международных инициатив в развитие исследований. Таблица интерактивна: можно одним кликом оставить на панели только публикации стран-участниц конкретной организации или, наоборот, исключить их.

Интенсивность цвета показывает плотность пересечения. Такая визуализация помогает выявить, например, взаимодействие между институтами: яркое пересечение ООН и Всемирного банка указывает, что в странах, принадлежащих обеим организациям, тематическое направление «Финансовая безопасность» особенно активно изучается.

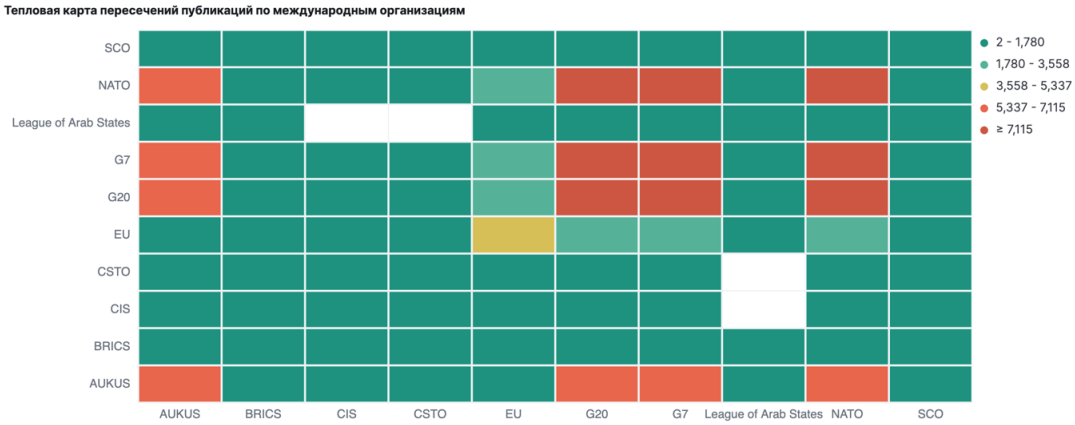


Рисунок 5.15 Тепловая карта пересечения публикаций международных организаций

При наведении на ячейку высвечивается информационная панель со значениями двух организаций и количеством публикаций, актуальными для выбранной ячейки (Рисунок 5.16).

Содержательная сторона исследований по «Финансовой безопасности» проанализирована с помощью методик работы с текстами.

Во-первых, построены две тепловые карты (Рисунок 5.17), где по строкам отложены топ-20 наиболее популярных ключевых слов, а по столбцам – международные организации (страны-участницы которых упоминаются в публикациях). Первая карта отражает ключевые слова, заданные самими авторами статей, вторая – наиболее часто встречающиеся термины из текстов публикаций. Эти карты позволяют сопоставить заявленную авторами тематику

с реальным содержанием работ и проследить, какие аспекты финансовой безопасности привлекают внимание исследователей из разных объединений стран.

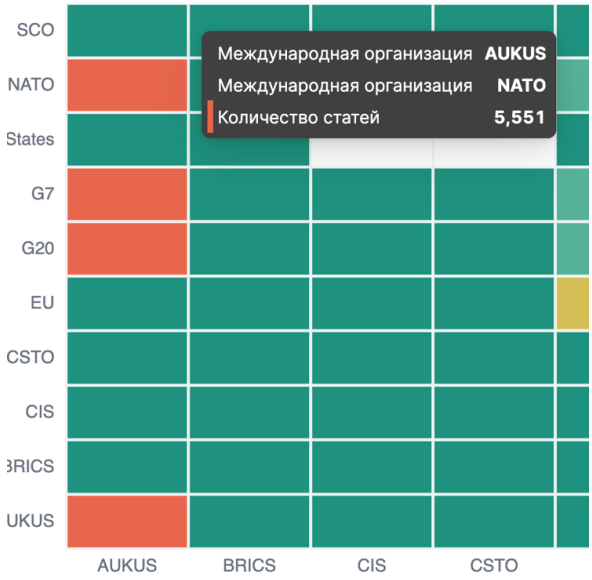


Рисунок 5.16 Фрагмент тепловой карты с количеством совместных публикаций AUKUS и NATO

В качестве строк заданы топ-20 ключевых слов: на первой тепловой карте – ключевые слова, заданные авторами публикаций, на второй – наиболее часто встречающиеся слова и термины в текстах публикаций, в качестве столбцов – международные организации (Рисунок 5.17).

Рисунок 5.17 Тепловые карты, на которых представлено пересечение ключевых слова публикаций и международных организаций

На рисунке 5.18 представлены два облака ключевых слов (Рисунок 5.18). Размер ключевого слова напрямую зависит от количества его встречаемости в публикациях: чем больше слово, тем чаще оно встречается в публикациях (Рисунок 5.18). В каждом облаке показаны ~150 наиболее встречаемых слов. Видно, что среди авторских тегов доминируют общие термины (“financial crime”, “risk management”), тогда как автоматический анализ текстов выявляет также специфические термины (“Basel III”, “cryptocurrency”, “AML regulation” и т.д.), указывающие на конкретные проблемы и регуляторные меры. Облака тегов интерактивны: щелчок по слову мгновенно фильтрует панель, оставляя только публикации, где этот термин фигурирует. Таким образом, исследователь может буквально в один клик сформировать выборку статей по интересующей узкой теме (например, все работы, связанные с криптовалютами и отмыванием денег).

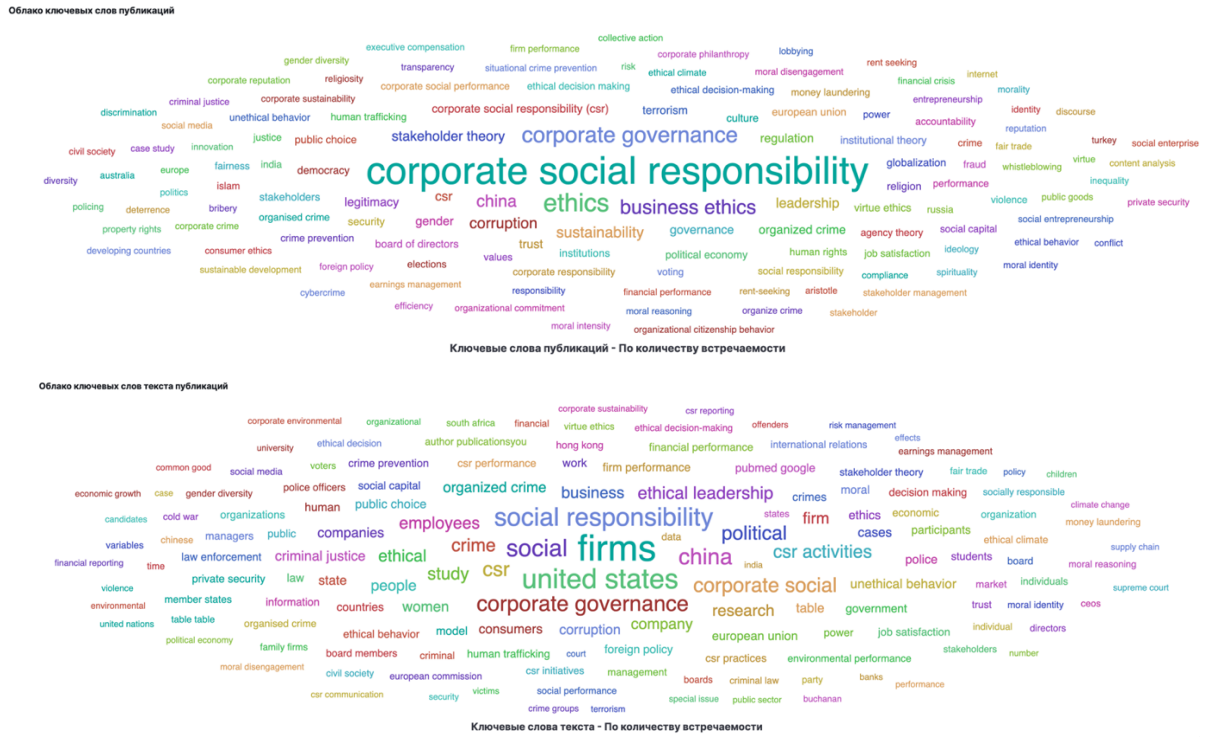


Рисунок 5.18 Облака ключевых слов публикаций

Для обеспечения прозрачности данных панель включает списочные представления. В табличном виде выведен перечень 1000 последних публикаций (Рисунок 5.19) с указанием названия, DOI и даты. Сортировка по дате позволяет отследить наиболее актуальные работы. Все DOI являются

гиперссылками на оригиналы статей, что облегчает переход к полным текстам. Рядом приведена таблица журналов по убыванию количества публикаций (Рисунок 5.20), которая демонстрирует, в каких изданиях тема финансовой безопасности публикуется чаще всего (в лидерах оказались Journal of Financial Crime, Security Journal, Journal of Money Laundering Control и др.). Эти таблицы также оснащены инструментами фильтрации: можно, например, оставить в списке только журналы от одного издателя или исключить определенное издание и пересчитать распределение.

Перечень публикаций (1000)

Название публикации	DOI	Дата публикации
A Decision Theory Perspective on Wicked Problems, SDGs and Stakeho...	https://doi.org/10.1007/s10551-0...	2022-06-27
A Critique of Vanishing Voice in Noncooperative Spaces: The Perspecti...	https://doi.org/10.1007/s10551-0...	2022-03-29
A Framework for Leader, Spiritual, and Moral Development	https://doi.org/10.1007/s10551-0...	2022-03-29
A Critique of Utilitarian Trust: The Case of the Dutch Insurance Sector	https://doi.org/10.1007/s10551-0...	2022-01-28
A Biographical Perspective on Processes of Radicalisation	https://doi.org/10.1007/s10610-0...	2021-09-30
A Framework for Authentic Ethical Decision Making in the Face of Gran...	https://doi.org/10.1007/s10551-0...	2021-09-30
A Crisis in Leadership: Transforming Opportunistic Leaders into Leader...	https://doi.org/10.1007/s10997-0...	2021-01-03
A Configurational Analysis of the Causes of Consumer Indirect Misbeha...	https://doi.org/10.1007/s10551-0...	2020-10-05
A Commons Strategy for Promoting Entrepreneurship and Social Capita...	https://doi.org/10.1007/s10551-0...	2020-07-07
1956 in Hungary: as I saw it then and as I see it now	https://doi.org/10.1007/s11127-0...	2020-05-08
A 5* Destination: the Creation of New Transnational Moral Spaces of R...	https://doi.org/10.1007/s10767-0...	2020-05-08
A Cultural Sociology of Populism	https://doi.org/10.1007/s10767-0...	2020-05-08
100% sure bets? Exploring the precipitation-control strategies of fixed-...	https://doi.org/10.1007/s10611-0...	2019-12-10

Рисунок 5.19 Перечень 1000 публикаций по убыванию даты публикации

Перечень изданий

Наименование издания	Количество публикаций
Journal of Business Ethics	5,697
Public Choice	1,523
Crime, Law and Social Change	734
International Politics	634
European Journal of Law and Economics	573
Security Journal	449
European Journal on Criminal Policy and Research	398
Journal of Management & Governance	312
International Journal of Politics, Culture, and Society	292
Asian Journal of Criminology	260
Trends in Organized Crime	260
Journal of Management and Governance	225
International Journal of Politics, Culture, and Society IJPS	26

Рисунок 5.20 Таблица со списком журналов по убыванию количества публикаций

Кроме того, что представленные визуализации являются интерактивными и позволяют задавать фильтры, возможна фильтрация по заданному пользователем периоду (Рисунок 5.21).

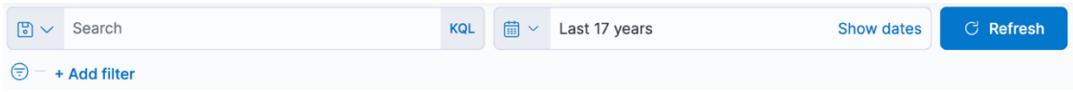


Рисунок 5.21 Инструменты управления панелью визуализации

Поисковая строка позволяет проводить поиск по всем полям, в частности для того, чтобы осуществить поиск по определенному полю, необходимо ввести название интересующего поля, далее двоеточие «:» и нужный термин/слово. Если интересующий термин содержит более одного слова, то его необходимо заключать в кавычки. Регистр в рамках запроса не имеет значение.

Пример запроса представлен на рисунке (Рисунок 5.22).

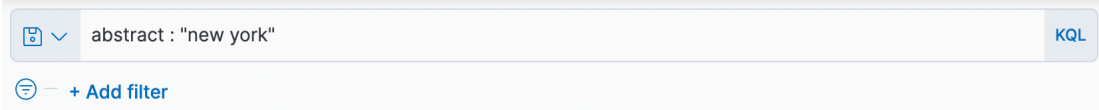


Рисунок 5.22 Пример запроса в поисковой строке

Справа от поисковой строки расположена панель управления временем. В рассматриваемой выборке имеется единственное поле даты, которая соответствует дате публикации статьи. С помощью панели управления временем возможно управлять панелью визуализации и оставлять лишь те публикации, которые соответствуют заданному временному промежутку. В панели представлены три вкладки: Absolute – инструмент для работы со временем в абсолютном формате (Рисунок 5.23); Relative – инструмент для работы со временем в относительном формате (Рисунок 5.24); Now – работа в настоящем времени.

Относительный формат позволяет выбирать промежуток времени в следующем виде: начальная точка – это сколько минут, часов, дней, недель или лет назад; конечная точка – это сколько минут, часов, дней, недель или лет назад или настоящий момент времени.

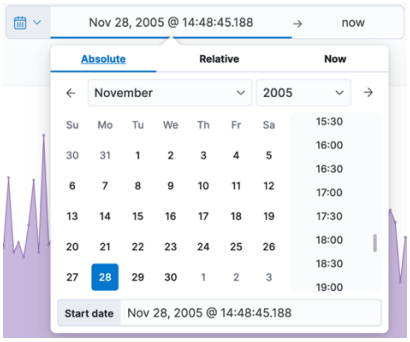


Рисунок 5.23 Пример абсолютного формата времени

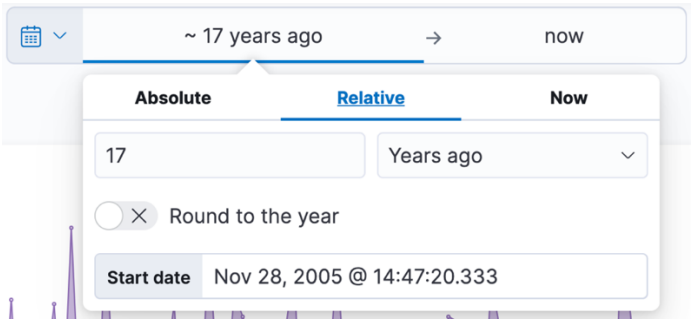


Рисунок 5.24 Пример относительного формата времени

Ниже поисковой строки расположена графа «Add filter», которая позволяет создавать фильтры с любым полем (Рисунок 5.25).

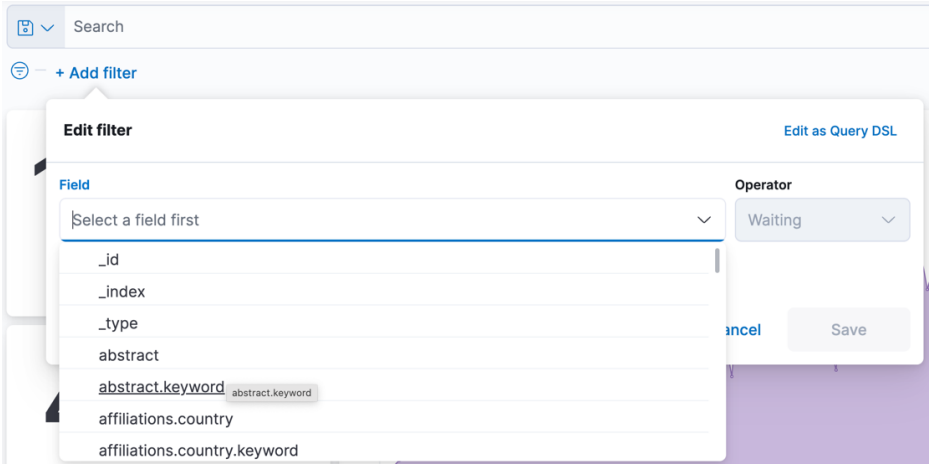


Рисунок 5.25 Инструмент для добавления фильтра

Данный инструмент имеет шесть операторов: is (является), is not (не является), is one of (один из), is not one of (не один из), exists (существует), does not exist (не существует). Оператор is после себя принимает значение, которое соответствует выбранному полю. Например, в поле affiliations.country представлены названия стран, поэтому в графе Value необходимо прописать интересующую страну (Рисунок 5.26). Добавив таким образом фильтр, панель выбирает из выборки только те публикации, которые бы соответствовали заданному фильтру; в случае оператора is, на панели остаются только те публикации, которые были написаны авторам из выбранном страны.

Оператор is not, принимает значения по тому же принципу, что и оператор is, но в результате из панели исключаются публикации, которые были написаны авторами из выбранной страны.

Операторы is one of и is not one of работают по тому же принципу, что и предыдущие два оператора с одним отличием: они принимают после себя несколько значений.

Операторы exists и does not exist не принимают после себя значений. Оператор exist оставляет публикации, в выбранном поле которых (например, affiliations.country) присутствует хоть какое-либо значение, то есть данное поле не является пустым. Оператор does not exist, напротив, оставляет публикации с пустым значением выбранного поля.

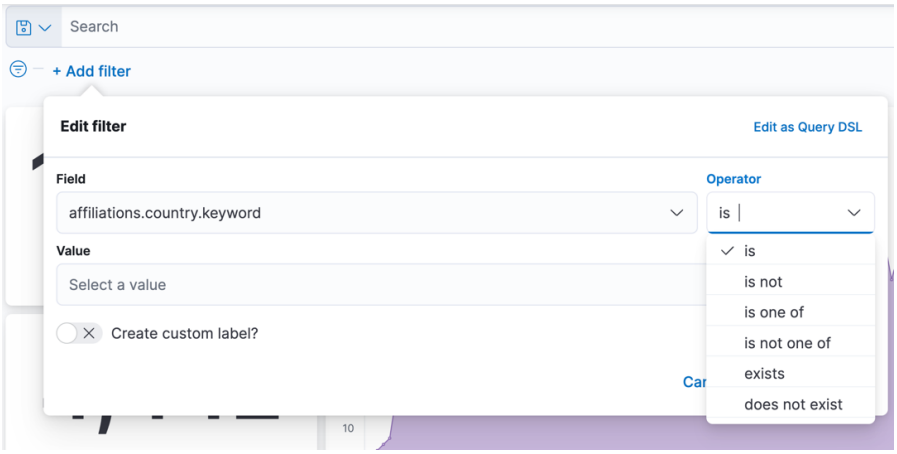


Рисунок 5.26 Инструмент для добавления фильтра с перечнем операторов

Все поля текстового типа представлены в двух видах: поле в изначальном виде и поле с добавлением keyword (например, abstract и abstract.keyword). Если пользователь задает фильтр с помощью обычного поля, то искомое значение ему необходимо вписать самостоятельно (Рисунок 5.27). В то время для полей в названии которых через точку прописано keyword, инструмент предлагает варианты для ввода (Рисунок 5.28).

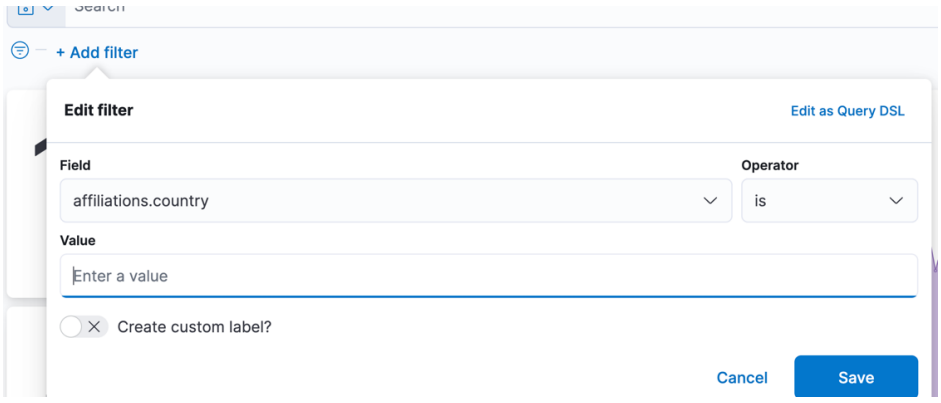


Рисунок 5.27 Добавление значения для полей без keyword в названии

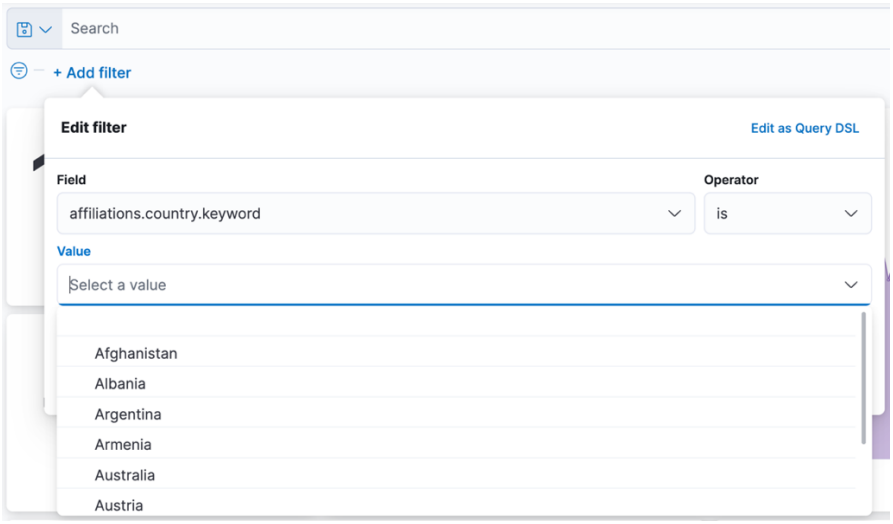


Рисунок 5.28 Добавление значения для полей с keyword в названии

В совокупности, построенное озеро данных и интерактивная аналитическая панель по «Финансовой безопасности» продемонстрировали значимость комплексного подхода для новой области знаний. Несмотря на междисциплинарный характер темы (пересечение экономики, права, ИТ), разработанная система успешно справилась с интеграцией данных и выявлением скрытых закономерностей. Более того, проведенное далее сравнение результатов по трем рассматриваемым направлениям – большим данным, медицине и финансовой безопасности – выявило интересные параллели и отличия.

Например, все три области показывают рост числа публикаций за последнее десятилетие, что согласуется с общей тенденцией на усложнение и цифровизацию мировых проблем. Однако характер международного сотрудничества разнится: в больших данных и медицине наблюдается сильная концентрация исследований в нескольких ведущих странах (США, Китай и др.), тогда как по финансовой безопасности вклад более равномерен между разными странами, отчасти благодаря координирующей роли международных организаций (что видно из анализа пересечений исследований организаций). Тематические тренды тоже имеют свою специфику: медицина фокусируется на конкретных проблемах здоровья и лекарствах, большие данные – на технологических решениях, а финансовая безопасность – на регуляторных и

правовых аспектах. Эти отличия подтверждают необходимость межотраслевого анализа: только рассмотрев несколько областей бок о бок, можно полноценно оценить универсальность и адаптивность инструментария интеллектуального анализа [128, 129, 125, 125, 125].

Апробация результатов работ происходила в рамках Государственного задания Министерства науки и высшего образования Российской Федерации №3466-22 «Создание учебно-методических материалов по финансовой безопасности для школьников и студентов, в том числе для передачи указанных учебно-методических материалов в зарубежные страны-партнеры Международного сетевого института в сфере противодействия отмыванию доходов, полученных преступным путем, и финансированию терроризма» [130, 131].

5.2 Интеграция и анализ публикационной активности по направлениям больших данных и медицинских исследований

5.2.1 Анализ публикационной активности в области большиех данных

Область больших данных была выбрана в качестве первой пилотной тематики для апробации разработанной системы. Данное направление характеризуется стремительным ростом числа исследований, охватывающих как академические публикации, так и отраслевые отчеты в информационной сфере. Был сформирован корпус данных из 34 062 статей по тематике больших данных, собранных с 2011 по 2021 год из 24 различных информационных ресурсов глобальной сети Интернет. В число источников вошли как ведущие академические журналы (например, PLoS ONE, Wiley Online Library), индексируемые реферативными базами (Scopus и др.), так и специализированные новостные порталы о данных и технологиях (например, Datanami и др.).

Для агрегирования этих разнородных данных использовались разработанные автором скрипты парсинга веб-страниц и API; полученные записи унифицировались по принятой информационной модели публикации (метаданные статьи, авторы, аффилиации, аннотация, ключевые слова, полные

тексты при наличии и т.д.). Каждая публикация снабжалась уникальным идентификатором и временными метками для отслеживания динамики цитирования и появления новых работ. В результате в хранилище была получена обширная подборка данных, на основе которой построена интерактивная аналитическая панель.

Разработанная интерактивная аналитическая панель по тематике больших данных (Рисунок 5.29) включает комплекс взаимодополняющих визуализаций для всестороннего анализа публикационной активности. В состав панели вошли:

1) линейный график временного ряда, отображающий динамику количества публикаций во времени (с шагом в четыре недели);

2) две круговые диаграммы – первая иллюстрирует вклад стран в общий массив публикаций (процент и число статей по странам), вторая отражает распределение публикаций по типам источников (академические журналы, конференции, онлайн-ресурсы и т. п.);

3) три таблицы данных: по наиболее продуктивным аффилиациям (организациям), авторам и источникам;

4) два облака тегов ключевых терминов (одно построено на основе ключевых слов, указанных авторами статей, другое – на основе частотного анализа терминов непосредственно в тексте публикаций).

Такое комбинирование метрик и визуальных представлений позволяет выявлять как количественные, так и содержательные аспекты исследуемой области. На рисунке (Рисунок 5.29) крупным планом видно, что среди ключевых слов, указанных авторами, доминируют общие технологические термины (на первом месте — “machine learning”, «машинное обучение»), тогда как в ключевых словах, автоматически извлеченных из текстов, чаще встречаются прикладные понятия (“data analytics”, «аналитика данных»). Этот факт указывает на отличие между авторским позиционированием работы и ее фактическим содержанием: авторы зачастую указывают ключевые слова, относящиеся к теме в целом, тогда как анализ текста выявляет упоминания

конкретных методов и областей применения исследования. Таким образом, облака тегов облегчают обнаружение скрытых тематических акцентов и междисциплинарных связей.

Все элементы интерактивно связаны: например, при выборе отдельной страны на диаграмме фильтруются записи во всех других представлениях, что позволяет детально изучать каждую из стран. На верхнем графике заметен экспоненциальный рост числа публикаций по большим данным после 2015 года, облака тегов и наглядно демонстрируют сдвиг терминологии.

Аналитическая панель предоставляет широкие возможности интерактивного исследования данных. Пользователь системы может в режиме реального времени применять фильтры по различным полям (страна, автор, организация, ключевое слово, год публикации и др.), комбинировать несколько фильтров одновременно, а также изменять временные диапазоны для анализа определенных периодов. Например, можно ограничить рассмотрение только начальным этапом развития области. При анализе публикаций 2008–2011 годов (77 работ) выявлено, что ~67% из них выполнены учеными США, а ~71% опубликованы на онлайн-порталах. Однако в последующие годы наблюдается интернационализация и академизация исследований: доля работ в научных журналах существенно возросла, в лидеры по числу статей кроме США и Китая вышли Великобритания, Индия, Германия и другие страны. Так, по совокупным данным 2011–2021 гг. топ-5 стран выглядят следующим образом: США – 7292 статьи (21,81%), Китай – 5899 (17,32%), Великобритания – 2659 (~7%), Индия – 1848 (5,42%), Россия – 1406 (3,77%).

По ведущим организациям, внесшим наибольший вклад, лидируют учреждения из США и Китая, однако география гораздо шире и охватывает десятки университетов и компаний по всему миру. При этом в некоторых странах отсутствует единый доминирующий центр исследований по большим данным – публикации рассредоточены между множеством организаций, что отражает децентрализованный характер развития данной отрасли.

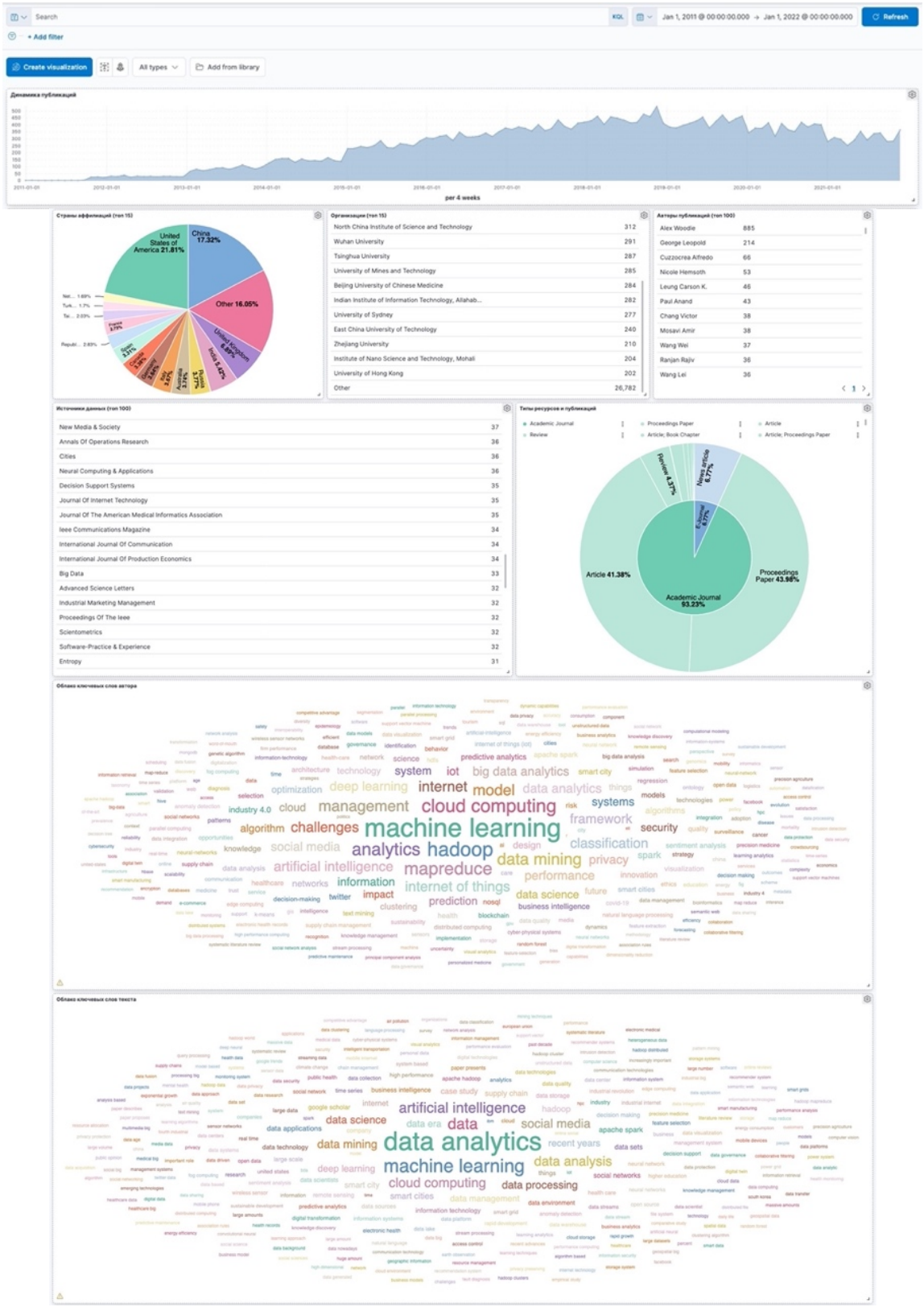


Рисунок 5.29 Интерактивная аналитическая панель публикаций с 2011 по 2021 год по направлению большие данные

Отдельного внимания заслуживает качественный анализ тематик и тенденций. Применение методов текстовой аналитики (облака тегов, графы со-ссылок ключевых слов и т. д.) позволило выявить эволюцию фокуса

исследований. В первые годы становления области (около 2010 г.) ведущими темами были вопросы аппаратного обеспечения и инфраструктуры для больших данных – в текстах часто встречались слова из области высокопроизводительных вычислений (“high performance computing”), упоминались крупные IT-корпорации (IBM, Oracle, Cloudera и др.), выпускающие решения для хранения и обработки данных.

Однако примерно с середины 2010-х наблюдается смещение интереса в сторону алгоритмов машинного обучения, аналитических платформ и приложений. Например, в 2016–2021 гг. среди популярных терминов появляются “data science”, «нейронные сети», «предиктивная аналитика», а также названия фреймворков (TensorFlow, PyTorch и т. д.), что отражает проникновение методов искусственного интеллекта в область больших данных. Такой переход четко прослеживается при сравнении облаков тегов ключевых слов для 2011 г. и 2021 гг.: за десятилетие на смену преимущественно инфраструктурной повестке пришли вопросы интеллектуального анализа данных, интеграции больших данных с технологиями AI и т. п. Кроме того, с развитием социальных медиа возник новый пласт исследований по анализу данных социальных сетей (например, тема “sentiment analysis” тесно связана в публикациях с анализом социальных медиа и твиттер-сообщений).

Эти наблюдения демонстрируют, как комплексный подход к аналитике публикаций – сочетание количественных и содержательных метрик – позволяет выявлять тренды в тематике исследований.

Интерактивные аналитические панели доказали свою полезность для быстрого обнаружения знаний: ключевые игроки (страны, организации, авторы), наиболее цитируемые источники, популярные темы и их динамика во времени стали наглядно видны исследователю.

5.2.2 Интеграция данных и публикаций в медикобиологической сфере

Третьим направлением, выбранным для апробации предложенной автором, стала медицина, а именно задача интеграции научных публикаций и

данных доказательной медицины для поддержки клинических решений. Данная область кардинально отличается от IT-сферы по характеру источников информации: помимо научных статей, существенный интерес представляют клинические рекомендации, данные о лекарственных средствах, клинические исследования, фармакологические базы знаний и т. п. Целью проводимой работы было разработка методов искусственного интеллекта для систем поддержки принятия врачебных решений (СППВР) на основе большого массива доказательных данных о применении лекарств при коморбидных (сочетанных) заболеваниях. Исследование проводилось в рамках гранта Российского научного фонда № 23-75-30012 «Снижение рисков полифармакотерапии с использованием искусственного интеллекта и анализа Больших данных о лекарственных препаратах и их взаимодействиях», в котором автор являлся ответственным исполнителем [132]. На основе предложенной автором аналитической модели цифрового объекта реализована интеграция разнородных медицинских ресурсов, включая:

1) национальные клинические рекомендации Минздрава РФ;
2) государственный реестр лекарственных средств (ГРЛС);
3) международные базы данных по лекарственным взаимодействиям (Drugs.com, Medscape);
4) библиографические базы научных публикаций (PubMed, PubMed Central);
5) базы данных о грантах и исследованиях (NIH RePORTER).

Мультиагентный подход позволил параллельно обрабатывать несколько источников, гибко реагируя на изменения структуры сайтов и содержания данных [133]. Например, один программный агент осуществлял сбор по каталогу клинических рекомендаций (разделяя их по рубрикам МКБ-10), другой – загружал тексты рекомендаций в формате PDF и извлекал из них структуры разделов и заключения, третий – формировал из полученной информации структурированный JSON файл. Аналогично, для ГРЛС был реализован программный агент, который автоматически переходмл по

каждому регистрационному номеру лекарственного препарата, собирал все связанные данные (атрибуты лекарства, список производителей, тексты инструкций) и сохранял их в соответсвующей структуре данных. На рисунке (Рисунок 5.30) представлена схема работы такого агента: он циклически перебирает ссылки на препараты, проверяет наличие страниц с информацией о взаимодействиях, собирает данные по каждому разделу (взаимодействия с другими лекарственными препаратами (ЛП), противопоказания, побочные эффекты и т.д.) и в итоге формирует сводный JSON по данному препарату.

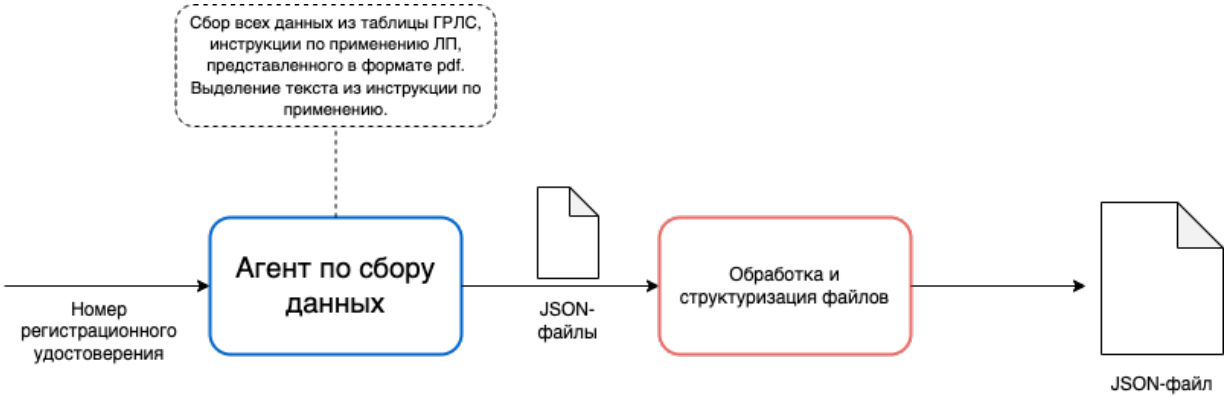


Рисунок 5.30 Схема сбора информации о ЛП из ИС ГРЛС

Полученный в результате интеграции набор медицинских данных является многокомпонентным. Во-первых, удалось собрать полный массив национальных клинических рекомендаций – 415 документов в структурированном виде (общим объемом ~126,4 МБ). Каждая рекомендация была автоматически разобрана на разделы, таблицы и списки рекомендаций, что облегчает последующую идентификацию упоминаний конкретных лекарственных препаратов и методов терапии. Во-вторых, из ГРЛС извлечены сведения обо всех зарегистрированных лекарственных средствах (на апрель 2023 г.) – несколько тысяч наименований. По каждому ЛП хранится его регистрационный номер, состав, производители, а также тексты инструкций по применению (если они доступны). В-третьих, из международных ресурсов по взаимодействию лекарств получены ценные данные о нежелательных сочетаниях препаратов. Так, с сайта Drugs.com было выгружено 5 433 JSON файлов (объемом ~4,11 ГБ) – по числу уникальных действующих веществ, для

которых указаны взаимодействия. Формат одного такого файла представляет собой словарь с ключами: name (название вещества), source (ссылка на страницу источника) и interactions (список всех найденных взаимодействий с указанием типа взаимодействия, степени риска, описания и пр.). Аналогично, по базе Medscape собрано 2 583 JSON файлов (152,4 МБ) со структурированной информацией о взаимодействиях каждого препарата (Приложение В).

Для анализа фронтира научных исследований были использованы данные PubMed и NIH RePORTER. Из PubMed по заданному тематическому запросу (сердечнососудистые заболевания) были извлечены все доступные библиографические записи, при этом из-за ограничения API (не более 10 000 записей за раз) запрос был автоматически разнесен по годам, и полученные частичные результаты объединены (сформировано 40 JSON файлов с суммарной информацией). Данные RePORTER (более 95 000 проектов) позволили дополнить картину сведениями о текущих исследованиях и грантах в данной области, включая бюджеты, географию проектов и руководителей исследований. В базе данных проекты занимают два индекса с количеством записей 390 054 (ключевые проекты) и 1 117 507 (отдельные проекты). Объем собранных данных составил 8,5 Гб.

Важно подчеркнуть, что все агрегированные медицинские данные преобразованы к единому форматно-структурному представлению (JSON), интегрированному в хранилище наряду с публикациями из других доменов. Это значительно упрощает их последующий анализ и сравнение. Так, для научных статей из PubMed применяются те же алгоритмы извлечения метрик (частот ключевых слов, географии авторов, сетей соавторства и т. д.), что и для статей по большим данным и финансовой безопасности. В то же время, наличие специальных полей (например, для препаратов: списки взаимодействующих веществ) открывает возможности для предметноориентированного анализа.

В рамках данного раздела работы были проведены пилотные эксперименты, демонстрирующие пользу комплексного подхода. Например, совмещение сведений о лекарственных взаимодействиях с данными PubMed позволило выделить темы исследований, наиболее актуальные с точки зрения безопасности полифармакотерапии. С помощью облаков тегов по публикациям 2018–2022 гг. выявлено, что часто вместе упоминаются, например, “anticoagulants” (антикоагулянты) и “antibiotic therapy” (антибактериальная терапия) – это сигнализирует о возросшем интересе к проблеме сочетания антикоагулянтов с антибиотиками у сложных пациентов. Такие выводы были бы труднодостижимы при раздельном анализе литературных данных и фармацевтических баз, однако в интегрированной системе их обнаружение заметно упрощается.

Важным результатом является демонстрация практической ценности собранных данных для разработки СППВР. На основе базы знаний, наполненной описанными данными, был создан прототип рекомендательной системы, предлагающей врачу информацию о возможных взаимодействиях назначаемых лекарств и о наличии клинических рекомендаций по интересующему случаю. Этот прототип подтвердил, что межотраслевой синтез (сочетание научной литературы и официальных руководств) дает более полную поддержку решений: например, система не только предупреждает о риске одновременного назначения двух лекарств, но и сразу ссылается на соответствующую клиническую рекомендацию Минздрава и на свежие исследования из PubMed по данному сочетанию.

Результаты проведенных исследований свидетельствуют о том, что предложенный автором подход успешно масштабируется на медико-биологическую область, обеспечивая качественно новый уровень интеграции научных знаний для практического здравоохранения.

5.3 Построение базы данных свойств облученных реакторных материалов

При сборе и анализе информации из различных информационных ресурсов возникают задачи оценки достоверности и надежности распространяемой информации. Особенно чувствительны эти вопросы при рассмотрении и анализе научно-технической информации, связанной с критическими технологиями развития государства, к которым относятся ядерные технологии, в частности, вопросы, связанные с поведением материалов в различных условиях. Неверная оценка материалов по данному направлению может привести к негативным последствиям не только в масштабах одной страны, но и всей планеты [134].

Ситуация усугубляется наличием достаточно большого количества опубликованных изданий и отчетов, не прошедших экспертизу. Последнее может приводить к достаточно серьезным проблемам при использовании научно-технической информации. Поэтому кроме решения задач сбора, классификации, хранения информации необходимо критически оценивать найденные исходные данные.

Критическая оценка надежности данных может проводиться поэтапно, начиная с компьютерной оценки надежности полученной информации с учетом библиометрических данных авторов информации и средств научно-технической информации, и, заканчивая экспертной оценкой специалистов.

Основываясь на данных положениях, можно сделать вывод о том, что при необходимости по каждому материалу научно-технической информации необходимо проводить двухфакторную проверку – проверка входных данных материала и экспертная оценка. Проверка данных материала может быть автоматизирована, базируясь на доступных выходных характеристиках, таких как рейтинг издания, рейтинг автора, аффилиация автора и т.д.

Экспертная оценка же должна представлять собой всестороннее рассмотрение данных учеными, имеющими достаточный опыт в конкретной области знаний. При рассмотрении данных, направленных на пополнение

существующих, или создание новых баз данных, эксперты должны учитывать множество факторов. Например, актуальность рецензируемых исследований, научную новизну, качество исследований и представляемого материала и т.д.

Компьютерная оценка является универсальным, относительно независимым источником анализа достоверности данных. Для решения задач оценки материалов в первую очередь необходимо выявить доверенные источники информации.

Рассмотрим решение задачи надежности материалов научно-технической информации при формировании профиля объекта типа сталь. Апробация методики проходила в рамках договора с ВНИИА им. Н.Л. Духова № 1707 от 29 августа 2022 г. по выполнению НИР «Разработка программы выборки данных по свойствам и структурам облученных реакторных материалов из мировых источников информации».

В данном случае предметом анализа стали публикации в международных реферативных базах данных – Web of Science и Scopus (в базы данных входит более 24 тысяч изданий). Выбор перечисленных баз данных обусловлен тем, что изданию необходимо удовлетворять ряду характеристик для того, чтобы войти в озвученные базы данных. В том числе это относится к рецензированию поступающих статей специалистами-предметниками в тематической области, международному составу редколлегии и т.д.

Для компьютерной оценки предложены следующие критерии.

Квартиль издания в реферативных базах данных является первым критерием оценки надежности информации. Чем выше квартиль издания, тем больше доверия к рассматриваемому материалу.

Вторым критерием оценки материала является индекс Хирша авторов издания. Индекс вычисляется на основе распределения цитирований работ исследователя. Чем выше индекс Хирша автора, тем более он считается признанным и авторитетным в научном мире. Анализ тематической области показал, что можно выделить следующие диапазоны индексов Хирша авторов, публикующих работы по тематике ядерных материалов (Таблица 5.1).

Таблица 5.1 Диапазоны индекса Хирша

№	Значение индекса Хирша автора	Степень надежности в промежутке [0;1]
1.	0 – 5	Низкая надежность
2.	6 – 20	Средняя надежность
3.	21+	Высокая надежность

Третьим критерием оценки надежности материала является страна, к которой относятся авторы. Можно выделить признанных лидеров в области исследования ядерных материалов – США, Российскую Федерацию, Францию, Китай, Японию. Страны можно разделить также на три группы по убыванию надежности (аналогично Таблице 5.1).

Четвертым критерием оценки надежности материала является аффилиация автора. По данному критерию выделяются ключевые центры, ведущие научно-исследовательские работы по ядерным материалам. Принцип оценки аффилиации аналогичен третьему критерию (деление на три группы).

Выбранные критерии являются наиболее объективными для оценки информационного материала. Возможно использование и других критериев, таких как количество цитирований материала, количество просмотров и т.д. Однако эти критерии сильно зависят от года публикации статьи, так что чем раньше статья была опубликована, тем больше у неё будет цитирований и просмотров. Расчет вычисляемой характеристики доверия к статье – f_{trust} проходит по формуле:

$$f_{trust} = w_1 Q + w_2 h_a + w_3 C + w_4 Aff \quad (5.1)$$

где Q — обратное значение квартилю издания,

h_a— индекс Хирша ведущего автора,

C— наибольший из коэффициентов доверия к странам по уровню исследований,

Aff— наибольший из коэффициентов доверия к аффилиациям по уровню исследований,

w_i — нормированные весовые коэффициенты, $i = 1, \dots, 4$, $\sum_{i=1}^{4} w_i = 1$.

В подразделе 5.2 рассматриваются аналитические задачи, базирующиеся на анализе выходных данных отдельных публикаций, которые являются базовым объектом.

В решении заданной аналитической задачи возникает более сложная задача, связанная с формированием и наполнением профиля цифрового объекта из полного текста материала.

Рассмотрим пример решения задачи составления базы данных характеристик облученных ядерных материалов на основе модели комплексного информационного объекта. В рамках задачи необходимо:

1. Разработать формальную модель комплексного информационного объекта для стали.
2. Сформировать критерии доверия к источникам (квартиль журнала, индекс Хирша авторов, страна происхождения, аффилиации авторов коллектива) и встроить их в алгоритм ранжирования публикаций.
3. Выполнить статистический и содержательный анализ выборки (география, журналы, ключевые слова, материалы) как основу для построения дальнейших прогностических моделей.
4. Сформировать итоговый документ для аналитиков.

Предметом автоматизированного мониторинга выступают публикации, содержащие результаты экспериментальных исследований по свойствам четырех больших групп конструкционных реакторных материалов:

1. Бейнитные стали (низколегированные корпусные стали российских и зарубежных водо-водяных реакторов, а также модельные сплавы близкого к ним состава).
2. Аустенитные стали (коррозионно-стойкие хромоникелевые конструкционные стали реакторов на быстрых нейтронах, а также стали этого класса, использующиеся во внутри корпусных устройствах реакторов).

3. Ферритно-мартенситные стали (хромистые стали мартенситного, ферритно-мартенситного и ферритного классов, использующиеся в реакторах на быстрых нейтронах, а также планирующиеся к использованию в реакторах будущего и термоядерных реакторах).

4. Никелевые сплавы (никель и сплавы на его основе, планируемые к использованию в реакторах будущего и использующиеся во внутри корпусных устройствах реакторов).

По итогам анализа их тематической принадлежности были отобраны 5 изданий, специализирующихся на тематике ядерных материалов. Перечисленные издания относятся к направлениям – Materials Science (miscellaneous), Nuclear and High Energy Physics, Nuclear Energy and Engineering, которые являются релевантными к заданной области исследований.

Отобранные издания:

1. Journal of Nuclear Materials (Q1).
2. Nuclear Materials and Energy (Q1).
3. Nuclear Instruments and methods Part B (Q2, Q3).
4. Nuclear Engineering and Design (Q1).
5. Fusion Engineering and Design (Q2).

Целевыми определены 5 классов свойств конструкционных реакторных материалов:

1) Hardening – влияние облучения на механические свойства сталей и никелевых сплавов, а также некоторые данные об охрупчивании данных материалов (пределы текучести и прочности, удлинение, сужение, температура хрупко-вязкого перехода и их изменения в зависимости от различных факторов (флюенса, температуры облучения и испытания, повреждающей дозы, содержания примесей и легирующих элементов и т.п.).

2) Phase stability – радиационное распухание (табличные данные и графики о вакансионном или газовом распухании, а также параметрах пористости (размеры и плотность пор и пузырьков и т.п.).

3) Radiation defects – формирование радиационных дефектов при облучении (табличные данные и графики о плотности, размерах дислокаций, дислокационных петель, вакансионных скоплений и т.п.).

4) Segregarity – радиационно-стимулированные изменения состава и структуры (данные атомно-зондового анализа, энерго-дисперсионного анализа и т.д.).

5) Swelling – формирование радиационно-стимулированных преципитатов (образование, рост и поведение дисперсных фаз), а также данные об их типе и составе.

Для решения поставленных задач автором разработана комплексная модель цифрового информационного объекта (см. раздел 2.3):

$$Steel = \langle S_{base}, S_{hardening}, S_{phasestab}, S_{rad\_def}, S_{segreg}, S_{swelling} \rangle \qquad (5.2)$$

В которой статические характеристики разбиты на 6 групп со следующей размерностью:

$$|S_{base}| = 27, |S_{hardening}| = 45, |S_{phasestab}| = 28, \qquad (5.3)$$
$$|S_{rad\_def}| = 42, |S_{segreg}| = 30\ |S_{swelling}| = 43$$

Множество S_{base} состоит из базовых характеристик стали, таких как название и легирующий состав. В решении поставленных аналитических задач характеристики множества динамических D и вычисляемых F характеристик пусты, а в Rel содержатся ссылки на соответствующие научные публикации.

Наполнение такого цифрового профиля происходит на основе данных публикаций с использованием процедур конвертации. В данном случае модель Article, предложенная в подразделе 3.1, насыщается данными из полнотекстовых материалов.

Расширенная модель публикации выглядит следующим образом:

$$Article_{full} = \langle S_{bibl}, S_{table}, S_{img} \rangle. \qquad (5.4)$$

Для осуществления преобразования данных из формата публикации в формат стали необходимо с использованием методов обработки изображений

и таблиц (подраздел 3.3.4) заполнить соответствующие характеристики согласно формуле (5.2).

Предложенный подход позволяет решать не только заданную аналитическую задачу, но и может быть масштабирован для других задач.

Всего было проанализировано более **40 000** статей, из которых в итоговую выборку вошли **534** публикации, удовлетворившие предметным и качественным фильтрам.

В результате анализа публикаций на предмет фотографий микроструктур, диаграмм получено и передано заказчику ~1650 изображений. В результате рассмотрения и анализа таблиц и графиков из научных статей получено и передано ~8700 точек, описывающих требуемые заказчиком свойства [135]. Реализована база данных с удобным интерфейсом поиска по материалам и параметрам, позволяющая специалистам быстро находить все доступные результаты по конкретному материалу или режиму облучения (Рисунок 5.31).

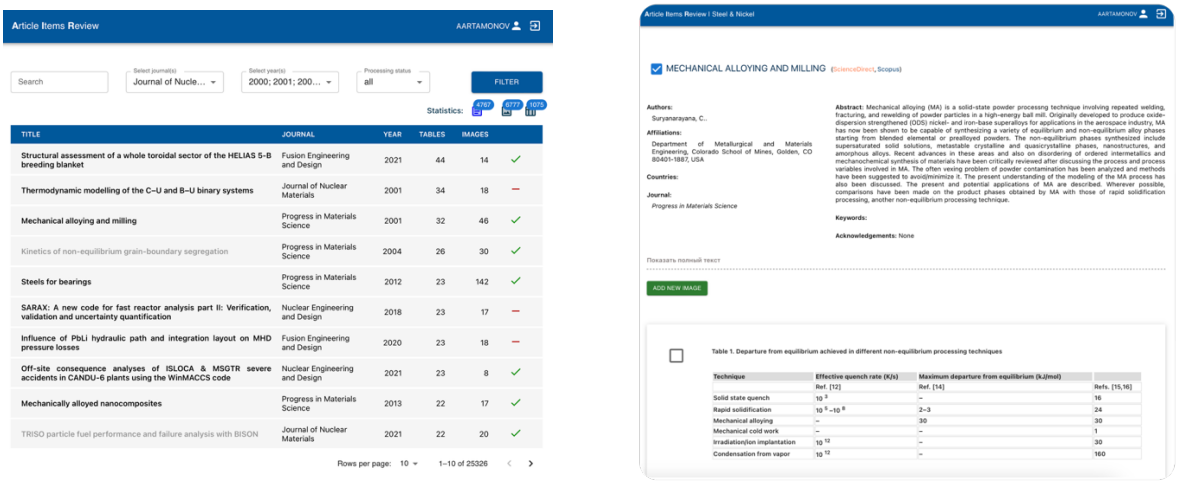


Рисунок 5.31 Примеры интерфейса БД по облученным реакторным материалам

Темп публикации релевантных работ демонстрирует отчётливый возрастающий тренд: с единичных статей в начале 2000-х до локального максимума 54 публикации в 2021 г. (Рисунок 5.32), что коррелирует с запуском крупных международных проектов и модернизацией исследовательских

реакторов. Однако следует учитывать, что рассматривались только статьи, содержащие графическую или табличную информацию.

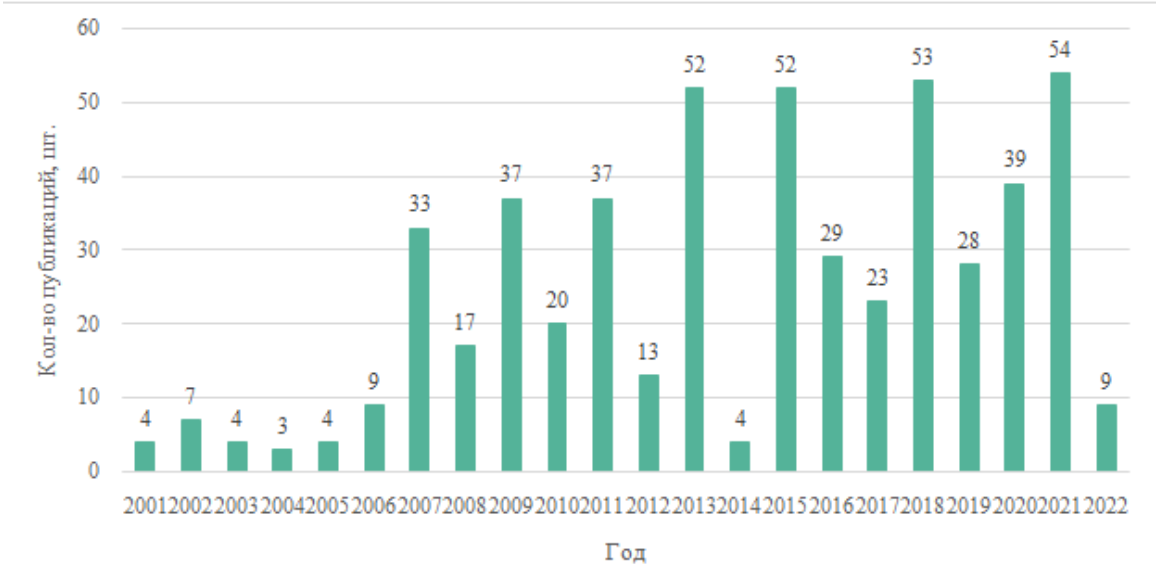


Рисунок 5.32 Распределение публикаций по годам

Публикации, вошедшие в выборку, были также рассмотрены на предмет распределения по странам (Рисунок 5.33).

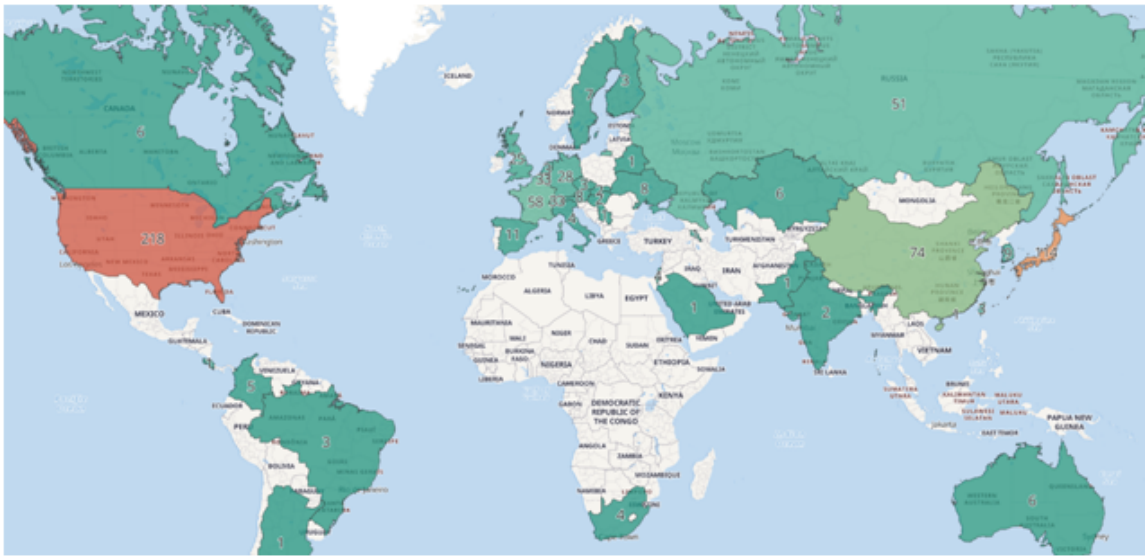


Рисунок 5.33 – Распределение публикаций по рассматриваемой тематике по странам

На рисунке представлено распределение в соответствии с указанными в публикациях странами в аффилиации авторов. В случае если у публикации было несколько авторов с различающимися странами аффилиации, то учитывались все указанные страны. Наибольшее количество публикаций

приходится на США (218 шт.). Также в пятерку лидеров вошли такие страны, как Япония (114 шт.), Китай (74 шт.), Франция (58 шт.) и Россия (51 шт.).

Подавляющее число статей (423 шт.) было опубликовано в журнале Journal of Nuclear Materials – ежемесячном рецензируемым научным журналом по исследованиям материалов для физики ускорителей, ядерной энергетики и приложений топливного цикла издательства Elsevier.

ВЫВОДЫ ПО ГЛАВЕ 5

1. Подтверждено, что методы интеллектуального анализа и построения интерактивных аналитических панелей, основанные на комплексной модели информационного цифрового объекта, применимы для систематизации и анализа больших массивов научных данных. Интеграция разнородных научно-технических объектов (публикации, патенты, данные проектов, отраслевые документы) в единую модель и позволяют сформировать аналитическую систему.

2. Предложена методика построения интерактивных аналитических панелей по тематическому направлению, на примере междисциплинарного направления «Финансовая безопасность», позволяющая проводить оперативный анализ больших данных в режиме реального времени для решения различных информационно-аналитических задач.

3. Разработана комплексная модель информационного цифрового объекта облученного реакторного материала, включающая физико-механические свойства, экспериментальные и медиа-данные (изображения микроструктур, таблицы точек) и методы ее заполнения с учетом оценки доверия исходного материала, учитывающего квартиль журнала, индекс Хирша автора, страновой и аффилиационный коэффициенты.

4. Подтверждена универсальность предложенных автором моделей и методов не только при решении социально значимых задач, но и в научно-технических задачах выявления закономерностей в разнородных данных, а именно, определение географии и динамики научных исследований, ключевых объектов и тематических трендов, что необходимо для управления научными

исследованиями, технологическим развитием и поддержки принятия решений по научно-техническим направлениям.

ГЛАВА 6 ПРОГРАММНЫЕ ИНСТРУМЕНТЫ РЕШЕНИЯ НАУЧНО-ТЕХНИЧЕСКИХ ЗАДАЧ

6.1 Программный инструмент выявления явных и неявных связей между цифровыми объектами

6.1.1 Метод выявления связей между объектами

В настоящее время актуальными направлением исследований являются методы графового анализа и визуальной аналитики, позволяющие выявлять скрытые закономерности в сложных, разнородных наборах информации. Обзор последних исследований показывает, что визуальная аналитика выступает ключевым направлением в обработке и интерпретации больших данных, находя применение в медицине, социальных и естественных науках, инженерии и компьютерных технологиях [136, 137]. Столкнувшись с огромными объёмами разнородных данных, исследователи отмечают необходимость трансформации «нечитаемых, сложных данных» в наглядные графики, схемы или диаграммы, понятные человеку. Именно поэтому в последние годы интенсивно разрабатываются разнообразные системы и инструменты визуализации данных, призванные облегчить аналитикам понимание сложных структур.

Визуальная аналитика играет роль связующего звена между методами интеллектуального анализа данных и когнитивными способностями человека, позволяя интерактивно исследовать данные и находить скрытые взаимосвязи [138].

Графовая визуализация широко применяется для отслеживания взаимосвязей между информационными объектами; в частности, с её помощью исследуются сложные сети взаимодействий (например, научные публикации и патенты), что позволяет изучать эволюцию знаний и технологий и поиска скрытых взаимосвязей в больших массивах данных [139, 140]. Например, в обзоре [141] отмечается, что графы демонстрируют мощные возможности при анализе научных данных, позволяя эффективно моделировать гетерогенные и сложные связи между сущностями. Это

подтверждается ростом использования технологий графовых баз данных и графовой аналитики.

В работе [142] предложен интерактивный инструмент визуализации для исследования неявных связей в реляционных наборах данных. Данный инструмент позволяет пользователю гибко изучать скрытые отношения между объектами в связанных таблицах, комбинируя графовое представление с традиционными подходами к анализу данных. В работе [143] представлен обзор современных тенденций в области визуализации сетей и графов, где подчёркивается важность адаптации методов графовой визуализации под конкретные аналитические задачи и типы связей.

Отечественные исследования отражают возрастающий интерес к графовой аналитике, например в работе [144] рассматривают применение графовых методов для решения задач цифровой экономики, демонстрируя эффективность графовых баз данных при анализе больших данных и поиске взаимосвязей в разнородной информации. В работе [145] исследуется применение графовых моделей в анализе социальных сетей, показывая, как графовый подход помогает выявлять структуру сообществ и ключевые связи между участниками.

Развитие методов графового машинного обучения (Graph ML) позволяет автоматически обнаруживать скрытые паттерны в сетевых данных, например, авторы [146] предложили модель DynaHGraph для обучения скрытым отношениям в динамических графах – эта модель сочетает семантическую и структурную информацию для выявления эволюционирующих зависимостей во времени.

Таким образом, современные научные публикации подтверждают актуальность создания гибких инструментов графового анализа, способных интегрировать данные различной природы и раскрывать как явные, так и неочевидные (скрытые) взаимосвязи между объектами [147, 148].

В статье [149] предлагается подход к определению тематической близости научных журналов и конференций посредством анализа графа

соавторства публикаций. Если один и тот же автор публикуется в разных журналах, между соответствующими вершинами (журналами) проводится ребро, причём ему присваивается вес, рассчитываемый по определённой формуле. Предполагается, что общие авторы указывают на тематическую близость изданий. Метод графа соавторства позволяет выявлять сообщества и связи в массивах библиографических данных, однако применим лишь для специфического вида объектов (публикации) и отношений (авторство). Его масштабируемость ограничена: при расширении на другие типы сущностей возникают сложности интерпретации ребёр и необходимость иных метрик близости.

Классический статистический метод выявления связей – вычисление коэффициента корреляции между показателями. В общем случае корреляционный метод оценивает степень взаимосвязи (линейной зависимости) между двумя переменными. Если коэффициент корреляции существенно отличен от нуля, можно говорить о наличии связи между параметрами. Однако корреляционный анализ предполагает, что данные представлены в упорядоченных наборах чисел. Он не учитывает структуру связей между разнородными объектами и не применим непосредственно к нечисловым данным. Более того, выявленная корреляция не всегда означает наличие прямой причинно-следственной зависимости.

В ряде работ [150] предлагаются специальные методы выявления причинности («выявление причинности» и «оценка причинности»), позволяющие установить, как изменение одного фактора влияет на другой. Эти методы концентрируются на оценке воздействия одного элемента на другой и формальном тестировании наличия причинно-следственной связи. Однако в контексте системного анализа данных применение таких методов ограничено: они ориентированы скорее на пары переменных и не учитывают сложную структуру объекта и множества взаимодействий, а также требуют специальных (возможно, экспериментальных) данных для подтверждения причинности.

Для моделирования сложных предметных областей широко используются семантические модели, включая онтологии, тезаурусы и т.п. [151, 152]. Онтологии позволяют описать структуру области: ввести классы сущностей, их свойства и связи, после чего представить базу знаний в виде графа (семантической сети). Семантический подход эффективен, когда априорно известна структура данных и отношения между сущностями, и он применяется для явного кодирования знаний (например, в системах искусственного интеллекта) [147]. Однако при анализе открытых неструктурированных данных онтологический подход затруднён: требуется либо заранее создать онтологию предметной области, либо извлекать факты с помощью методов обработки текста. Кроме того, онтологии обычно фиксируют только заведомо известные (явные) связи.

Традиционные подходы не позволяют в полной мере работать с гетерогенными данными – ситуацией, когда объекты различной природы и структуры могут быть связаны разнообразными отношениями. В реальных информационных системах данные разнородны: например, профиль организации включает текстовые описания, списки сотрудников, показатели активности; научная публикация содержит метаданные, полный текст, ссылки, списки авторов с аффилиациями и т.д. Возникает потребность в универсальном подходе, способном независимо от структуры и типа входных данных строить единую модель связей и выделять как прямые, так и скрытые пути взаимодействия объектов.

Обозначенную проблему предлагается решить посредством графового представления совокупности объектов и их характеристик. Графы естественным образом моделируют сети связей: вершины могут обозначать сущности или значения характеристик, а рёбра – наличие определённого отношения между ними. Визуальное представление данных в виде графа предоставляет аналитику наглядную «карту» взаимосвязей, позволяющую интерактивно исследовать структуру данных.

Пусть Q – множество объектов $AObj$, то есть

$$Q = \{AObj = (a_1, a_2, \dots, a_M) | a_i - \text{характеристика объекта } AObj, \quad (6.1)$$
$$i = 1,2, \dots, M\},$$

где $a_i \in S \cup D \cup F \cup Rel$.

Каждому объекту $AObj = (a_1, a_2, \dots, a_M) \in Q$ поставим в соответствии граф $G =< V, E >$, где множество вершин

$$V = \{v_1, v_2, \dots, v_T\}, T \geq M, \quad (6.2)$$

составляют значения характеристик $a_1, a_2, \dots, a_M$ объекта $AObj$, и множество ребер

$$E = \{(v_i, v_j) \mid \exists\, AObj \in Q : AObj(a_i) = v_i \text{ и } AObj(a_j) = v_j\}, \quad (6.3)$$

то есть две вершины $v_i, v_j \in V$ соединены ребром, если найдется объект $AObj = (a_1, a_2, \dots, a_M) \in Q$, такой что значение характеристики a_i объекта $AObj$ равно v_i и значение характеристики a_j объекта $AObj$ равно v_j, где характеристики $a_i, a_j \in S \cup D \cup F \cup Rel$. Иными словами, вершины связываются ребром, если соответствующие характеристики объектов совпадают по значениям.

Правило построения графа определяется аналитиком, то есть заранее указывается, какие именно характеристики и каким образом следует связывать между собой для выявления интересующих связей. Например: при анализе научных публикаций можно определить типы вершин «Автор», «Статья», «Журнал» и связать вершины типа «Автор» с вершинами типа «Статья», а вершины «Статья» – с вершинами «Журнал». Тогда граф отразит отношения авторства и публикации в журнале.

Каждой вершине из множества V поставим в соответствие ее тип – характеристику из множества $\{a_1, a_2, \dots, a_M\}$. Тогда множество вершин

$$V = K_1 \cup K_2 \cup \dots \cup K_M, \quad (6.4)$$

где $K_i = \{AObj(a_i) | AObj \in Q\}$ – множество вершин типа a_i, $i = 1,2, \dots, M$.

Таким образом, граф $G =< V, E >$ можно представить в виде гетерогенного графа

$$G =< K_1 \cup \dots \cup K_L, E >, \quad (6.5)$$

где $L \in \{1,2,\dots,M\}$, при этом если $v_i, v_j \in K_W$, $W \in \{1,\dots,L\}$, то $(v_i, v_j) \notin E$ (характеристики и их количество L задает аналитик).

Пусть K_L и K_P – множества вершин типа a_L и a_P соответственно, $L \neq P$.

Приведем следующие определения:

Явная связь между вершинами $v_i \in K_L$ и $v_j \in K_P$ существует, если есть ребро $(v_i, v_j) \in E$. Явные связи образуются между вершинами разных типов (поскольку ребро проводится при совпадении значений двух разных характеристик у некоторого объекта). Например, для графа публикаций между конкретным автором и конкретной статьёй имеется явная связь (автор написал статью); между статьёй и журналом – также явная связь (статья опубликована в журнале). Явные связи отражают непосредственные отношения между объектами (или их компонентами), представленные в данных.

Неявная связь между вершинами $v_i \in K_L$ и $v_j \in K_L$ существует, если есть вершина $v_x \in K_P$, такая что имеется ребро $(v_i, v_x) \in E$ и ребро $(v_j, v_x) \in E$. Например, два автора могут быть связаны неявно, если они оба являются соавторами одной и той же статьи: хотя между вершинами-авторами нет ребра, существует вершина-статья, соединённая явными связями с каждым из них, что свидетельствует о косвенной связи между авторами (они работали над одним проектом).

Неявная связь порядка n между вершинами $v_i \in K_L$ и $v_j \in K_L$ существует, если есть вершина $v_x \in K_P$, такая что существует явные связи $(v_i, v_x) \in E$ и $(v_j, v_x) \in E$, порядок хотя бы одной из которой равен n. В частности, неявная связь 1-го порядка соответствует определению выше (через одну промежуточную вершину).

Например, если один научный журнал публикует тематический спецвыпуск, объединяющий две конференции, то эти конференции оказываются неявно связаны через данный выпуск; но фактически признаком связи выступает один и тот же объект (выпуск), содержащий обе конференции, поэтому такую связь можно считать «очевидной». В рассмотренной модели

очевидные неявные связи проявляются, когда одна вершина (значение характеристики) включает в своё значение два других значения – тогда для вершин одного типа формально получается неявная связь 1-го порядка, но она тривиальна.

Следует отметить, что данные определения справедливы для неориентированных графов. Если граф ориентированный (ребра имеют направление), то типизация вершин для выявления неявных связей не требуется, поскольку направление дуги уже задаёт различие сущностей в паре. В работе рассматриваются неориентированные графы (симметричные отношения), поэтому разделение на типы и введение понятий явных/неявных связей оправдано.

Предложенный автором подход выявления неявных связей между объектами заключается в визуализации связей между сущностями внутри объектов, при которой аналитик самостоятельно определяет критерии связывания сущностей, а на основе полученной структуры связей делает выводы об их взаимоотношениях. Метод позволяет пользователю сконструировать графовую модель на основе неструктурированных данных, указав, какие поля объектов и как именно должны быть взаимосвязаны. Это обеспечивает универсальность подхода независимо от структуры исходных данных.

В разработанном авторе методе выделены четыре возможных случая взаимосвязи полей в рамках иерархически организованных данных (в каждом случае реализуется свой алгоритм извлечения и связывания значений полей):

1) Связывание объектов, не имеющих общих предков;

2) Связывание объектов, находящихся на одном уровне имеющих общего предка;

3) Связывание объектов, находящиеся на разном уровне, но имея общего предка;

4) Связывание объектов одного поля.

Рассмотрим файл со следующей структурой (Рисунок 6.1). Для получения объектов из такой структуры данных и последующего связывания интересующих полей будет достаточно обратиться напрямую к полю. Если интересующее поле находится на n-ом уровне вложенности, то данное поле получает предка, который описывает путь получения данных по полю. Предком считается предшествующее поле в данных. Если структура данных представляет сильную вложенность, то вводится понятие ближайшего предка. Ближайший предок – поле в данных, при обращении к которому выводится интересующее поле. Для одного поля может существовать только один ближайший предок, таким образом каждое поле будет иметь путь, состоящий из предков и одного ближайшего предка. Ближайший предок может относится к нескольким полям, например поле affiliation является ближайшим предком для “name” и “country”, по которым можно получить название аффилиации и страну.


```
{
  "Title": "article 1",
  "author": {
    "full_name": "Author's name",
    "affiliation": {
      "name": "National Resaerch Nuclear University MEPhI",
      "country": "Russian Federation"
    },
    "published_year": 2021,
    "keyword": "React"
  }
}
```


Рисунок 6.1 Пример файла с обычной вложенностью

Общим предком считается поле, по которому необходимо обратиться для получения искомых данных. Стоит отметить, что ближайший предок может быть и общим, при условии, что уровень вложенности равен одному. В противном случае общим предком будет считаться поле находящиеся на верхнем уровне вложенности, а ближайший предок на предпоследнем уровне вложенности.

Рассмотрим файл с вложенной структурой. На рисунке (Рисунок 6.2) представлен пример данных, в котором по одному полю можно получить

список данных с одинаковыми полями. В этом случае обратится напрямую к полю и связать их не получится, так как при обращению к предку выводится список значений. Для такого случая необходимо получить данные из списка и работать с элементами списка по отдельности используя один и тот же алгоритм связи.


```
{
  "Title": "article_1",
  "author": [
    {
      "full_name": "Author's name",
      "affiliation": [
        {
          "name": "National Research Nuclear University MEPhI",
          "country": "Russian Federation"
        }
      ]
    },
    {
      "full_name": "Another Authors's name",
      "affiliation": [
        {
          "name": "National Research Nuclear University MEPhI",
          "country": "Russian Federation"
        },
        {
          "name": "Moscow State University",
          "country": "Russian Federation"
        }
      ]
    }
  ],
  "published_year": 2021,
  "keyword": [
    "React",
    "Python"
  ]
}
```


Рисунок 6.2 Пример файла со вложенностью данных

Для таких структур рассмотрены следующие примеры связи полей:

Связывание объектов, не имеющих общих предков (1-ый вид связи). В этом случае выбранные поля находятся в разных ветвях структуры данных и не принадлежат одному вложенному объекту. Например, если в структуре данных поле *Title* (название документа) находится отдельно, а поле *author.full_name* вложено в структуру автора, то у них нет общего родительского объекта кроме самого документа верхнего уровня. Для связи таких полей достаточно напрямую сопоставлять их значения. При структуре

данных, изображённой на рисунке (Рисунок 6.1), для связывания доступны следующие пары полей:

- *Title-author.full_name,*
- *Title-author.affiliation.name,*
- *Title-author.affiliation.country,*
- *Title-published_year,*
- и другие варианты, при которых узлы без общих предков будут связываться.

Если значение одного из выбранных полей представлено списком, то для связывания такого рода необходимо составить декартово произведение множеств их значений. Например, при связи списка авторов со списком ключевых слов сначала перебираются все авторы каждого документа (значения поля $author.full\_name$), затем для каждого автора перебираются все ключевые слова (значения поля $keyword$).

Связывание объектов, находящихся на одном уровне, имеющих общего предка (2-ый вид связи). Данный случай позволяет связать поля, находящиеся на одном уровне вложенности, в пределах одного родительского объекта. Например, можно связать название аффилиации и страну аффилиации, если они являются атрибутами одного и того же вложенного объекта *affiliation*. Для связывания полей, расположенных на одном уровне, необходимо получить данные общего ближайшего предка и соединить их попарно.

Такая связь обеспечивает связь характеристик одного объекта, что позволяет сделать поэлементное соединение (каждый элемент одного списка сопоставляется соответствующему элементу другого списка). Например, имея структуру данных, представленную на рисунке (Рисунок 6.2) (где каждому автору сопоставлен вложенный объект *affiliation* с полями *name* и *country*), можно связать поля author.affiliation.name и *author.affiliation.country* друг с другом (название организации с соответствующей страной той же организации).

Связывание объектов, находящихся на разных уровнях, но имеющих общего предка (3-ый вид связи). Этот случай представляет собой комбинацию первого и второго видов. Чтобы связать два поля, расположенных на разных уровнях вложенности, но внутри одного объекта ($CAObj$), необходимо сначала привести данные к одному уровню (в рамках общего предка) как во втором случае, а затем получить информацию по второму полю и найти декартово произведение, как для первого вида.

Предположим, требуется связать поле верхнего уровня с вложенным полем, имеющим общего предшественника. Например, пользователь хочет связать *author.full_name* (имя автора, поле верхнего уровня авторов статьи) и *author.affiliation.country* (страну аффилиации автора, вложенное поле). Для этого необходимо преобразовать структуру данных: получить для каждого автора его страны аффилиаций (то есть поднять *affiliation.country* на уровень автора), после чего связать каждое имя автора со всеми странами его аффилиаций. При структуре данных, аналогичной рисунку (Рисунок 6.2), возможно связать, например, пары: *author.full_name – author.affiliation.country* и *author.full_name – author.affiliation.name* (если у автора несколько аффилиаций, его имя свяжется с каждой страной и названием организации из списка аффилиаций).

Связывание объектов одного поля (4-ый вид связи). Этот случай представляет собой частный случай первого вида связи, когда в качестве двух выбранных полей выступает одно и то же поле. Такой вид связи используется при связывании элементов внутри одной записи по одной и той же характеристике. Например, если выбрать поле “*full_name*” авторов и попытаться связать его само с собой в рамках одного объекта (одной статьи), то в результирующем графе все авторы данной статьи окажутся связанными друг с другом. В частности, для статьи *«Title 1»* все указанные авторы будут попарно связаны между собой. Таким образом, получаются узлы со следующими связями (здесь под *«Author 1»*, *«Author 2»* и *«Author 3»* условно подразумеваются разные авторы внутри одной записи):

- *«Author 1 — Author 1»*;
- *«Author 1 — Author 2»*;
- *«Author 2 — Author 3»*.

Подобные связи позволяют отразить в графе, например, факты совместной авторства: все соавторы одной публикации будут соединены между собой через общую публикацию (что, по определению, является неявной связью первого порядка, но здесь она введена явно в граф как связь элементов одного множества).

Для рассмотренных четырех видов связи можно сформировать универсальные шаблоны связывания объектов внутри одного документа (записи). На рисунке (Рисунок 6.3) приведён пример структуры данных с обозначением видов связей: каждый вид пронумерован согласно приведённой классификации (1-ый вид связи, 2-ой вид связи, 3-ий вид связи, 4-ый вид связи) для соответствующих пар полей внутри одной и той же записи.

В зависимости от выбранного вида связи будет применяться соответствующий алгоритм связывания полей. Для каждого вида реализован отдельный модуль программного кода на основе предложенного алгоритма, который позволяет проходить по всей структуре данных и получать необходимые значения по полям, заранее конвертируя их в список узлов для графа. Таким образом, автором разработана алгоритмическая основа, охватывающая все перечисленные шаблоны: от простого связывания полей верхнего уровня до сложных случаев с несколькими уровнями вложенности. Это обеспечивает независимость методики построения графового представления от структуры входных данных.

На основе предложенной методики разработан программный инструмент, состоящий из трех компонентов (Рисунок 6.4): графического интерфейса, программного интерфейса и базы данных. Графический интерфейс позволяет пользователю взаимодействовать с программным интерфейсом посредством использования определенных фильтров. В свою очередь программный интерфейс взаимодействует с базой данных. При

выгрузке данных программный интерфейс преобразует их в формат для построения графа.


```
[
  {
    "Title": "article 1",
    "authors": [
      {
        "full_name": "Authors full name",
        "affiliations": [
          {
            "name": "National Research Nuclear University MEPhI",
            "country": "Russian Federation"
          }
        ]
      },
      {
        "full_name": "Authors full name",
        "affiliations": [
          {
            "name": "National Research Nuclear University MEPhI",
            "country": "Russian Federation"
          },
          {
            "name": "Moscow State Universirty",
            "country": "Russian Federation"
          }
        ]
      }
    ],
    "published_year": 2021,
    "keyword": [
      "React",
      "Python"
    ]
  },
  {
    "Title": "article 2"
```


4 2 3 1

Рисунок 6.3 Пример структуры данных

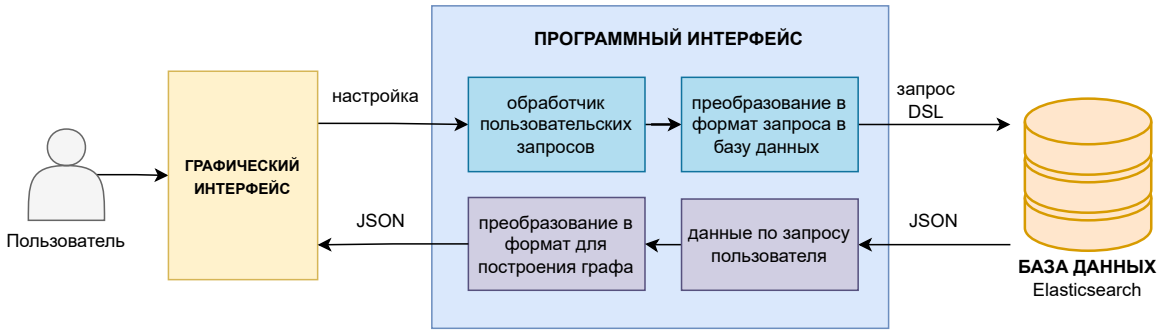


Рисунок 6.4 Архитектура инструмента для построения графа

Графический интерфейс (GUI) представляет собой веб-приложение, предоставляющее пользователю интерактивные средства для настройки параметров графа и визуального отображения результатов [153]. Программный интерфейс (Server API) отвечает за обработку данных: формирует запросы к базе данных, преобразует полученные результаты в формат графа и применяет заданные пользователем фильтры, правила и т.д. Связующим звеном выступает спецификация запроса, формируемая в GUI и передаваемая на сервер для выполнения.

В качестве frontend-библиотеки выбран React; для визуализации графа используется специализированная библиотека react-force-graph, а также Three.js (для 3D-рендеринга узлов и рёбер) и d3-force-3d (для реализации «физики» графа: силы притяжения/отталкивания узлов и проч.) [154]. Эти решения позволяют формировать крупномасштабные графы в браузере и настроить внешний вид узлов/связей. Для верстки и готовых компонентов интерфейса использована библиотека MUI [46], предоставляющая набор готовых React-компонентов и тем самым ускоряющая разработку пользовательского интерфейса [155, 156].

Пользовательский интерфейс предоставляет интерактивную форму для конфигурирования графа (Рисунок 6.5). Пользователь последовательно выбирает источник данных (индекс в базе данных), интересующие поля для связывания и отображения на графе, задаёт условия фильтрации и правила визуального оформления. После ввода всех параметров пользователь запускает процесс построения графа.

При нажатии кнопки построения графа графический интерфейс отправляет запрос в программный интерфейс. Программный интерфейс реализован на языке Python с использованием фреймворка FastAPI [157] для организации RESTful API. Программный модуль принимает входной запрос, извлекает из него указания: какой индекс (набор данных) использовать, какие поля связывать, какие фильтры применить и т.д. Далее программный интерфейс формирует запрос к базе данных (в нашем случае используется

документоориентированная СУБД Elasticsearch) с учётом всех условий, заданных пользователем.

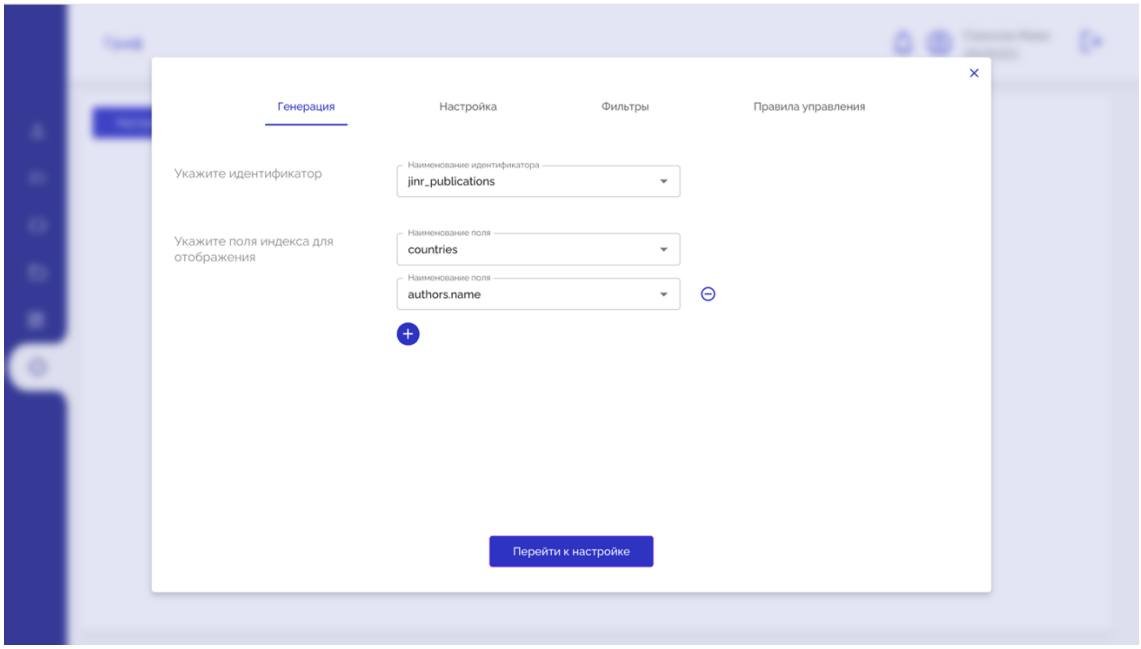


Рисунок 6.5 Графический интерфейс системы, начальная настройка графа

На стороне базы данных выполняется поисковый запрос, возвращающий отфильтрованный набор документов (объектов) вместе с их данными. Программный интерфейс получает эти данные и трансформирует их во внутреннюю структуру графа – JSON объект, содержащий список узлов и список рёбер графа. Формат структуры показан в таблице (Таблица 6.1).

Таблица 6.1 Пример структуры узлов и ребер графа

Структура	Описание ключей

{
"nodes": [
{
"id": "{id}",
"name": "{name}",
"color": "{color}",
"val": "{value}",
"highlighted": "{bool}"
}
],
"links": [
{
"source": "{source}",
	«id» – идентификатор объекта,
	«name» – название узла,
	«color» – цвет узла,
	«val» – вес узла, который используется для его размера,
	«highlighted» – логическое значение для визуального выделения определенных узлов,
	«source» – идентификатор узла (откуда выходит ребро),
	«target» – идентификатор узла (куда входит ребро).

"target": "{target}" }] }	

В программном инструменте реализован неориентированный граф, взвешенный по вершинам (вес вершины определяется количеством инцидентных ей рёбер) [158]. Это означает, что связи между узлами считаются двунаправленными, а значимость узла может быть вычислена на основе числа связей (что удобно для визуального выделения крупных узлов).

Данные хранятся в виде JSON документов произвольной вложенной структуры в ElasticSearch. Перед построением графа система получает от ElasticSearch описание структуры индекса – mapping, содержащий перечень полей и типов данных каждого поля. Mapping позволяет узнать вложенность интересующих полей, однако напрямую не задаёт отношений между полями.

При работе с большими базами данных важно иметь возможность ограничивать объём выводимых данных. В разработанном под руководством автора программном инструменте реализованы два режима фильтрации: стандартные фильтры и расширенный режим. Стандартные фильтры позволяют задать простые условия (например, «год публикации» > 2015, «страна» = Россия и т.п.) через GUI – они будут автоматически включены в запрос к Elasticsearch. Расширенный режим предоставляет пользователю возможность вручную написать DSL-запрос (Domain Specific Language) для Elasticsearch, задавая произвольные условия фильтрации. Оба подхода могут комбинироваться, что обеспечивает многоуровневую фильтрацию данных перед построением графа [159].

Кроме фильтров, предусмотрен механизм правил стилизации узлов. Пользователь может указать группу узлов (например, все узлы типа «Организация» или узлы, удовлетворяющие некоторому условию) и задать для них особый стиль: цвет, размер, форму значка и т.д. Также можно определить условие применения правила (например, подсветить узлы-журналы, если «число публикаций» > 100). Правила позволяют визуально выделить

существенные элементы графа, тем самым упростив анализ (по сути, выполняется элементарная кластеризация или классификация узлов по заданному признаку). В программном инструменте реализован простой декларативный язык правил, понятный пользователю (например: if type = "Organization" and count > 100 then color = "yellow"). При построении графа правила применяются последовательно: проверяются условия для каждого узла, и при выполнении назначаются указанные атрибуты отображения.

Реализованный программный инструмент может быть интегрирован в существующие информационные системы. Использование веб-технологий (React, FastAPI) делает систему кроссплатформенной и расширяемой, а открытые библиотеки (d3, three.js) позволяют модифицировать функциональность под новые требования [160].

Рассмотрим применение предложенного метода для решения конкретных аналитических задач, а именно визуальный анализ публикационной активности научной организации и выявление международного сотрудничества научной лаборатории.

6.1.2 Визуальный анализ публикационной активности организации

Разработанный программный инструмент использован для решения ряда актуальных научно-технических задач, продемонстрировавших его практическую применимость. Для экспериментов использовались реальные данные Объединённого института ядерных исследований (ОИЯИ) – крупного международного межправительственного научного центра.

Были собраны данные 48 835 научных публикаций, связанных с сотрудниками Института за 1957-2024 года. На основе этих данных построен граф по двум полям за 2023-2025 гг.:

1) affiliations.name (название аффилиации, организации),
2) keyword (ключевые слова публикации).

Вершины первого типа (организации) отображены красным цветом, второго типа (ключевые слова) – розовым. Всего в графе 7088 узлов и 89575 ребер (Рисунок 6.6). Полученный граф отражает взаимосвязи организаций по

тематическим направлениям исследований: узлы-организации соединены с узлами-терминами, что означает участие данной организации в работах по соответствующей тематике. Видно, что граф разбивается на несколько кластеров, каждый из которых соответствует определённой тематике исследований. Анализ этих кластеров позволяет выявить, с какими внешними организациями ОИЯИ взаимодействует по каждой тематике.

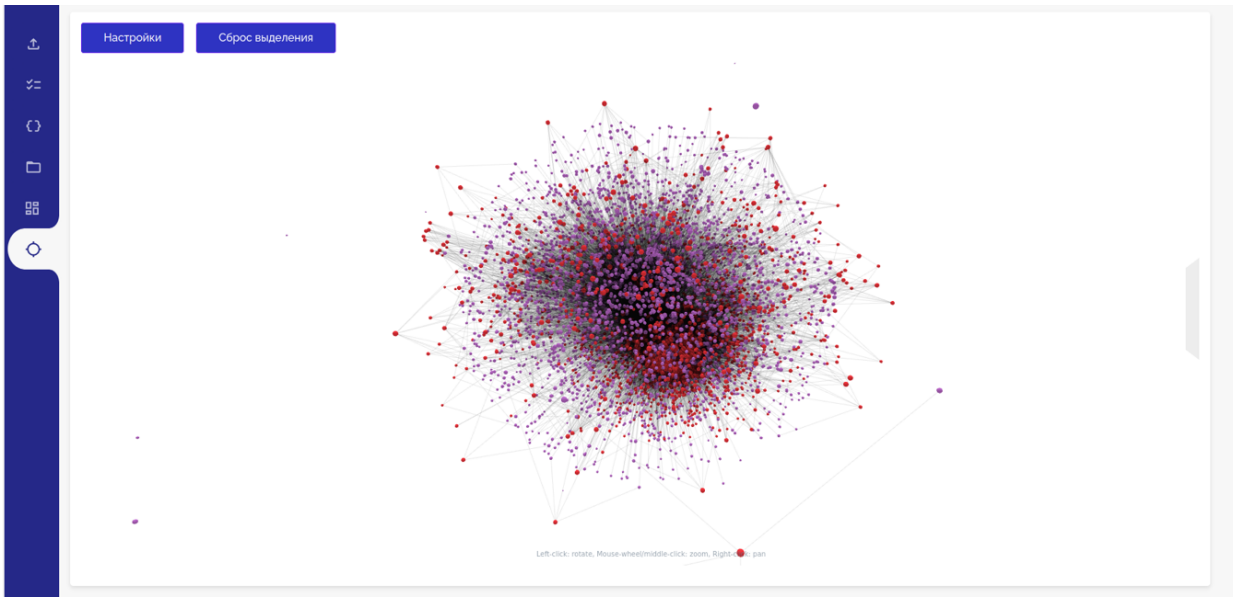


Рисунок 6.6 Граф публикационной активности ОИЯИ

Каждый кластер на рисунке (Рисунок 6.6) представляет совокупность организаций, связанных через общие ключевые слова, то есть совместную научную тематику. Например, один из крупнейших кластеров соответствует направлению физики частиц. Для его детального изучения был использована функциональность фильтрации по запросу: «quark gluon plasma» OR «jets» (Рисунок 6.7).

В результате получен граф из 887 узлов и 15419 рёбер, содержащий узлы ОИЯИ и организаций-партнёров, а также соответствующие ключевые слова. В этом графе прослеживаются связи ОИЯИ с внешними организациями.

Рисунок 6.7 Фильтрация по запросу: «quark gluon plasma» OR «jets»

На отфильтрованном графе кластеры интерпретируются как тематико-географические группы: каждая группа узлов объединяет определённую тему исследований и организации, совместно работающие по этой теме.

6.1.3 Выявление международного сотрудничества научной лаборатории

В дополнение к обзору общей активности ОИЯИ рассмотрим частный случай – научную деятельность конкретного подразделения. В качестве объекта выбрана Лаборатория информационных технологий им. М.Г. Мещерякова ОИЯИ. Собран набор данных о публикациях сотрудников ЛИТ. Построен граф (Рисунок 6.8), в котором присутствуют узлы трёх типов:

1) *affiliations.name* – название аффилиации (фиолетовый),
2) *keywords* – ключевые слова (розовые),
3) *countries* – страна (красные).

В графе 305 узлов и 1377 рёбер. Такая визуализация позволяет проследить, с какими организациями и из каких стран сотрудничает лаборатория, и по каким научным направлениям.

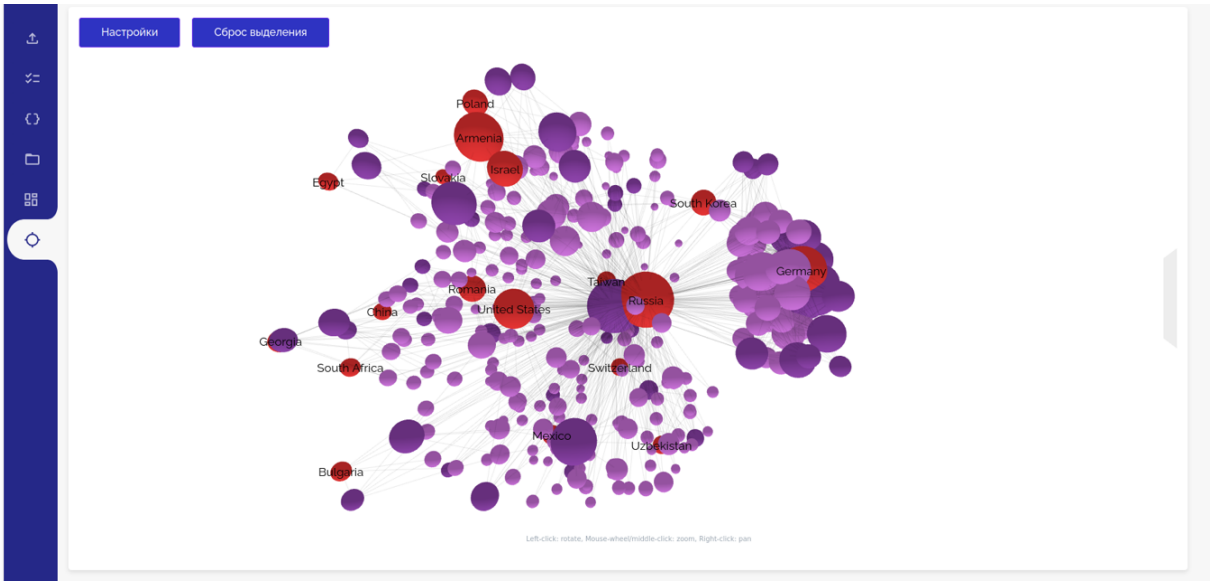


Рисунок 6.8 Граф взаимосвязи ЛИТ ОИЯИ с другими организациями

Визуализация демонстрирует явные и неявные связи ЛИТ с другими организациями и странами. Например, один из кластеров объединяет узлы, связанные с тематикой, близкой к исследованиям ЛИТ, где партнёрами выступают организации из Германии и Южной Кореи. Из этого кластера видно, что по выбранной тематике ЛИТ чаще всего публикуется совместно с учреждениями из Германии (большинство организаций в кластере – немецкие), но также присутствуют совместные работы с учёными из Южной Кореи. Это может указывать на потенциальную трёхстороннюю коллаборацию.

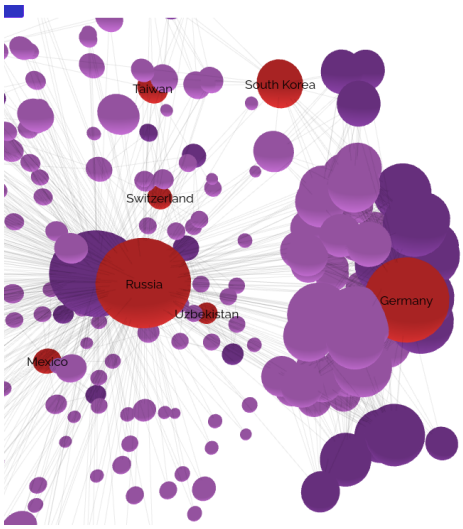


Рисунок 6.9 Кластер ЛИТ ОИЯИ с Германией и Южной Кореей

Из построенного графа (Рисунок 6.9) видно, что Южная Корея написала две коллаборационных статьи вместе с Россией и Германией (Event reconstruction in the RICH detector of the CBM experiment at FAIR [161]) и вместе с Россией и Китаем (Тайванем) (Simulations of Wave Motions in Magnetically Polarized Gas-Dust Interstellar Media [162]).

Из отфильтрованного становится возможным определить организации и исследовательские направления, на которых фокусируются страны, что в свою очередь позволяет сформировать список приоритетных партнеров для будущего взаимодействия. На рисунке (Рисунке 6.10) представлен граф по публикациям ЛИТ ОИЯИ, построенный на данных по стране и ключевых словах для выявления совместных исследований.

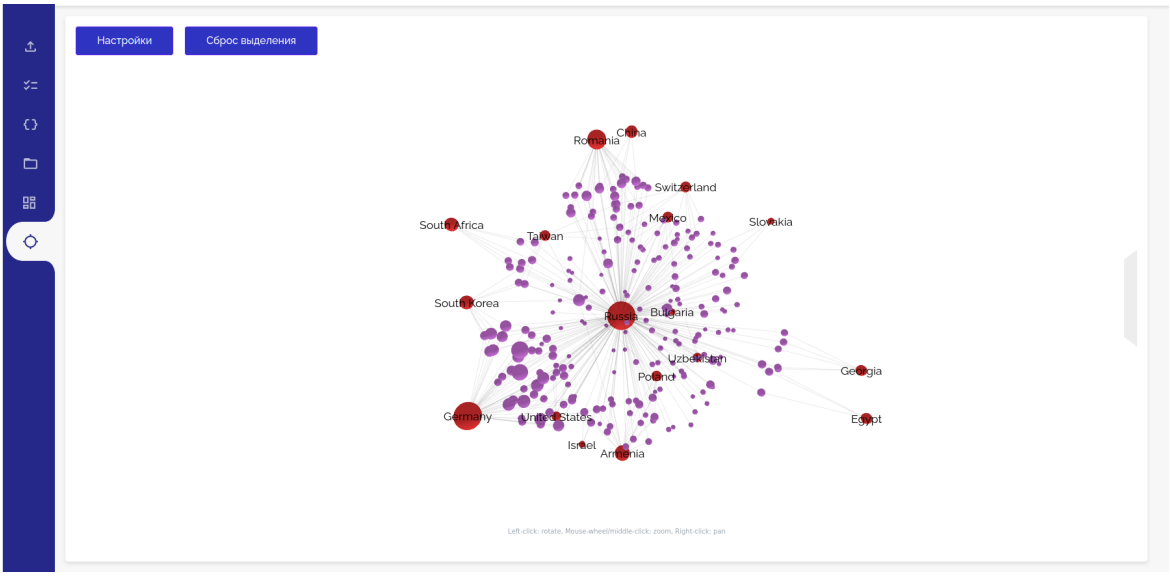


Рисунок 6.10 Граф по странам с часто исследуемыми направлениями

Выделяются несколько географических кластеров, которые проводят исследования совместно с ЛИТ ОИЯИ:

1) Германия и Южная Корея,
2) Армения, Израиль и США,
3) Румыния и Китай.

Такое разбивание показывает, что научное сотрудничество ЛИТ ОИЯИ распределяется по нескольким географическим направлениям.

Был построен граф, отражающий межинституциональные связи на уровне аффилиаций авторов (Рисунок 6.11).

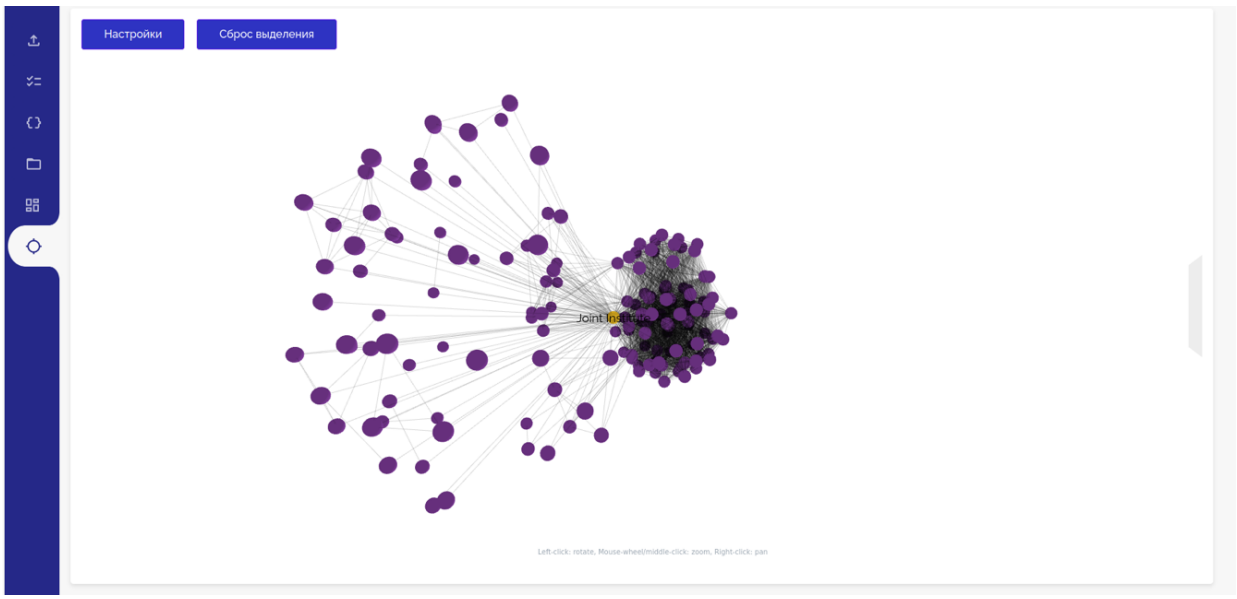


Рисунок 6.11. Граф взаимосвязи ЛИТ ОИЯИ с другими организациями

В этом графе узлы представляют организации (аффилиации), а рёбра – совместное авторство: две организации соединены ребром, если их сотрудники совместно публиковали научную работу. Такой граф позволяет выявить наиболее частые комбинации организаций, выступающих соавторами. Подобные визуализации полезны для анализа структуры научных коллективов и планирования новых коллабораций.

Использование разработанного инструмента для анализа деятельности ЛИТ ОИЯИ позволило определить ключевых партнёров лаборатории за рубежом, основные совместные направления исследований, а также выявить скрытые связи между партнёрами (например, через третьи страны или общие проекты). Эти сведения имеют практическую ценность для формирования стратегии международного сотрудничества: зная, какие организации и страны наиболее активно вовлечены в совместные публикации, можно планировать укрепление связей с ними, а также искать новые точки роста (например, тематические ниши, где есть потенциал для расширения коллаборации).

В рамках проведенной апробации разработанного графового инструмента были рассмотрены различные наборы научно-технических данных. Экспериментальная проверка подтвердила, что инструмент способен обнаруживать и наглядно визуализировать как явные, так и неявные (скрытые)

связи между информационными объектами. Аналитикам предоставлены гибкие механизмы фильтрации данных и настройки отображения графов, что позволяет адаптировать визуализацию под конкретные задачи исследования. В результате работы получен программный прототип, который на основе произвольного набора данных из документоориентированной базы с разнородной структурой способен построить графовую модель связей. В ходе тестирования были успешно опробованы механизмы многоуровневой фильтрации, применение правил для выделения групп узлов, а также интерактивное получение дополнительной информации (метаданных) при выборе узлов графа пользователем.

6.2 Программный инструмент построения научно-технологического ландшафта

6.2.1 Методика построения научно-технологического ландшафта

Появление инструментов для научного картирования (science mapping) и визуализации больших данных стимулировало многочисленные исследования, посвящённые методам представления научных ландшафтов. Так, Ч. Чен в работе [163] провёл обзор методов научного картирования, а в обзоре “Atlas of Knowledge” [164] продемонстрированы способы визуального отображения структуры знаний на различных уровнях от индивидуальных исследований до глобальной науки. Развитию направлений визуальной аналитики способствовали фундаментальные работы по интеграции методов визуализации и анализа данных: например, [165] определили концепцию визуальной аналитики, её процесс и основные задачи.

Библиометрические и визуальные анализы научных ландшафтов активно применяются, к примеру, в медицине и материаловедении. Так, в работе [166] выполнена визуализация научного поля, посвящённого взаимосвязи аденомиоза и бесплодия, на основе данных 2000–2024 гг., что позволило выявить ключевые направления исследований и их взаимосвязи. В работе [167] с помощью библиометрического анализа составили карту научного знания по проблематике «старение на месте» (aging in place), выделив

основные кластеры исследований и эволюцию тематики. Также предпринимаются попытки совместить библиометрический анализ и патентный анализ для получения целостного научно-технологического ландшафта. Например, в области сенсорных технологий проведён патентно-библиометрический анализ использования графеновых биосенсоров, позволивший проследить научный ландшафт данной технологии и перспективы её развития [168].

Одной из ключевых задач при визуализации ландшафта является классификация и агрегирование огромного числа научных публикаций по тематическим областям. В последние годы предложены методы автоматического анализа тенденций на основе ключевых слов. В работе [169] представлен подход к анализу исследовательских трендов, автоматически выделяющий ключевые слова из множества статей и строящий на их основе структуру исследовательского поля. Подобные методы позволяют формировать так называемые карты знаний или ландшафты конкретных научных направлений, показывая кластеры тем и их эволюцию.

Научно-технологический ландшафт (НТЛ) постоянно меняется, и его динамичность затрудняет полноценное отражение традиционными методами обзора литературы или экспертного анализа. Понимание текущего состояния и тенденций НТЛ необходимо как государственным деятелям для принятия взвешенных решений о распределении приоритетов, так и организациям для стратегического планирования проектов и опережения конкурентов в быстро меняющейся среде. Традиционные подходы (анализ публикаций вручную, экспертные консультации) не успевают за темпами роста информации и не позволяют оперативно выявлять новые тенденции. В таких условиях требуется разработка специальных программных инструментов, снимающих с исследователя трудоёмкие процессы поиска и обработки данных за счёт автоматизации и предоставления наглядного представления результатов.

Отсутствие или недостаток данных, несоответствие качества информации, сложность работы с разнообразными и многомерными данными,

неоднозначность интерпретации полученных результатов, динамичные изменения в сфере технологий, и высокая сложность моделирования научных и технологических процессов – все эти факторы требуют от исследователей глубоких знаний, опыта и тщательной методологии анализа [170].

Основная цель создания инструмента анализа НТЛ – помощь аналитикам и руководителям в принятии обоснованных решений и стратегическом планировании, достигаемая через наглядную визуализацию.

Под ***научно-технологическим ландшафтом*** понимается текущее состояние и развитие научно-технических областей в конкретном регионе или стране, то есть коллективная экосистема научных знаний, технологических достижений, организаций и специалистов, определяющих прогресс в этих областях [171]. Иными словами, НТЛ отражает структуру и динамику науки и технологий в определённых границах.

Результатом разработанной автором модели НТЛ является системное графическое представление структуры, текущего состояния и динамики развития области/областей научного знания в пределах страны или группы стран. Визуальная модель ландшафта позволяет наблюдателю определить, какие тематические направления активно развиваются, как они распределены во времени и каково их относительное значение.

Каждый исходный цифровой объект ($Obj, AObj, CAObj$) соотносится с определённым временным интервалом и тематикой, а также вносит вклад в счётчик соответствующей тематической категории и даты. В результате совокупность записей представляет собранные статистики вида (тема, дата) → количество публикаций. Именно такие агрегированные данные необходимы для построения графиков НТЛ.

Обозначим множество входных объектов (научная публикация/патент) через A, а множество данных, которые необходимо получить из него, – через B. Таким образом,

$$A = \{Obj = (a_1, a_2, \dots, a_N) | a_i - \text{характеристика объекта } Obj, \quad i = 1,2, \dots, N\}, \tag{6.6}$$

где $a_i \in S \cup D$.

Множеству объектов A поставим в соответствие множество $B = \{(v_1, v_2, v_3)\}$ следующим образом:

из каждого объекта $Obj \in A$ выделим набор значений (v_1, v_2, v_3), где

$v_1 = DATETIME\big(Obj(a_d)\big)$ – значение характеристики $a_d("date")$ объекта Obj в формате DATETIME (временной интервал), пример: 31.01.2025.

$v_2 = Obj(a_t)$ – значение характеристики $a_t("topic")$ объекта Obj в формате string (тематика), пример: «Biochemistry & Molecular Biology».

$v_3 = \begin{cases} 1, \text{если } (v_1, v_2, C) \notin B, \text{где } C \in \mathbb{N} \\ C + 1, \text{если } (v_1, v_2, C) \in B \end{cases}$ – количество объектов Obj во временном интервале v_1 по тематике v_2.

Для построения НТЛ необходимо разработать методы программной классификации цифрового объекта, классификацию можно проводить относительно рубрик (в основе классификации лежат термины) и относительно более крупных академических рубрик (группы рубрик) [172, 173]. Соответственно, в первом случае, когда мы относим документ к рубрике (классу), мы говорим о рубрикации первого порядка; во втором же случае, когда документ относится к группе рубрик (классов), речь идет о рубрикации второго порядка.

«*Статистический классификатор первого порядка*» (СКПП) позволяет определить распределение релевантности поступающего документа среди описанных классов на основе терминологической базы и весовых характеристик, которые обозначают принадлежность термина к тому или иному классу.

«*Статистический классификатор второго порядка*» (СКВП) позволяет определить распределение релевантности поступающего документа среди описанных групп на основе уже имеющихся данных о степени релевантности описанных классов в СКПП.

Рассмотрим механизм построения СКПП на базе классификатора «Web of Science», основными этапами разработки являются:

1. Составление списка рубрик – классов, по которым должна проводиться классификация входных документов.

2. Составление словаря характерных для определенных рубрик понятий – терминов.

3. Формализованное выражение «принадлежности» (отношения) каждого термина к рубрикам.

Терминологическая база для разрабатываемого СКПП «Web of Science» включает в себя выборку из публикаций «Web of Science» за последние пять лет объемом 235 рубрик и 27 326 уникальных терминов.

К каждому из терминов составлен поисковый образ. Поисковой образ – это шаблон представления термина на языке регулярных выражений, который регламентирует положение термина в тексте с учетом его морфологии, регистра отдельных символов и вхождения лексических единиц между его составными частями. Это необходимо для того, чтобы распознавать термин, несмотря на вариативность его написания. Например, поисковый образ для термина *honor* имеет вид *honou?rs?* и позволяет найти соответствие в тексте следующим значениям: *honor*, *honour*, *honors*, *honours*.

Отношения терминов к рубрикам являются нормированными значениями и выражаются действительными числами от нуля до единицы. Таким образом, если термин встречается исключительно в документах, классифицированных под рубрику «А», то его вес (weight) равен 1 для рубрики «А» и 0 для всех остальных рубрик. То есть, чем чаще термин встречается в документах, классифицированных под данную рубрику, тем ближе к 1 определяется его вес по отношению к данной рубрике.

Подсчет весовых характеристик производился путем статистического анализа употребления ключевых слов в реферативной базе данных «Web of Science» за пять лет по методике TF-IDF.

Таким образом в СКПП насчитывается 72 493 связей терминов с рубриками. Способом хранение данной информации является файл в формате JSON, который загружается в базу данных для дальнейшего использования на стадиях выявления рубрики анализируемого документа. Образец одного из терминов СКПП «Web of Science» изображен на рисунке (Рисунок 6.12).


```
{
  "hippocampal volume": {
    "rubrics": [
      ["Geriatrics & Gerontology", 0.04],
      ["Neuroimaging", 0.548],
      ["Psychiatry", 0.302],
      ["Psychology, Clinical", 0.222]
    ],
    "image": "(?<!\\w)hippocampal(.{0,3}?|(\\s?\\w+\\s){0,2})volume"
  }
}
```


Рисунок 6.12 Образец объекта статистического классификатора первого порядка (формат файла – JSON)

Рубрикатор состоит из объектов – терминов (на Рисунке 6.12 представлен родительский элемент «hippocampal volume»), каждый из которых включает в себя следующие дочерние элементы: *«image»* и *«rubrics»*. Элемент *«image»* располагает значением поискового образа в виде строки, а элемент *«rubrics»* располагает значениями связей, характеризующих отношение термина к рубрикам с весовыми характеристиками, где каждая связь представлена массивом из двух элементов: название рубрики и весовая характеристика. Всего таких объектов в классификаторе, как и общее количество терминов, – 27 326.

Пример файла с данными, полученными после обработки единичного документа при помощи классификатора, представлен на рисунке (Рисунок 6.13).

Полученный документ включает в себя два головных элемента: термины (*«termins»*) и рубрики (*«groups»*). Внутри элемента *«termins»* располагаются термины, объединенные в соответствующие рубрики. Каждый термин обладает собственным названием, своим поисковым образом, весом по

отношению к данной рубрике, количеством упоминаний в обработанном тексте и значением, определяющим относительный вес, характеризующий значимость термина, при определении документа к данной рубрике.


```
{"termins": {
  "Biochemical Research Methods": [
    ["normalization", "(?<!\\w)normalization", 0.39, 10, 0.5394190871369294],
    ["replication", "(?<!\\w)replication", 0.63, 3, 0.26141078838174275]
  ],
  "Geochemistry & Geophysics": [
    ["boundary", "(?<!\\w)boundary", 0.69, 1, 0.5798319327731092],
    ["high pressure", "(?<!\\w)high(.{0,3}?|(\\s?\\w+\\s){0,2})pressure", 0.502, 1, 0.42016806722689076]
  ]
},
  "groups": [
    [
      "Web Of Science",
      1,
      [
        ["Biochemical Research Methods", 0.0037],
        ["Geochemistry & Geophysics", 0.0006]
      ]
    ]
  ]
}
```

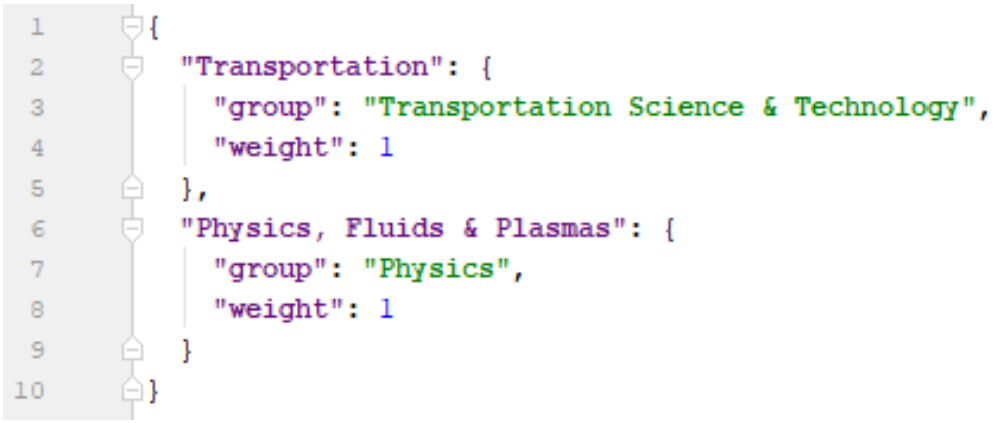

Рисунок 6.13 Образец JSON файла на выходе классификатора первого порядка

Дочерними элементами объекта *«groups»* являются группы, в которые объединены рубрики классификатора с относительными весовыми характеристиками, так как статистический классификатор является СКПП, то определена всего одна группа – сам классификатор «Web of Science», где общая весовая характеристика равна единице.

Такое совокупное представление обозначено как статистический (распределенный) классификатор, поскольку каждый его объект относится хотя бы к одному классу (рубрике), но при этом имеется распределение, поскольку объекты с разным весом (вероятностью) относятся к разным рубрикам.

Классификация второго порядка необходима для повышения точности распределения документов среди рубрик, потому что некоторые категории, такие как «Философия», обладают множеством общих терминов, в таком случае почти каждый поступающий цифровой объект можно отнести к «Философии», что не является удовлетворительным результатом работы классификатора. Однако если объединить технические рубрики «Web of Science» в группы, то такие группы могут составить, так называемую «конкуренцию», общим направлениям, таким как «Философия».

СКВП так же, как и СКПП, состоит из набора объектов [174]. Каждый объект представляет собой сопоставление рубрики СКПП с группой в СКВП. Таким образом, название элемента – это рубрика, описанная в СКПП, а его элементы «group», «weight» обладают значениями группы СКВП и весом отношения рубрики к данной группе соответственно (Рисунок 6.14).



```
{
  "Transportation": {
    "group": "Transportation Science & Technology",
    "weight": 1
  },
  "Physics, Fluids & Plasmas": {
    "group": "Physics",
    "weight": 1
  }
}
```


Рисунок 6.14 Образец объекта статистического классификатора второго порядка (тип файла – JSON)

После обработки входного документа при помощи СКВП результаты обработки сохраняются в файл формата JSON, структура которого представлена на (Рисунке 6.15).

```
{
 "groups": [
  ["Energy Science & Engineering",
   0.1265,
   [
    ["Nuclear Science & Technology", 0.0475],
    ["Energy & Fuels", 0.0449],
    ["Engineering, Petroleum", 0.0341]
   ]
  ]
 ],
 "termins": {

  "Engineering, Petroleum": [
   ["reactor", "(?<!\\w)reactor", 0.71, 91, 0.9656254670452847],
   ["steam", "(?<!\\w)steam", 0.72, 2, 0.02152144671947392],
   ["removal", "(?<!\\w)removal", 0.864, 1, 0.01285308623524137]
  ],
  "Nuclear Science & Technology": [
   ["reactor", "(?<!\\w)reactor", 0.986, 91, 0.9607066381156317],
   ["nuclear power plant", "(?<!\\w)nuclear(.{0,3}?|(\\s?\\w+\\s){0,2})power(.{0,3}?|(\\s?\\w+\\s){0,2})plant", 0.364, 6, 0.023340471092077087],
   ["removal", "(?<!\\w)removal", 0.91, 1, 0.009743040685224838],
   ["pwr", "(?<!\\w)pwr", 0.576, 1, 0.0062098510706638]],
  "Energy & Fuels": [
   ["reactor", "(?<!\\w)reactor", 0.936, 91, 0.9644474637681157],
   ["steam", "(?<!\\w)steam", 0.646, 2, 0.014605978260869562],
   ["removal", "(?<!\\w)removal", 0.946, 1, 0.010756340579710142],
   ["electricity market", "(?<!\\w)electricity(.{0,3}?|(\\s?\\w+\\s){0,2})market", 0.122, 5, 0.006906702898550723],
   ["generator", "(?<!\\w)generator", 0.146, 2, 0.003283514492753622]
  ]
 }
}
```


Рисунок 6.15 Образец JSON файла на выходе классификатора первого порядка

Результаты обработки – это два головных элемента: «groups», «termins», где в элементе «groups» описано отношение исходного документа к группе СКВП и рубрикам СКПП, а в элементе «termins» – отношение терминов к рубрикам СКПП. Каждый объект представляет собой список из пяти элементов:

1) первый элемент – это название термина (например, "reactor"),

2) второе – его поисковый образ (например, "(?<!\\w) reactor"),

3) третий элемент – его вес по отношению к рубрике (например, 0.936),

4) четвертый элемент – частота употребления данного термина в исходном документе (например, 91),

5) пятый элемент – это нормированный вес термина, отображающий его значительность в формировании данной рубрики (например, 0.964447463768115 71).

После осуществления классификации всех документов по рубрикам осуществляется сохранение итоговых данных. Каждая запись цифрового объекта дополняется характеристикой «*target_field»* – названием научной области, к которой отнесена данная публикация по результатам рубрикации, и значением соотнесения цифрового объекта с названием научной области «*target_field_accuracy»* (вероятностью принадлежности к области).

Разработанный под руководством автора программный инструмент построения НТЛ обладает следующими функциональными возможностями, направленными на поддержку аналитического процесса посредством визуализации необходимых пользователю данных:

1. Выбор источника данных (индекса в базе) – позволяет выбирать конкретный индекс (набор) данных в хранилище Elasticsearch, определяя тем самым, с каким массивом публикаций работать (например, данные определённой страны или тематической коллекции).

2. Выбор временного диапазона – дает возможность задать интересующий период (год, квартал, месяц или произвольные даты начала и

конца). Это позволяет отслеживать изменения во времени и выявлять тенденции в НТЛ.

3. Выбор страны/группы стран – обеспечивает фильтрацию данных по географическому признаку. Пользователь может проанализировать ландшафт одной выбранной страны либо сравнить одновременно несколько стран.

4. Выбор ключевых слов – позволяет ввести одно или несколько ключевых слов/фраз для тематической фильтрации. Это актуально при изучении определённых тематик или направлений исследований; система поддерживает автодополнение и коррекцию ввода, чтобы облегчить выбор релевантных терминов.

Для получения необходимых данных пользователю достаточно выбрать индекс (например, «Статьи Японии»), указать временной диапазон (скажем, с 01.01.2015 по 31.12.2020), а затем отфильтровать по странам и ключевым словам. Если выбрано несколько стран, инструмент может либо построить их совмещённый НТЛ, либо отобразить два графика раздельно для сравнения. Если ключевые слова не заданы – выводится сводный график по всем тематикам.

Взаимодействие компонентов происходит следующим образом. Пользователь через GUI формирует набор фильтров; далее frontend формирует запрос к backend (методом POST) с указанными параметрами. Например, запрос на поиск статей по ключевому слову "ядерные технологии" и году 2022 будет отправлен как:

```
POST /search/articles
{
  "query": {
    "bool": {
      "must": [
        { "match": { "title": "ядерные технологии" } },
        { "match": { "year": 2022 } }
```

] }

} }

Графический интерфейс настроек построения НТЛ приведен на рисунке (Рисунок 6.16).

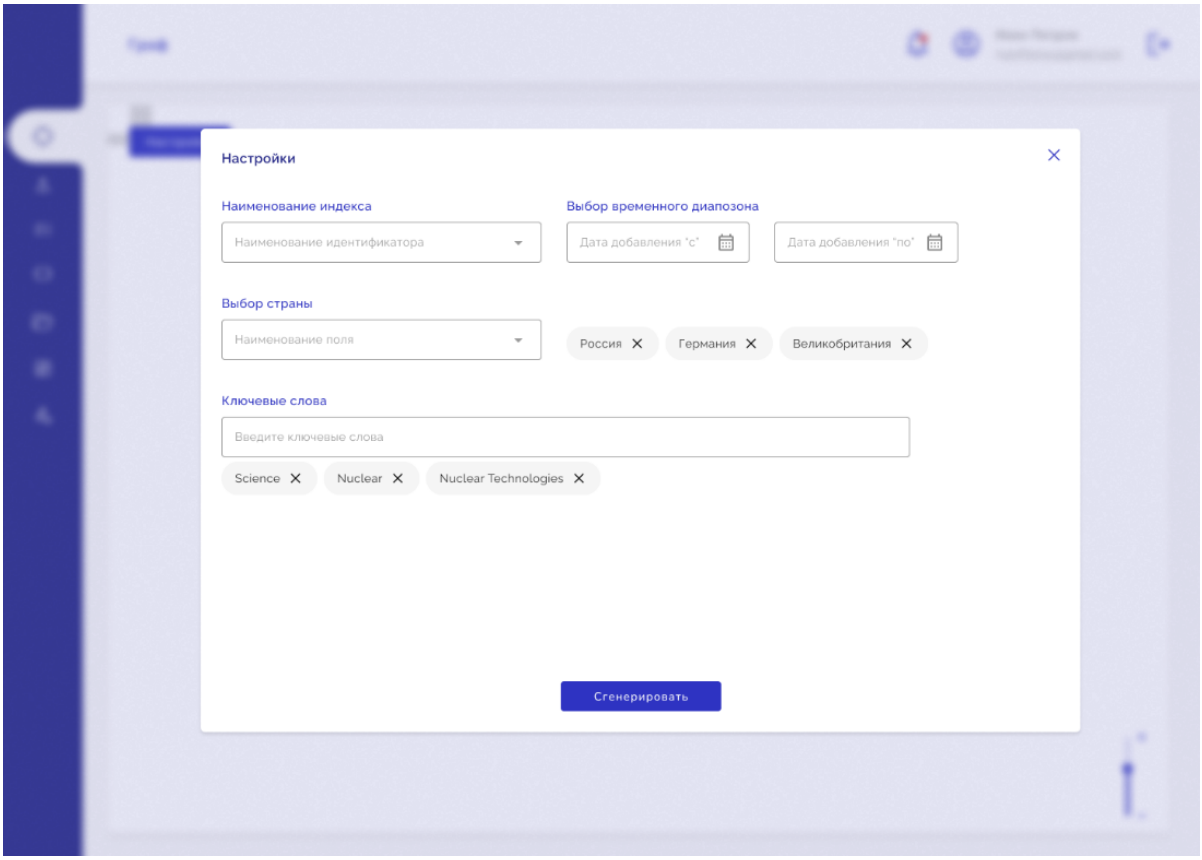


Рисунок 6.16 Графический интерфейс настроек построения НТЛ

При применении фильтра на несколько стран возможно построить как их совместный НТЛ, так и провести сравнительный анализ двух графиков по-отдельности.

Все перечисленные настройки в совокупности создают гибкий механизм настройки вывода больших данных, позволяющий аналитику визуализировать графики как научных областей в рамках одной страны, так и рассматривать научно-технологические ландшафты регионов, включающих в себя несколько стран.

Графический интерфейс позволяет пользователю не только задавать параметры, но и непосредственно взаимодействовать с результирующими графиками. Например, можно вращать 3D-график, изменять масштаб,

наводить курсор на точку поверхности для получения подробной информации (значения количества публикаций, названия тематики, даты). Кроме того, предусмотрен вывод дополнительной информации в виде таблицы – краткой сводки по выбранным фильтрам (например, список наиболее продуктивных авторов или организаций после применения фильтров). Подобное сочетание интерактивной графики и табличных деталей обеспечивает более глубокую аналитическую функциональность.

6.2.2 Практическое применение программного инструмента построения научно-технологического ландшафта

Для оценки работоспособности инструмента были проведены эксперименты на различных наборах данных научно-технической направленности. Экспериментальная проверка подтвердила способность системы преобразовывать большие неструктурированные массивы публикаций в информативные визуализации, а также показала эффективность встроенных механизмов фильтрации и настройки отображения под задачи аналитика. Были протестированы несколько типовых сценариев использования.

Временной диапазон для одной страны. В этом случае пользователь выбирает, например, одну страну и короткий промежуток (несколько месяцев или один год). Система строит 3D-ландшафт, который практически вырождается в двухмерный график (по оси времени мало точек). Такой режим позволяет детально рассмотреть, какие тематики были активны в конкретный период. При выборе одного года реализована автоматическая подстройка: строится 2D-столбчатая диаграмма (ось X – тематики, ось Y – число публикаций за год), что нагляднее для сравнения тематик внутри года (Рисунок 6.17).

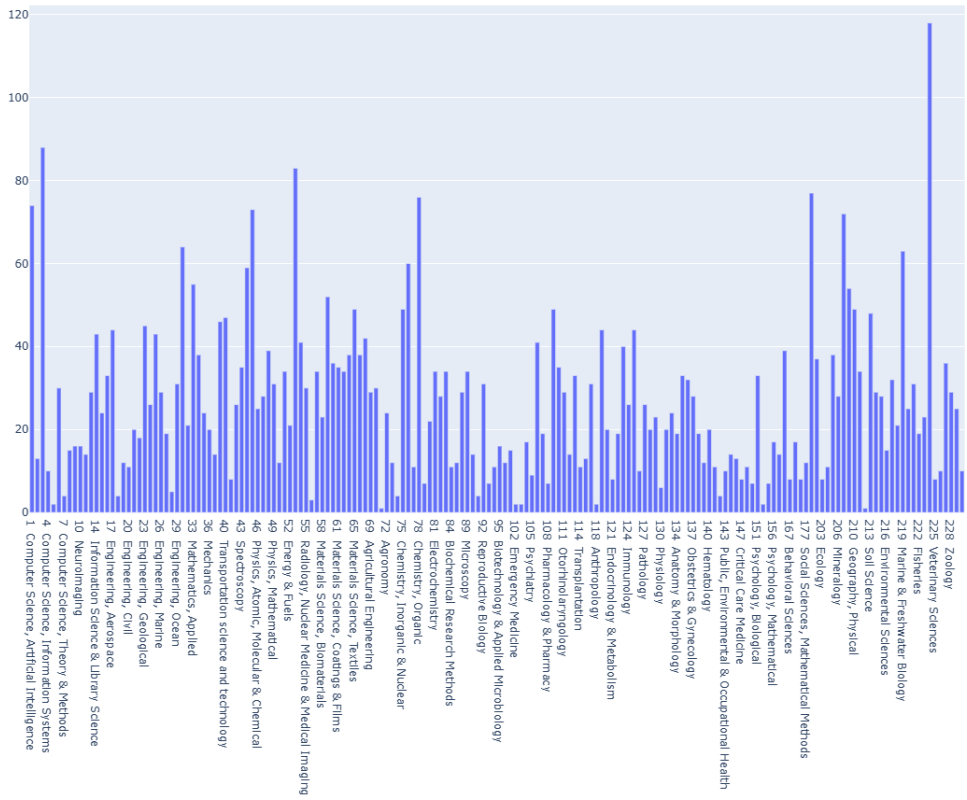


Рисунок 6.17 Пример фильтрации по одному году.

Совместная визуализация ряда стран. Пользователь может выбрать сразу несколько стран и формировать их совмещённый НТЛ за заданный период. Инструмент объединяет данные по выбранным странам и отображает единую поверхность, показывающую суммарный ландшафт. Это полезно для оценки общей динамики научной активности, например, региона (Северная Европа, Азия и пр.) (Рисунок 6.18).

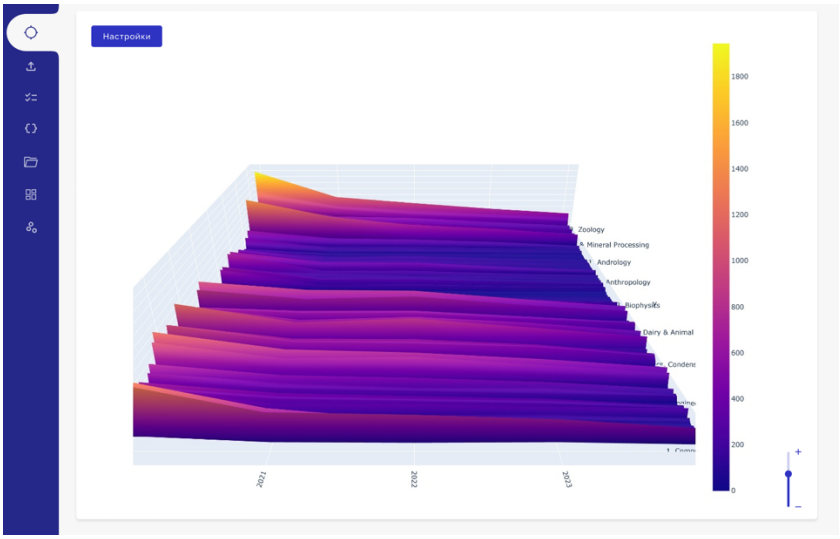


Рисунок 6.18 Визуализация НТЛ по нескольким странам

Сравнительный анализ двух стран. Для непосредственного сравнения двух наборов данных предусмотрен режим с двумя графиками (Рисунок 6.19). Например, построены параллельно ландшафты Японии и Республики Корея за 2010–2021 гг. на отдельных графиках. Анализируя их, эксперт может увидеть различия: в какие годы в каждой стране был пик публикаций по определённой теме, какая тематика имеет больший относительный вес в одной стране по сравнению с другой и т.д. Оказалось, что подобное визуальное сравнение очень эффективно для выявления национальной специфики научных приоритетов.

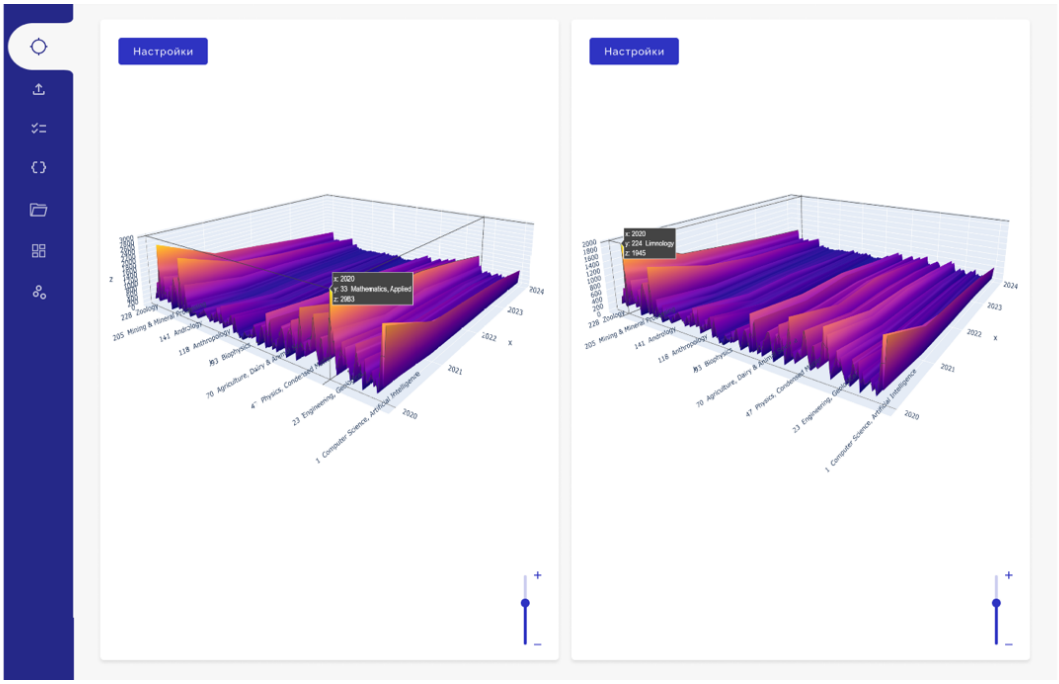


Рисунок 6.19 Визуализация сравнения НТЛ двух стран (Япония и Южная Корея)

Анализ одной научной области по разным странам. Инструмент позволяет фильтровать не только по странам, но и по тематике. Например, можно рассмотреть публикации стран БРИКС по направлению «Искусственный интеллект», по запросу «AI OR Artificial Intelligent OR Big Data». Был построен ландшафт по коллекции публикаций стран БРИКС с 2020-2004 из исходного количества 2,4 млн публикаций, собранных из ведущих реферативных изданий на момент 27.11.2024, по запросу было отобрано – 7120 публикаций (Рисунок 6.20).

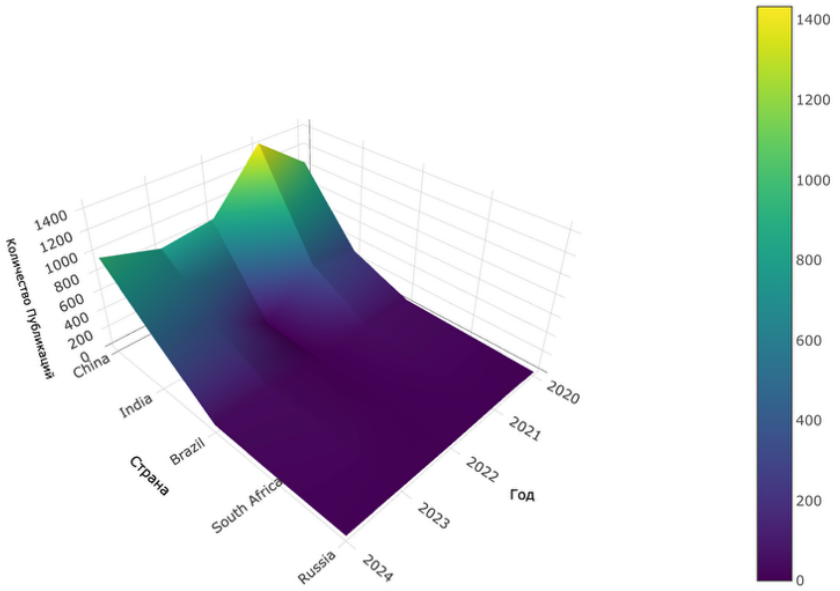


Рисунок 6.20 Научно технологический ландшафт стран БРИКС по направлению «Искусственный интеллект» за 2020-2024 гг.

Практические примеры подтверждают, что инструмент позволяет увидеть динамику научной активности, выявить ключевые темы и проследить изменение количества публикаций по тематикам исследований во времени. Благодаря многоуровневой фильтрации, можно переключаться от макро-анализа (все области науки по стране) к микро-анализу (отдельная тема по нескольким странам). Успешная работа инструмента опирается на предложенную модель цифрового объекта и структуры представления данных.

6.3 Система интеллектуального анализа информационных объектов в решении прикладных научно-технических и социально значимых задач

Разработанные автором в диссертационной работе модели цифровых объектов ($Obj, AObj, CAObj$), методы преобразования данных из разнородных информационных ресурсов; функции, алгоритмы расчета характеристик, методы визуализации образуют систему интеллектуального анализа информационных объектов в научно-технических и социально значимых задачах, графическое представление дано на рисунке (Рисунок 6.21).

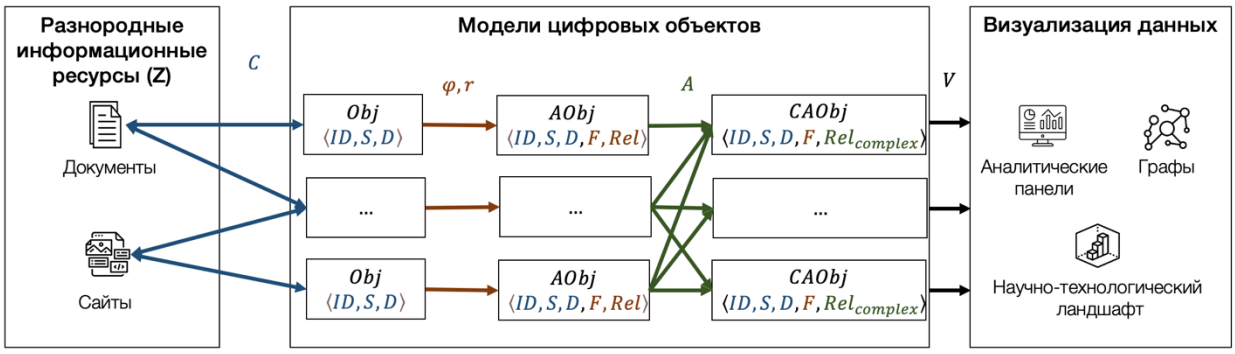


Рисунок 6.21 Графическое представление системы интеллектуального анализа информационных объектов

Предложенную систему интеллектуального анализа информационных объектов можно представить в виде следующий математической модели:

$$\begin{cases} C: Z \to Obj, \text{ где} & Z - \text{множество разнородных информационных ресурсов,} \\ & Obj - \text{базовая модель цифрового объекта,} \\ & C - \text{методы преобразования данных;} \\ M: Obj \to CAObj, \text{ где} & CAObj - \text{модель комплексного цифрового объекта,} \\ & M = A \circ r \circ \varphi - \text{функции, алгоритмы расчета характеристик;} \\ & (CAObj = AObj \text{ при } A = \emptyset) \\ V: CAObj \to K, \text{ где} & K - \text{решенные аналитические задачи,} \\ & V - \text{методы визуализации.} \end{cases} \quad (6.7)$$

На основе отдельных методов, алгоритмов, программных средств решен ряд актуальных научно-технических задач в интересах Федеральных органов исполнительной власти, организаций контура Госкорпорации «Росатом».

По государственному заданию Министерства образования и науки РФ № 2.12611.2018/12.1 ***«Обеспечение каталогизации и хранения научно-технической информации, полученной из различных неструктурированных источников»*** реализованы методики составления статистических классификаторов первого и второго порядка, а также методы и программное средство потоковой рубрикации научно-технической

информации. При тестовой обработке полнотекстовых статей, входящих в реферативную базу данных «Web of Science», время обработки 1304 файлов (1,76 Гб) составило 58 минут (1,82 Гб / час), из них классифицировано 1292 (99%).

В рамках выполнения НИР ***«Разработка программы выборки данных по свойствам и структурам облученных реакторных материалов из мировых источников информации»*** (договор № 1707 от 29 августа 2022 г. с ВНИИА им. Н.Л. Духова) предложенная система интеллектуального анализа данных позволила существенно сократить время на выполнения аналитических исследований. За три месяца работы в автоматизированном режиме был обеспечен сбор более 40 тысяч публикаций, проведен анализ их на предмет наличия целевых изображений и табличных данных из них было отобрано 534 целевые публикации. В результате анализа таблиц и изображений из целевых публикаций экспертами на предмет фотографий микроструктур, диаграмм получено 8700 записей, описывающих свойства материалов, а также 1650 изображений микроструктур [175]. Решение подобной задачи в ручном режиме сопоставимым составом экспертов заняло бы более года.

По договору №23313/13 от 26.09.2023 г. на выполнение НИР между ФГУП «РФЯЦ-ВНИИТФ им. академ. Е.И. Забабахина» и НИЯУ МИФИ по теме ***«Создание методики обнаружения признаков нарушения обязательств по ядерному нераспространению государством-импортером реакторов на быстрых нейтронах с замкнутым ядерным топливным циклом на основе компьютерного анализа открытой информации»*** разработана и реализована методика обнаружения признаков нарушения обязательств по ядерному нераспространению государством-импортером реакторов на быстрых нейтронах с замкнутым ядерным топливным циклом на основе компьютерного анализа открытой информации, базирующаяся на предложенной автором комплексной цифровой модели информационного объекта. Разработаны программные инструменты классификации и выделения

научных публикации по чувствительным технологиям для дальнейшего анализа экспертами [83, 176].

В деятельности ***Исследовательского центра по искусственного интеллекта по направлению «Транспорт и логистика» НИЯУ МИФИ,*** реализующего программу мероприятий по договору от 27.12.2023г. № 70-2023-001309 с АНО «Аналитический центр при Правительстве Российской Федерации», разработанные методы преобразования данных из разнородных информационных ресурсов сформировали базис для аналитического фреймворка обработки и представления научно-технической информации [177].

По Государственному заданию Министерства образования и науки РФ №2.12915.2018.12.1 ***«Разработка и апробация информационной системы комплексной антисуицидальной интернет-профилактики»*** проведено исследование и разработка цифровых профилей персон, характеризующихся деструктивным поведением. Проведена апробация на выборке из 1 миллиона цифровых профилей объектов, выявлено, что разработанная методика успешно идентифицируют цифровые объекты склонных к деструктивным действиям. Полученные результаты неоднократно докладывались на совещаниях Межведомственной рабочей группы при Министерстве образования и науки Российской Федерации.

На основе разработанных моделей и методов реализован программный комплекс, обеспечивающий решение задач сбора, обработки, насыщения и визуализации данных в едином контуре в рамках договора №349ГС1ЦТС10-D5/80243 от 12.12.2022 ***«Разработка и тестирование прототипа мультиагентной системы обработки и представления неструктурированных массивов данных»*** с Фондом содействия инновациям спроектирован и реализован программный комплекс интеллектуального анализа научно-технической информации СИА.Атташе [178], позволяющий:

- загружать пользовательские файлы в информационное хранилище;

- осуществлять потоковый сбор данных из информационных ресурсов сети Интернет;
- оперативно масштабировать систему хранения цифровых объектов (проведено масштабирование на научно-технический объект – патент и новостное сообщение);
- проводить насыщение данных на основе разработанных методов и программных средств (выделять физические, химические элементы и их значения);
- систематизировать набор документов по выделенным сущностям и атрибутам;
- формировать аналитические панели по хранилищу для определения встречаемости сущностей (атрибутов);
- проводить аналитические исследование при помощи графовых представлений и анализа научно-технологического ландшафта.
- создавать и обрабатывать сложные поисковые запросы.

С использованием программного комплекса интеллектуального анализа научно-технической информации выполнен ряд заказных работ в том числе по договорам № 2024-sia-dgk-1 от 15.04.2024 и № 2024-sia-dgk-2 от 15.05.2024.

Предложенная автором система интеллектуального анализа информационных объектов и разработанный на ее основе программный комплекс позволяют решать широкий комплекс научно-технических и социально значимых задач.

ВЫВОДЫ ПО ГЛАВЕ 6

1. Предложенные модели, методы и реализованный на их основе программный комплекс позволяют решать широкий спектр научно-технических и социально значимых задач, включая мониторинг социальных процессов, анализ развития научных направлений, выявление скрытых закономерностей в больших данных и поддержку принятия решений на основе этих знаний.

2. Предложен и разработан программный инструмент выявления явных и неявных (скрытых) взаимосвязей между цифровыми информационными объектами, основанный на методах графового анализа данных и визуальной аналитики. Программный инструмент позволяет находить скрытые взаимосвязи и осуществляет связь между сложными алгоритмами интеллектуального анализа данных и когнитивными возможностями человека, позволяя интерпретировать полученные результаты: благодаря графикам и диаграммам специалист быстрее замечает аномалии и кластеры в данных.

3. Разработан программный инструмент построения научно-технологического ландшафта, предназначенный для анализа больших массивов данных (публикаций, патентов, отчетов) с целью определения тематической близости между различными научными направлениями, странами, что формирует целостную картину научно-технологического ландшафта в выбранной области, выделят центральные узлы (ведущие издания, ключевых авторов) и оценивает плотность взаимодействия между различными направлениями науки.

4. Подтверждена практическая ценность предложенных автором методов, технологий и разработанного программного комплекса для интеллектуального анализа научно-технической информации в рамках выполненных заказных работ. Благодаря модульному подходу, отдельные компоненты успешно адаптированы под конкретные задачи: от анализа связей до мониторинга технологических трендов в атомной отрасли. Решён ряд актуальных научно-технических задач в интересах федеральных органов, связанных со сбором, обработкой и анализом больших массивов разнородных данных, что демонстрирует универсальность и прикладную полезность разработанного инструментария.

ЗАКЛЮЧЕНИЕ

В диссертации разработаны методы, модели и технологии интеллектуального анализа цифровых информационных объектов в научно-технических и социально значимых задачах, агрегирующая модели представления цифрового объекта, методы извлечения и насыщения данных, программные инструменты визуализации разнородной информации.

Основные результаты диссертационной работы заключаются в следующем:

1. Разработана система интеллектуального анализа данных, объединяющая формальную модель представления цифровых объектов, методы автоматизированной обработки данных и специализированные аналитические инструменты. Предложенная система обеспечивает единый подход к работе с разнородной научно-технической и социальной информацией, что подтверждено её успешным применением в нескольких предметных областях.

2. Предложена модель комплексного цифрового информационного объекта, включающая уникальный идентификатор, статические, динамические и вычисляемые характеристики, а также связи (отношения) с другими объектами. Показано, что использование данной модели повышает полнота представления информации о каждом объекте, облегчает слияние данных из разных источников и снижает неоднозначность интерпретации сведений.

3. Разработан набор методов извлечения и насыщения данных из слабоструктурированных информационных источников (PDF-документов, веб-страниц, социальных сетей). В их числе: алгоритмы парсинга текстов научных статей с восстановлением структуры, методы выделения значимых сущностей (ключевых слов, физических величин) из текста, процедура унификации аффилиаций авторов и геокодирования организаций, а также алгоритмы выделения изображений и таблиц из документов. Комплексное применение этих методов позволяет автоматизированно преобразовывать

массивы документов в структурированную базу знаний, пригодную для дальнейшего анализа.

4. Разработана методика аналитического описания социальных объектов и выявления целевых групп пользователей. В рамках методики предложено представлять цифровой профиль пользователя как совокупность статических и динамических характеристик, ранжировать эти характеристики по значимости и вычислять интегральный критерий принадлежности профиля к искомой категории. Методика апробирована при решении практических задач и показала высокую эффективность: автоматическая идентификация целевых профилей достигла точности, сопоставимой с экспертной, значительно сократив при этом объём ручной работы специалистов.

5. Выполнен комплексный анализ научно-технической информации с использованием разработанных инструментов. Реализованы интерактивные панели визуализации, отражающие динамику и географию развития таких областей, как технологии больших данных, биомедицинские исследования, финансовая безопасность и др. Получены новые научно-практические выводы: определены лидирующие страны и организации в указанных областях, выявлены основные тематические кластеры и тренды эволюции тематик, показана степень международного сотрудничества. Полученные результаты подтверждают, что интеграция разнородных данных (научные публикации, проекты, нормативные документы) в едином информационном пространстве значительно расширяет возможности экспертного анализа и принятия решений.

6. Применение программного комплекса в проектах по анализу научно-технической информации (например, создание базы данных свойств материалов для ГК «Росатом») позволило сократить время обработки данных в несколько раз и извлечь из текстов десятки тысяч структурированных фактов. В социально значимых задачах внедрение алгоритмов привело к повышению оперативности реагирования и более точному выявлению целевой аудитории.

7. Реализованы специализированные программные средства интеллектуального анализа данных, в части построения графовых представлений, позволяющих выявлять неявные связи между объектами и автоматизированного построения научно-технологического ландшафта, позволяющего определить и сравнить динамику изменения научно-технологической области стран.

8. Разработан программный комплекс интеллектуального анализа данных, обеспечивающий автоматизированный сбор данных из сети Интернет, их хранение в документно-ориентированном хранилище, обогащение метаданными и аналитическими признаками, а также интерактивную визуализацию результатов через веб-интерфейсы. Структура программного комплекса спроектирована с учётом горизонтальной масштабируемости и модульности, что подтверждено испытаниями: система успешно обработала свыше 50 тысяч документов и легко расширяется для новых типов объектов.

9. Результаты диссертационной работы апробированы в рамках выполнения научно-исследовательских и опытно-конструкторских работ по Государственным заданиям Министерства науки и высшего образования Российской Федерации, Фонда перспективных исследований, организаций контура Госкорпорации Росатом (НИИ «Графит», ФГУП «РФЯЦ-ВНИИТФ им. академ. Е.И. Забабахина», ФГУП ВНИИА им. Н.Л. Духова), Российского энергетического агентства. Научные и технические положения, изложенные в диссертации, использованы в проектах других систем, имеющих специальное назначение.

ПЕРЕЧЕНЬ ИСПОЛЬЗУЕМЫХ СОКРАЩЕНИЙ

БЯМ	Большая языковая модель
ГРЛС	Государственный реестр лекарственных средств
ИАД	Интеллектуальный анализ данных
ИИ	Искусственный интеллект
ЛИТ ОИЯИ	Лаборатория информационных технологий им. М.Г. Мещерякова ОИЯИ
ЛП	Лекарственный препарат
НТИ	Научно-техническая информация
НТЛ	Научно-технологический ландшафт
ОИЯИ	Объединённый институт ядерных исследований
СИ	Международная система единиц
СКВП	Статистический классификатор второго порядка
СКПП	Статистический классификатор первого порядка
СППВР	Система поддержки принятия врачебных решений
СУБД	Система управления базами данных
API	Application Programming Interface
BFI	Big Five Inventory
CSS	Cascading Style Sheets
CSV	Comma Separated Values
DOI	Digital Object Identifier
DOM	Document Object Model
DSL	Domain Specific Language
GPE	Geopolitical entity
GROBID	GeneRation Of BIbliographic Data
GUI	Graphical user interface
IoT	Internet of things
ISO	International Organization for Standardization
JSON	JavaScript Object Notation

LLM	Large language models
NEO-PI	Revised NEO Personality Inventory
NER	Named Entity Recognition
NLP	Natural Language Processing
NoSQL	Not Only SQL
PDF	Portable Document Format
SNA	Social Network Analysis
URL	Uniform Resource Locator
XML	Extensible Markup Language
XPath	XML Path Language
YAKE!	Yet Another Keyword Extractor
YOLO	You Only Look Once

СПИСОК ЛИТЕРТАТУРЫ

1. Castillo C. Big Crisis Data. Social Media in Disasters and Time-Critical Situations // Cambridge University Press, 2016.
2. Vanhala M., Lu C., Peltonen J., Sundqvist S., Nummenmaa J., Järvelin K. The usage of large data sets in online consumer behaviour: A bibliometric and computational text-mining–driven analysis of previous research // Journal of Business Research, Vol. 106, January 2020. pp. 46-59.
3. Lozano M.G., Brynielsson J., Franke U., Rosell M., Tjörnhammar E., Varga S., Vlassov V. Veracity assessment of online data // Decision Support Systems, Vol. 129, 2020.
4. Savall H., Zardet V. The Qualimetrics Approach: Observing the Complex Object. Information Age Publishing, 2011.
5. Huo R., Zeng S., Wang Z., Shang J., Chen W., Huang T., Wang S., Yu F.R., Liu Y. A Comprehensive Survey on Blockchain in Industrial Internet of Things: Motivations, Research Progresses, and Future Challenges // IEEE Communications Surveys & Tutorials, Vol. 24, 2022. pp. 88–122.
6. Gupta S., Kar A.K., Baabdullah A., Al-Khowaiter W.A.A. Big data with cognitive computing: A review for the future // International Journal of Information Management, Vol. 42, 2018. pp. 78-89.
7. Zhang C., Li W., Zhang H., Zhan T. Recent Advances in Intelligent Data Analysis and Its Applications, 2nd Edition // Electronics, Vol. 14, January 2025. P. 228.
8. Sharma S. Markov Chain Monte Carlo Methods for Bayesian Data Analysis in Astronomy // Annual Review of Astronomy and Astrophysics, Vol. 55, August 2017. pp. 213–259.
9. Li G., Lu Z., Wang J., Wang Z. Machine Learning in Stellar Astronomy: Progress up to 2024, February 2025.

10. Peres R.S., Rocha A.D., Leitao P., Barata J. IDARTS – Towards intelligent data analysis and real-time supervision for industry 4.0 // Computers in Industry, Vol. 101, 2018. pp. 138-146.
11. Ferraris C., Amprimo G., Cerfoglio S., Masi G., Vismara L., Cimolin V. Machine-Learning-Based Validation of Microsoft Azure Kinect in Measuring Gait Profiles // Electronics, Vol. 13, November 2024. P. 4739.
12. Sun X., Zhao L., Chen J., Cai Y., Wu D., Huang J.Z. Non-MapReduce computing for intelligent big data analysis // Engineering Applications of Artificial Intelligence, Vol. 129, 2024. P. 107648.
13. Al-Zaiti S.S., Alghwiri A.A., Hu X., Clermont G., Peace A., Macfarlane P., Bond R. A clinician's guide to understanding and critically appraising machine learning studies: a checklist for Ruling Out Bias Using Standard Tools in Machine Learning (ROBUST-ML) // European Heart Journal - Digital Health, Vol. 3, April 2022. pp. 125–140.
14. Kontsewaya Y.; Antonov E.; Artamonov A. 2020 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence: Eleventh Annual Meeting of the BICA Society // Evaluating the Effectiveness of Machine Learning Methods for Spam Detection. 2021. Vol. 190. pp. 479-486.
15. Прокопец Т.Н., Синюк Т.Ю., Рыбалко Ю.А. Методы интеллектуального анализа данных URL: https://rep.polessu.by/bitstream/123456789/26509/1/Metody.pdf#:~:text=Интеллектуальный%20анализ%20данных%20,–%20это%20процесс%20обнаружения%20в
16. Lindley D.V. Regression and Correlation Analysis. London: Palgrave Macmillan UK, 1990.
17. Mughal M.J. Data mining: Web data mining techniques, tools and algorithms: An overview // International Journal of Advanced Computer Science and Applications. 2018. Vol. 9. No. 6.

18. Mughal M.J.H. Data Mining: Web Data Mining Techniques, Tools and Algorithms: An Overview // International Journal of Advanced Computer Science and Applications, Vol. 9, No. 6, Июнь 2018.
19. Meher S.K., Panda G. Deep learning in astronomy: a tutorial perspective // The European Physical Journal Special Topics, Vol. 230, July 2021. pp. 2285–2317.
20. Smith M.J., Geach J.E. Astronomia ex machina: a history, primer and outlook on neural networks in astronomy // Royal Society Open Science, Vol. 10, May 2023.
21. Choudhary S., Sharma K., Bajaj M. Social Networks Analysis and Machine Learning: an Overview of Approaches and Applications // 2023 International Conference on Sustainable Computing and Smart Systems (ICSCSS). June 2023. pp. 123–128.
22. Chu M.N., Huang X., Hsu J.L., Tu H.L. A Social Network Analysis on the Danmaku of English-Learning Programs // Applied Sciences, Vol. 15, February 2025. P. 1948.
23. Koshlan D.I., Tretyakov E.S., Korenkov V.V., Onykij B.N., Artamonov A.A. Proceedings of the VIII International Conference "Distributed Computing and Grid-technologies in Science and Education" (GRID 2018), // Agent technology situational express analysis in assessment of technological development level of the BRICS countries. Dubna, Moscow region, Russia. 2018. Vol. 2267. pp. 436-440.
24. Artamonov A.A., Ananieva A.G., Tretyakov E.S., Kshnyakov D.O, Onykiy B.N., Pronicheva L.V. A three-tier model for structuring of scientific and technical information // Journal of Digital Information Management. 2016. Vol. 14. No. 3. pp. 184-193.
25. ИИ против болезней: как машинное обучение меняет медицину [Электронный ресурс] // Habr.com: [сайт]. [2025]. URL: https://habr.com/ru/companies/magnus-tech/articles/878456/

26. Ajibade S.S.M., Alhassan G.N., Zaidi A., Oki O.A., Awotunde J.B., Ogbuju E., Akintoye K.A. Evolution of machine learning applications in medical and healthcare analytics research: A bibliometric analysis // Intelligent Systems with Applications, Vol. 24, December 2024. P. 200441.
27. Bykanov A.E., Danilov G.V., Kostumov V.V., Pilipenko O.G., Nutfullin B.M., Rastvorova O.A., Pitskhelauri D.I. Artificial Intelligence Technologies in the Microsurgical Operating Room (Review) // Sovremennye tehnologii v medicine, Vol. 15, April 2023. P. 86.
28. Rai R., Tiwari M.K., Ivanov D., Dolgui A. Machine learning in manufacturing and industry 4.0 applications // International Journal of Production Research, Vol. 59, August 2021. pp. 4773–4778.
29. Hector I., Panjanathan R. Predictive maintenance in Industry 4.0: a survey of planning models and machine learning techniques // PeerJ Computer Science, Vol. 10, May 2024. P. e2016.
30. Hernndez J., Calvet N., Briceo C., Hartmann L., Berlind P. Spectral Analysis and Classification of Herbig Ae/Be Stars // The Astronomical Journal, Vol. 127, March 2004. pp. 1682–1701.
31. Sen S., Agarwal S., Chakraborty P., Singh K.P. Astronomical big data processing using machine learning: A comprehensive review // Experimental Astronomy, Vol. 53, January 2022. pp. 1–43.
32. Ulizko M., Artamonov A., Fomina J.,Antonov E., Tukumbetova R. Proceedings of the International Conference on Computer Graphics and Vision “Graphicon” // Clustering Thematic Information in Social Media. Ryazan. 2022. pp. 403-413.
33. Jordan M.I. , Mitchell T.M.. Machine learning: Trends, perspectives, and prospects // Science, Vol. 349, No. 6245, 2015.
34. Ulizko M.S., Antonov E.V., Artamonov A.A., Tukumbetova R.R. Visualization of graph-based representations for analyzing related multidimensional objects // Scientific Visualization. 2020. Vol. 12. No. 4. pp. 133-142.

35. Ulizko M. S., Antonov E. V., Grigorieva M. A., Tretyakov E.S., Tukumbetova R.R., ArtamonovA.A.. Visual analytics of twitter and social media dataflows: a casestudy of Covid-19 rumors // Scientific Visualization. 2021. Vol. 13. No. 4. pp. 144-163.
36. Ouyang W. Data Visualization in Big Data Analysis: Applications and Future Trends // Journal of Computer and Communications, Vol. 12, 2024. pp. 76–85.
37. Артамонов А.А., Ананьева А.Г., Третьяков Е.С., Оныкий Б.Н., Проничева Л.В., Ионкина К.В., Суслина А.С. Визуализация результатов экспресс-анализа ситуаций // Научная визуализация. 2016. Т. 8. С. 25-34.
38. Van Eck N. J.; Waltman L. Software survey: VOSviewer, a computer program for bibliometric mapping // Scientometrics, Vol. 84, No. 2, 2010.
39. Артамонов А.А., Галин И.Ю., Леонов Д.В., Михина Е.К., Оныкий Б.Н., Соколина К.А. Поисковые агентные технологии с многоязычным тезаурусом // Вестник Национального исследовательского ядерного университета "МИФИ". 2015. Т. 4. № 4. С. 369-376.
40. Chun-houh Chen, Wolfgang Hardle, Antony Unwin. Handbook of Data Visualization. Berlin: Springer-Verlag, 2008. 954 pp.
41. Эйхольц М.П. Программные средства визуализации данных о массовом распространении вирусов // Информационные технологии в науке и производстве : Материалы X Всероссийской молодежной научно-технической конференции, Омск, 18 апреля 2023 года. Омск. 2023. pp. 77-82.
42. Sievert C. Interactive Web-Based Data Visualization with R, plotly, and shiny. 1st ed. Chapman and Hall/CRC, 2020.
43. Gunawan R. Proceedings of the 2018 International Conference on Industrial Enterprise and System Engineering (IcoIESE 2018) // Comparison of Web Scraping Techniques : Regular Expression, HTML DOM and Xpath. 2019. pp. 283-287.

44. Zhang R., Meng Z., Wang H., Liu T., Wang G., Zheng L., Wang C. Hyperscale data analysis oriented optimization mechanisms for higher education management systems platforms with evolutionary intelligence // Applied Soft Computing, Vol. 155, April 2024. P. 111460.
45. Иванов Д.В., Петров И.Ю. Программное выделение данных из неструктурированных источников: обзор методов // Информационные технологии. 2023. Т. 7. С. 12-25.
46. Xu D., Chen W., Peng W., Zhang C., Xu T., Zhao X., Wu X., Zheng Y., Wang Y., Chen E. Large language models for generative information extraction:a survey // Frontiers of Computer Science. 2024. Vol. 18. No. 6. P. 186357.
47. Dagdelen J., Dunn A., Lee S., Walker N., Rosen A.S., Ceder G., Persson K., Jain A. Structured information extraction from scientific text with large language models // Nature Communications. 2024. Vol. 15.
48. Антонов Е.В., Ионкина К.В., Кондратько В.О., Смирнова Е.А., Артамонов А.А. База данных сгенерированных объектов информационно-учебного полигона для подготовки специалистов в сфере международных отношений, Свидетельство о государственной регистрации базы данных 2023621854, Jun 07, 2023.
49. Vicentiy A.V. The Geoimage Generation Method for Decision Support Systems Based on Natural Language Text Analysis // Lecture Notes in Networks and Systems. 2021. Vol. 230. pp. 609-619.
50. Vicentiy A.V., Shishaiv M.G. The Technology of Spatial Relations Visualization Based on the Analysis of Natural Language Texts // Lecture Notes in Networks and Systems. 2021. Vol. 232. pp. 971-980.
51. Пилецкий Б.М. Распознавание пространственных данных из естественно языковых текстов с целью визуализации // Труды Кольского научного центра РАН. Информационные технологии. 2021. Т. 12. № 4. С. 50-56.

52. Li S., Guo H., Tang X., Tang R., Hou L., Li R., Zhang R. Embedding Compression in Recommender Systems: A Survey // ACM Computing Surveys. 2024. Vol. 56. pp. 1-21.
53. Sokolov I.; Antonov E.; Artamonov A. Evaluation of Named Entity Recognition Software Packages for Data Mining // Physics of Particles and Nuclei. Vol. 55. No. 3. pp. 557-559.
54. Артамонов А.А., Тукумбетова Р.Р., Антонов Е.В., Сафиканов Д.И. Технологии и средства создания и ведения онтологий в информационно-аналитических системах (учебно-методическое пособие). Москва: НИЯУ МИФИ, 2023. 1-42 с.
55. Курнаев В.А., Оныкий Б.Н., Галин И.Ю., Соколина К.А., Курнаев А.А., Николаев В.С., Артамонов А.А., Баламутенко А.В., Леонов Д.В., Проничева Л.В., Фомина Ю.Е. Тезаурус по физике плазмы в международном стандарте TMX 1 4b specification, Свидетельство о государственной регистрации базы данных 2015620043, Jan 12, 2015.
56. Fernandes M. B. et al. Classification of neurologic outcomes from medical notes using natural language processing // Expert Systems with Applications, Vol. 214, 2023. P. 119171.
57. Москалев И.В., Кротова О.С., Хворова Л.А. Автоматизация процесса извлечения структурированных данных из неструктурированных медицинских выписок с применением технологий интеллектуального анализа текстов // Высокопроизводительные вычислительные системы и технологии, Т. 4, № 1, 2020. С. 163-167.
58. Schegoleva L., Burdin G. Chatbot for Applicants on University Admission Issues // Conference of Open Innovations Association, FRUCT., Vol. 29, Щсещиук 2021. pp. 491-494.

59. Гончаров А.С., Гончарова М.А. Роботизация обработки обращений граждан по вопросам социального и бытового обслуживания // Наука без границ, Т. 6, № 58, 2021. С. 40-52.
60. Abdullah M., Abujaber D., Al-Qarqaz A., Abbott R., Hadzikadic M. Combating propaganda texts using transfer learning // Int. J. of Artificial Intelligence. 2023. Vol. 12. No. 2. pp. 956-965.
61. Zelling H. Distributional Structure // WORD. 1954. Vol. 10. No. 2. pp. 146-162.
62. Mikolov T., Chen K., Corrado G., Dean J. Efficient Estimation of Word Representations in Vector Space // International Conference on Learning Representations. 2013.
63. Pennington J., Socher R., Manning C. Glove: Global Vectors for Word Representation // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP. 2014. pp. 1532-1543.
64. Devlin J., Lee K., Chang M.W., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). 2019. pp. 4171-4186.
65. Lee J., Yoon W., Kim S., Kim D., Kim S., Ho SO C., Kang J. BioBERT: a pre-trained biomedical language representation model for biomedical text mining // Bioinformatics. 202. Vol. 36. No. 4. pp. 1234-1240.
66. Yet Another Keyword Extractor (Yake) [Электронный ресурс] URL: https:// github.com/LIAAD/yake
67. Jun L.S., Siau K. A review of data mining techniques // Industrial Managment & Data Systems. 2001. Vol. 101. No. 1. pp. 41-46.
68. Wu X., Kumar V., Quinlan J.R., Ghosh J., Yang Q., Motoda H., McLachlan G., Ng A., Liu B., Yu P., et al. Top 10 algorithms in data mining // Knowledge and information systems. 2008. Vol. 14. pp. 1-37.

69. Seiferi J.W. Data mining: An overview // National security issues. 2004. pp. 201-217.

70. Artamonov A., Vasilev M., Tukumbetova R., Ulizko M. 2022 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence: The 13th Annual Meeting of the BICA Society // Multiagent System for Monitoring, Analysis and Classification of Data from Procurement Services. 2022. Vol. 213. pp. 96-100.

71. Ионкина К.В., Оныкий Б.Н., Артамонов А.А., Ананьева А.Г., Проничева Л.В., Галин И.Ю., Ушмаров И.А., Суслина И.В., Петровский В.Н., Быковский Д.П., Соколина К.А., Горяинова А.Е. Управляющая база данных для агентного поиска в интернет новостной информации по тематическому направлению "лазерные технологии", Свидетельство о государственной регистрации базы данных 2016620164 , Dec 04, 2016.

72. Жучкова, С. В. Автоматическое извлечение текстовых и числовых веб-данных для целей социальных наук // Социология: методология, методы, математическое моделирование., No. 50-51, 2020. pp. 141-183.

73. Барахнин В.Б., Кожемякина О. Ю., Мухамедиев Р. И., Борзилова Ю. С. Якунин К. О. Проектирование структуры программной системы обработки корпусов текстовых документов // Бизнес-информатика, № №4, 2019.

74. Петровский В.Н., Ионкина К.В., Артамонов А.А., Галин И.Ю., Третьяков Е.С., Оныкий Б.Н., Проничева Л.В., Данилова В.В., Кшняков Д.О., Суслина А.С., Черкасский А.И., Быковский Д.П. Интегральная база данных агентного поиска информации в сети интернет по тематическому направлению аддитивные технологии, Свидетельство о государственной регистрации базы данных 2017620244, Feb 27, 2017.

75. Chang C.H. A Survey of Web Information Extraction Systems // IEEE Transactions on Knowledge and Data Engineering, Т. 18, № 10, 2006. С. 1411-1428.

76. Zhao B. Web Scraping // In: Encyclopedia of big data. Springer International Publishing, 2017. pp. 1-3.
77. Chapagain A. Hands-On Web Scraping with Python: Perform advanced scraping operations using various Python libraries and tools such as Selenium, Regex, and others. Packt Publishing Ltd ed. 2019.
78. Gunawan R. et al. Comparison of web scraping techniques: regular expression, HTML DOM and Xpath // 2018 International Conference on Industrial Enterprise and System Engineering (ICoIESE 2018)., 2019. pp. 283-287.
79. Тукумбетова Р.Р., Артамонов А.А., Улизко М.С., Антонов Е.В., Ионкина К.В., Васильев М.И. Программа автоматизированного сбора тендеров по заданной тематике, Свидетельство о государственной регистрации программы для ЭВМ 2022664709 , Aug 03, 2022.
80. Ananieva A.G., Artamonov A.A., Galin I.U., Tretyakov E.S., Kshnyakov D.O. Algorithmization of search operations in multiagent information-analytical systems // Journal of Theoretical and Applied Information Technology. 2015. Vol. 81. pp. 11-17.
81. Артамонов А.А.; Галин И.Ю.; Ионкина К.В.; Курнаев В.А.; Соколина К.А.; Черкасский А.И. Тематические тезаурусы в агентных технологиях поиска научно-технической информации в интернете (на примере тезауруса по теме "физика плазмы") // Математическое моделирование. 2015. Т. 27. № 7. С. 4-9.
82. Бастрикина В.В. Проектирование веб-скрапера для получения данных с сайтов книжных издательств // Актуальные проблемы авиации и космонавтики., Т. №14, 2018.
83. Антонов Е.В., Артамонов А.А., Ионкина К.В., Кучинов В.П., Соколов И.Д., Тукумбетова Р.Р., Улизко М.С., Черкасский А.И. Программа бинарной классификации текстовой информации на основе современных

нейросетевых технологий , Свидетельство о государственной регистрации программы для ЭВМ 2024668072 , Aug 01, 2024.

84. Anish C. Hands-on web scraping with Python perform advanced scraping operations using various Python libraries and tools such as Selenium, Regex, and others // Packt Publishing Ltda. 2019.
85. Bourhis P., Reutter J., Suarez F., Vrgoc D. JSON: data model, query languages and schema specification // Proceedings of the 36th ACM SIGMOD-SIGACT-SIGAI symposium on principles of database systems. 2017. pp. 123-135.
86. Artifex [Электронный ресурс] // «Module fitz» 2015-2023: [сайт]. URL: https://pymupdf.readthedocs.io/en/latest/module.html. (дата обращения: 04.06.2025).
87. GitHub [Электронный ресурс] // «GROBID» 2008--2023: [сайт]. URL: https://github.com/kermitt2/grobid (дата обращения: 04.06.2025).
88. Requests: HTTP for Humans™ — Requests 2.32.2 documentation [Электронный ресурс] URL: https://requests.readthedocs.io/en/latest/ (дата обращения: 21.05.2024).
89. Yin S., Fu C., Zhao S., Li K., Sun X., Xu T., Chen E. A survey on multimodal large language models // National Science Review. 2024. Vol. 11.
90. Jiang P. et al. A Review of Yolo algorithm developments // Procedia Computer Science, Vol. 199, 2022. pp. 1066-1073.
91. Zamfir A.V., Carabas M., Carabas C., Tapus N. Systems monitoring and big data analysis using the elasticsearch system // IEEE, 2019. pp. 188-193.
92. Kelvin M. Factors influencing data saturation in qualitative studies» // International Journal of Research in Business and Social Science, , Vol. 11, No. 4, 2022. pp. 414-420.
93. Антонов Е. В., Артамонов А. А., Орлов А. В., Николаев В. С., Захаров В. П., Хохлова М. В., Концевая Ю. М., Бонарцев А. П., Воинова В. В. Обработка научно-технической информации в междисциплинарных

исследованиях методами математико-лингвистического направленного поиска на примере области изучения биоматериалов для тканевой инженерии // International Journal of Open Information Technologies, No. №11, 2022. pp. 134-140.

94. Третьяков Е.С., Ионкина К.В., Данилова В.В., Кшняков Д.О., Артамонов А.А., Оныкий Б.Н., Проничева Л.В., Суслина А.С., Суслина И.В., Толстая П.М. Программа автоматизированного квазиреферирования научно-технической информации, Свидетельство о государственной регистрации программы для ЭВМ 2017613798 , Mar 31, 2017.
95. Хвостова М.О., Антонов Е.В., Тукумбетова Р.Р., Соколов И.Д., Тремасов Г.М., Артамонов А.А., Матвеева А.Р., Андреев М.Н. Программа автоматизированного выделения значений и единиц измерения физических величин из полнотекстовых материалов, Свидетельство о государственной регистрации программы для ЭВМ 2024616882, Mar 26, 2024.
96. Fomina J., Safikanov D., Artamonov A., Tretyakov E. Postproceedings of the 10th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2019 (Tenth Annual Meeting of the BICA Society), held August 15-19, 2019 in Seattle, Washington, USA // Parametric and semantic analytical search indexes in hieroglyphic languages. Washington. 2020. Vol. 169. pp. 507-512.
97. OpenStreetMap [Электронный ресурс] // OpenStreetMap: [сайт]. [2023]. URL: https://www.openstreetmap.org. (дата обращения: 02.08.2025).
98. API поиска по организациям [Электронный ресурс] // Яндекс: [сайт]. [2023]. URL: https://yandex.ru/dev/maps/geosearch (дата обращения: 02.08.2025).
99. Pycountry [Электронный ресурс] // Github: [сайт]. [2025]. URL: https:// github.com/flyingcircusio/pycountry

100. Stonebraker M. SQL databases v. NoSQL databases // Communications of the ACM, Vol. 53, No. 4, 2010. pp. 10-11.

101. Artamonov A.A., Ionkina K.V., Kirichenko A.V., Lopatina O.L., Tretyakov E.S., Cherkasskiy A.I. Agent-based search in social networks International // Journal of Civil Engineering and Technology. 2018. Vol. 9. No. 13. pp. 28-35.

102. Тимонин А.Ю., Бождай А.С. Методы анализа гетерогенных данных для построения социального профиля // Russian journal of management, Т. 5, № 3, 2017. С. 481-489.

103. Оныкий Б.Н.; Артамонов А.А.; Третьяков Е.С.; Черкасский А.И.; Ионкина К.В. Индуктивные модели обучения поисковых агентов, работающих в социальных сетях // Системы высокой доступности. 2020. Т. 16. № 1. С. 5-13.

104. Cherkasskiy A.I., Cherkasskaya M.V., Artamonov A.A., Galin I.Y. User Group Classification Methods Based on Statistical Models // Studies in Computational Intelligence. 2022. Vol. 1032. pp. 69-74.

105. Figueroa C., Guillén V., Huenupán F., Vallejos C., Henríquez E., Urrutia F., Sanhueza F., Alarcón E. Comparison of Acoustic Parameters of Voice and Speech According to Vowel Type and Suicidal Risk in Adolescents // Journal of Voice, August 2024.

106. Taku K., Arai H. Roles of values in the risk factors of passive suicide ideation among young adults in the US and Japan // Frontiers in Psychology, Vol. 14, August 2023.

107. Haghish E.F., Nes R.B., Obaidi M., Qin P., Stänicke L.I., Bekkhus M., Laeng B., Czajkowski N. Unveiling Adolescent Suicidality: Holistic Analysis of Protective and Risk Factors Using Multiple Machine Learning Algorithms // Journal of Youth and Adolescence, Vol. 53, November 2023. pp. 507–525.

108. Воронкова Я.Ю., Радюк О.М., Басинская И.В. «Большая пятѐрка», или пятифакторная модель личности // Смысл, функции и значение разных

отраслей практической психологии в современном обществе: сборник научных трудов. 2017. С. 39-45.

109. Frydenberg E., Lewis R. Coping scale for adults (CSA-2): User manual. ACER Press, 2014.
110. Mota M.S.S., Ulguim H.B., Jansen K., Cardoso T.D.A., Souza L.D.D.M. Are big five personality traits associated to suicidal behaviour in adolescents? A systematic review and meta-analysis // Journal of Affective Disorders, Vol. 347, February 2024. pp. 115–123.
111. Shin S., Kim K. Prediction of suicidal ideation in children and adolescents using machine learning and deep learning algorithm: A case study in South Korea where suicide is the leading cause of death // Asian Journal of Psychiatry, Vol. 88, October 2023. P. 103725.
112. Wang H., Yuan H., Zhang Y., Wang Q., Gao Z., Zhao M. Suicide risk prediction for Korean adolescents based on machine learning // Scientific Reports, Vol. 15, April 2025.
113. Cohen J., Wright-Berryman J., Rohlfs L., Wright D., Campbell M., Gingrich D., Santel D., Pestian J. A Feasibility Study Using a Machine Learning Suicide Risk Prediction Model Based on Open-Ended Interview Language in Adolescent Therapy Sessions // International Journal of Environmental Research and Public Health, Vol. 17, November 2020. P. 8187.
114. Pizzoli S.F.M., Monzani D., Conti L., Ferraris G., Grasso R., Pravettoni G. Issues and opportunities of digital phenotyping: ecological momentary assessment and behavioral sensing in protecting the young from suicide // Frontiers in Psychology, Vol. 14, June 2023.
115. Fernandez-Fernandez J., Jiménez-Treviño L., Andreo-Jover J., Ayad-Ahmed W., Bascarán T.B., Canal-Rivero M., Cebria A., Crespo-Facorro B., De la Torre-Luque A., Diaz-Marsa M., et al. Network analysis of influential risk

factors in adolescent suicide attempters // Child and Adolescent Psychiatry and Mental Health, Vol. 18, November 2024.

116. Rashed A.E.E., Atwa A.E.M., Ahmed A., Badawy M., Elhosseini M.A., Bahgat W.M. Facial image analysis for automated suicide risk detection with deep neural networks // Artificial Intelligence Review, Vol. 57, September 2024.
117. Guo J.W., Kimmel J., Linder L.A. Text Analysis of Suicide Risk in Adolescents and Young Adults // Journal of the American Psychiatric Nurses Association, Vol. 30, February 2022. pp. 169–173.
118. Cañón Buitrago S.C., Pérez Agudelo J.M., Narváez Marín M., Montoya Hurtado O.L., Bermúdez Jaimes G.I. Predictive model of suicide risk in Colombian university students: quantitative analysis of associated factors // Frontiers in Psychiatry, Vol. 15, May 2024.
119. Cherkasskiy A., Artamonov A., Cherkasskaya M., Leonova N. 2020 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence: Eleventh Annual Meeting of the BICA Society // Methods for identifying an information object in social networks. 2021. Vol. 190. pp. 137-141.
120. Оныкий Б.Н., Черкасский А.И., Проничева Л.В., Суслина А.С., Артамонов А.А., Ионкина К.В., Третьяков Е.С., Инкина В.А., Лопатина Е.О. База данных маркированных аудио-объектов, Свидетельство о государственной регистрации базы данных 2018621999, Dec 11, 2018.
121. Сафиканов Д.И., Артамонов А.А., Фомина Ю.Е., Черкасский А.И. Статистическая модель поиска целевых объектов в социальной сети // International journal of open information technologies, Т. 12, № 10, 2024. С. 71-77.
122. Ulizko M.; Pronicheva L.; Artamonov A.; Tretyakov E. Brain-Inspired Cognitive Architectures for Artificial Intelligence: BICA*AI 2020 // Complex Objects Identification and Analysis Mechanisms. 2020. pp. 517-526.

123. Onykiy B.N.; Artamonov A.A.; Tretyakov E.S.; Ionkina K.V. Visualization of large samples of unstructured information on the basis of specialized thesauruses // Scientific Visualization. 2017. Т. 9. № 5. С. 54-58. 2017. Vol. 9. No. 5. pp. 54-58.

124. Antonov E.V., Artamonov A.A., Rudik A.V., Malugin M.I.. Trend Visualization of Academic Field: Proposed Method and Big Data Review // Scientific Visualization. 2022. Vol. 14. No. 2. pp. 62-76.

125. Ulizko M.S., Tukumbetova R.R., Artamonov A.A., Antonov E.V., Ionkina K.V. Biologically Inspired Cognitive Architectures 2023 // Data Preparation for Advanced Data Analysis on Elastic Stack. 2024. pp. 884-893.

126. Малугин М.И., Антонов Е.В., Артамонов А.А. Разработка системы для отслеживания публикационной активности научных организаций // Физика элементарных частиц и атомного ядра. 2024. Т. 55. № 3. С. 665.

127. Malugin M., Antonov E., Artamonov A. Designing a System for Monitoring the Publication Activity of the Scientific Organization // Physics of Particles and Nuclei. 2024. Vol. 55. No. 3. pp. 554-556.

128. Onykiy B.N., Antonov E.V., Artamonov A.A. Tretyakov E.S.. Information Analysis Support for Decision-Making in Scientific and Technological Development // International Journal of Technology.. 2020. Vol. 11. No. 6. pp. 1125-1135.

129. Tretyakov E.S., Tukumbetova R.R., Artamonov A.A. Methodology of analysis of similar objects with the use of modern visualization tools // Mechanisms and Machine Science. 2020. Vol. 80. pp. 113-119.

130. Норкина А.Н., Артамонов А.А., Морозов Н.В., Антонов Е.В., Улизко М.С., Ионкина К.В., Соколов И.Д., Мальцев М.В. Рецензированные учебно-методические материалы по финансовой безопасности по укрупненным группам специальностей, Свидетельство о государственной регистрации базы данных 2023621163, Apr 11, 2023.

131. Норкина А.Н., Артамонов А.А., Морозов Н.В., Антонов Е.В., Ионкина К.В., Тукумбетова Р.Р., Улизко М.С., Соколов И.Д. Программный комплекс формирования и экспертизы учебно-методических материалов, Свидетельство о государственной регистрации программы для ЭВМ 2023617975 , Apr 18, 2023.
132. Тремасов Г.М., Антонов Е.В., Артамонов А.А., Тукумбетова Р.Р., Хвостова М.О., Соколов И.Д., Патрушев К.А., Чупрыгин С.С. Программа автоматизированного сбора и обработки публикаций в области биомедицины, Свидетельство о государственной регистрации программы для ЭВМ 2024616345, Mar 19, 2024.
133. Inkina V.A., Antonov E.V., Artamonov A.A., Ionkina K.V., Tretyakov E.S., Cherkasskiy A.I. Proceedings of the 27th International Symposium Nuclear Electronics and Computing (NEC'2019) // Multiagent information technologies in system analysis. Budva, Becici, Montenegro. 2019. pp. 195-199.
134. Силаев Н.Ю., Артамонов А.А., Бондарев И.М., Аршба Б.Б., Гладышева А.И., Громяк И.В., Жабина Д.П., Лебедев Д.В., Таран В.Е., Тукумбетова Р.Р., Улизко М.С., Фомин М.Ю. База данных компаний незападных стран – крупнейших торговых партнеров России, Свидетельство о государственной регистрации базы данных 2024623588, Aug 16, 2024.
135. Антонов Е.В., Тукумбетова Р.Р., Чернов И.И., Михальчик В.В., Улизко М.С., Стальцов М.С., Артамонов А.А., Рудик А.В., Малугин М.И. База данных изображений и метаданных научных публикаций по облученным ядерным материалам за 2014-2018 года, Свидетельство о государственной регистрации базы данных 2022620590, Mar 18, 2022.
136. Масальский Л.С., Арефьева Д.Я. Теплотехника и информатика в образовании, науке и производстве : сборник докладов XI Всероссийской научно-практической конференции студентов, аспирантов и молодых учёных (ТИМ'2023) с международным участием (Екатеринбург, 18–19

мая 2023 г.) // Применение научных визуализаций для анализа вариативности и динамики системы. Екатеринбург. 2023. С. 201-205.

137. Ulizko M., Antonov E., Artamonov A., Tukumbetova R. Graph Visualization of the Characteristics of Complex Objects on the Example of the Analysis of Politicians // CEUR Workshop Proceedings - 2020, Vol. 2744. 2020. Vol. 2744. pp. short8-1 — short8-9.
138. Ulizko M., Tretyakov E., Tukumbetova R., Artamonov A., Esaulov M. Visualization of Dataflows: a Casestudy of COVID-19 Rumors // CEUR Workshop Proceedings. 2021. Vol. 3027. pp. 259-267.
139. Wang Y., Qian Y., Qi X., Cao N., Wang D. InnovationInsights: A Visual Analytics Approach for Understanding the Dual Frontiers of Science and Technology // IEEE Transactions on Visualization and Computer Graphics. Aug 2023.
140. Артамонов А.А., Леонов Д.В., Николаев В.С., Оныкий Б.Н., Проничева Л.В., Соколина К.А., Ушмаров И.А. Визуализация семантических отношений в мультиагентных системах // Научная визуализация. 2014. Т. 6. № 3. С. 68-76.
141. Das R., Soyiu M. A key review on graph data science: The power of graphs in scientific studies 2023. Vol. 249.
142. Zhao J., Collins C., Chevalier F., Balakrishnan R. Interactive Exploration of Implicit and Explicit Relations in Faceted Datasets // IEEE Transactions on Visualization and Computer Graphics. 2013. Vol. 19. No. 12. pp. 2080-2089.
143. Filipov V., Arleo A., Miksch S. Are We There Yet? A Roadmap of Network Visualization from Surveys to Task Taxonomies // Computer Graphics Forum. Apr 2023. Vol. 42. No. 6.
144. Korovin D.I., Romanova E.V., Muminova S.R., Osipov A.V., Pleshakova E.S., Mazutskiy N.M., Gataullin T.M., Gataullin S.T. Graph analytics for digital economy tasks // ИТиВС. 2023. Vol. 3. pp. 33-45.

145. Бреслер М.Г. Сетевой принцип формирования элит в информационном/цифровом обществе // Вопросы элитологии. 2024. Т. 2. С. 91-111.

146. Patel L., Shuler K. Conference: Platform for Advanced Scientific Computing // DynaHGraph: Learning Hidden Relationships in Dynamic Graphs. Zurich, Switzerland. 2024.

147. Артамонов А.А.; Тукумбетова Р.Р.; Ионкина К.В.; Улизко М.С. Современные технологии и средства построения графа знаний (учебно-методическое пособие). Москва: НИЯУ МИФИ, 2023. 1-44 с.

148. Ulizko M.S.; Artamonov A.A.;Tukumbetova R.R.; Antonov E.V.; Vasilev M.I. Critical Paths of Information Dissemination in Networks // Scientific Visualization. 2022. Vol. 14. No. 2. pp. 98-107.

149. Козицын А.С. Нахождение скрытых зависимостей между объектами на основе анализа больших массивов библиографических данных // Материалы конференции VI Международная конференция Актуальные проблемы системной и программной инженерии (АПСПИ 2019). Москва. 2019.

150. Georgi M. Methods for Mining Causality from Observations in Artificial Intelligence // Izvestiya SFedU. Engineering Sciences, Vol. 3, No. 192, 2023. pp. 125-134.

151. Батищев С.В., Искварина Т.В., Скобелев П.О. Методы и средства построения онтологий для интеллектуализации сети интернет // Известия Самарского научного центра РАН , Vol. 4, No. 1, 2002. pp. 91-103.

152. Батура Т.В. Методы и системы семантического анализа текстов // Международный журнал Программные продукты и системы, Vol. 12, 2016.

153. Что такое GUI? Значение аббревиатуры GUI [Электронный ресурс] URL: https://animatika.ru/info/gloss/GUI.html (дата обращения: 20.05.2024).

154. Three.js – JavaScript 3D Library [Электронный ресурс] URL: https://threejs.org/ (дата обращения: 20.05.2024).

155. MUI Core: Ready-to-use React components, free forever [Электронный ресурс] URL: https://mui.com/core/ (дата обращения: 21.05.2024).

156. Мурзин Д. Г., Ярова А. В., Мурзин В. М. Библиотека для создания пользовательских интерфейсов React.js // Современные проблемы радиоэлектроники и телекоммуникаций, Т. 4, 2021. С. 201.

157. FastAPI [Электронный ресурс] URL: https://fastapi.tiangolo.com/ (дата обращения: 21.05.2024).

158. Camacho D., Panizo Lledot Á., Bello Orgaz G., Gonzalez-Pardo A., Cambria E. The Four Dimensions of Social Network Analysis: An Overview of Research Methods, Applications, and Software Tools. Singapore: School of Computer Science and Engineering Nanyang Technological University, 2020.

159. Query DSL | Elasticsearch Guide [8.13] | Elastic [Электронный ресурс] URL: https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl.html (дата обращения: 21.05.2024).

160. Круглик Р.И. Создание веб-приложения с помощью библиотеки React.js // Постулат, Т. 1-1, № 39, 2019. С. 125.

161. Adamczewski J., Becker K.H., Belogurov S., Boldyreva N., Chernogorov A., Deveaux C., Dobyrn V., Dürr M., Eom J., Eschke J., et al. Event reconstruction in the RICH detector of the CBM experiment at FAIR // Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment, Vol. 766, 2014. pp. 250-254.

162. Bastrukov S., Yang J., Lai P.Y. Simulations of Wave Motions in Magnetically Polarized Gas-Dust Interstellar Media // Journal of Computational Methods in Sciences and Engineering, Vol. 2, No. 1-2, 2002. pp. 13-20.

163. Chen C. Science Mapping: A Systematic Review of the Literature // Journal of Data and Information Science, Vol. 2, No. 2, 2017.

164. Börner K. Atlas of Knowledge: Anyone Can Map. MIT Press, 2015.

165. Keim D.; Kohlhammer J.; Ellis G.; Mansmann F. Visual Analytics: Definition, Process, and Challenges // Information Visualization, Vol. 8, No. 4, 2008.

166. Yang Q., Zheng X., Zhong F., Chen L., Hong J., Liu X., Jiang J. Scientific landscape and visualization analysis of the link between adenomyosis and infertility from 2000 to 2024 // Front Med (Lausanne). Jan 2025. pp. 2:1488866.

167. Jamshidi S., Hashemi S. The Scientific Landscape of the Aging-in-Place Literature: A Bibliometric Analysis // Journal of Ageing and Longevity. 2024. Vol. 4. pp. 417-432.

168. Litvinova O., Mickael M.E., Gerger G., Yeung A.W.K., Fatimi A., Haick H., Atanasov A.G., Willschke H. Patent and bibliometric analysis of the scientific landscape of the use of graphene-based biosensors and their prospects in digital medicine // World Patent Information. 2025.

169. Kim H., Kim S.H., Kim J., Kim E.H., Gu J.H., Lee D. A keyword-based approach to analyzing scientific research trends: ReRAM present and future // Scientific Reports. 2025. Vol. 15.

170. Small H. Co-citation in the scientific literature: A new measure of the relationship between publications // Journal of the American Society for Information Science, Vol. 24, No. 4, 1973.

171. Рудик А.В., Антонов Е.В., Артамонов А.А. 33-я Международная конференция по компьютерной графике и машинному зрению // Инструменты оценки научно-технологического ландшафта страны. Москва. 2023. С. 256-265.

172. Artamonov A.A.,Kshnyakov D.O., Danilova V.V., Cherkasskiy A.I., Galin I.Y. 8th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2017 (Eighth Annual Meeting of the BICA Society), held August 1-6, 2017 in Moscow, Russia // Methodology for the Development of

Dictionaries for Automated Classification System. Moscow. 2018. Vol. 123. pp. 57-62.

173. Artamonov A.A., Ionkina K.V., Tretyakov E.S., Timofeev A.I. Postproceedings of the 9th Annual International Conference on Biologically Inspired Cognitive Architectures, BICA 2018 (Ninth Annual Meeting of the BICA Society), held August 22-24, 2018 in Prague, Czech Republic // Electronic document processing operating map development for the implementation of the data management system in a scientific organization. Prague, Czech Republic. 2018. Vol. 145. pp. 248-253.
174. Оныкий Б.Н., Артамонов А.А., Ионкина К.В., Третьяков Е.С., Свистунов А.С., Лопатина Е.О., Зиновьева М.Г., Черкасский А.И., Проничева Л.В., Суслина А.С., Иванченко А.М. Конвейерное определение тематической принадлежности научно-технической документации, Свидетельство о государственной регистрации программы для ЭВМ 2018665872 , Dec 11, 2018.
175. Артамонов А.А., Стальцов М.С., Антонов Е.В., Чернов И.И., Улизко М.С., Тукумбетова Р.Р., Ионкина К.В. Программа выборки данных по свойствам и структурам облученных реакторных материалов , Свидетельство о государственной регистрации программы для ЭВМ 2022669767 , Oct 25, 2022.
176. Тукумбетова Р.Р., Улизко М.С., Коренькова Т.В., Артамонов А.А. Сравнение методов классификации данных в машинном обучении на примере научных публикаций по ядерному топливному циклу // Системы высокой доступности. 2025. Т. 21. № 1. С. 25-38.
177. Артамонов А.А., Тукумбетова Р.Р., Соколов И.Д., Зрелова Д.П., Коренькова Т.В., Хвостова М.О., Вуйкович А.Д., Антонов Е.В., Улизко М.С., Черкасский А.И. Аналитический фреймворк обработки и представления научно-технической информации, Свидетельство о

государственной регистрации программы для ЭВМ 2024690979 , Dec 04, 2024.

178. Николаев В.С., Артамонов А.А., Улизко М.С., Антонов Е.В., Кателевский Д.Н. Мультиагентная система сбора; обработки и анализа слабоструктурированных данных, Свидетельство о государственной регистрации программы для ЭВМ 2024690021, Dec 11, 2024.

ПРИЛОЖЕНИЕ А АКТЫ ВНЕДРЕНИЯ РЕЗУЛЬТАТОВ ДИССЕРТАЦИОННОЙ РАБОТЫ

Общество с ограниченной ответственностью
«СИСТЕМЫ ИНФОРМАЦИОННОЙ АНАЛИТИКИ»

«УТВЕРЖДАЮ»
Генеральный директор
Николаев В.С.
«09» 09 2025 г.

АКТ

внедрения в СИА результатов диссертационной работы Артамонова Алексея Анатольевича на соискание ученой степени доктора технических наук по теме «Модели, методы и технологии интеллектуального анализа информационных объектов в научно-технических и социально значимых задачах»
по специальности 2.3.1. «Системный анализ, управление и обработка информации, статистика»

Настоящий акт подтверждает, что результаты диссертационной работы Артамонова Алексея Анатольевича на тему «Модели, методы и технологии интеллектуального анализа информационных объектов в научно-технических и социально значимых задачах» использованы в деятельности по разработке системы Общества с ограниченной ответственностью «Системы информационной аналитики» (СИА).

Комиссия в составе:

В.С. Николаев – председатель комиссии, генеральный директор;

М.С. Улизко – член комиссии, разработчик программного обеспечения;

Д.Н. Кателевский – член комиссии, системный аналитик

рассмотрела материалы диссертационной работы Артамонова А.А., представленной на соискание ученой степени доктора технических наук и отмечает, что в ходе разработки системы СИА.Атташе по договору с Фондом содействия инновациям № 349ГС1ЦТС10-D5/80243 от 12.12.2022, а также в ходе использования разработанного программного комплекса для выполнения информационно-аналитических работ использованы, предложенные автором модели, методы и технологии интеллектуального анализа научно-технической информации.

Предложенные автором модели, методы и механизмы визуализации внесли определяющий вклад в разработку программный комплекс интеллектуального анализа

научно-технической информации, обеспечивающей горизонтальную масштабируемость системы для сбора, обработки и хранения разнотиповых входящих данных.

Реализованы специализированные методы насыщения данных научно-технической информации, что способствует значительному улучшению качества последующего анализа и выполнения аналитических работ. С использованием предложенных Артамоновым А.А. методов и технологий интеллектуального анализа научно-технической информации выполнен ряд заказных работ в том числе по договорам № 2024-sia-dgk-1 от 15.04.2024 и № 2024-sia-dgk-2 от 15.05.2024.

Председатель комиссии:

Генеральный директор  В.С. Николаев

Члены комиссии:

Разработчик программного обеспечения  М.С. Улизко

Системный аналитик  Д.Н. Кателевский

МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ РОССИЙСКОЙ ФЕДЕРАЦИИ

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ЯДЕРНЫЙ УНИВЕРСИТЕТ «МИФИ» (НИЯУ МИФИ)

«УТВЕРЖДАЮ»
Первый проректор НИЯУ МИФИ
О.В. Нагорнов

«28» 08 2025 г.

АКТ

внедрения в НИЯУ МИФИ результатов диссертационной работы Артамонова Алексея Анатольевича на соискание ученой степени доктора технических наук по теме «Модели, методы и технологии интеллектуального анализа информационных объектов в научно-технических и социально значимых задачах»
по специальности 2.3.1. «Системный анализ, управление и обработка информации, статистика»

Настоящий акт подтверждает, что результаты диссертационной работы Артамонова Алексея Анатольевича на тему «Модели, методы и технологии интеллектуального анализа информационных объектов в научно-технических и социально значимых задачах» использованы в деятельности Исследовательского центра в сфере искусственного интеллекта по направлению «Транспорт и логистика» НИЯУ МИФИ (ИЦИИ НИЯУ МИФИ).

Комиссия в составе:

А.Н. Петровский – директор ИЦИИ НИЯУ МИФИ;

Е.Н. Бажанова – ведущий программист ИЦИИ НИЯУ МИФИ;

А.И. Черкасский – ведущий программист ИЦИИ НИЯУ МИФИ

рассмотрела материалы диссертационной работы Артамонова А.А., представленной на соискание ученой степени доктора технических наук и отмечает:

1. Предложенная автором обобщённая модель комплексного цифрового информационного объекта, основанная на статических, динамических, вычисляемых характеристиках, а также учитывающая связи между объектами позволила спроектировать и реализовать информационно-аналитическую систему агрегирующую в себе научно-

техническую информацию и нормативно-правовые документы по исследованиям, разработкам по технологиям искусственного интеллекта использующимися в транспорте и логистике.

2. Внедренные оригинальные методы автоматизированного извлечения и насыщения данных, предложенные Артамоновым А.А., адаптированы для работы с научно-технической информацией. К ним относятся методы распознавания и стандартизации физических величин, а также извлечения и структурирования данных из таблиц и подписей к рисункам. За счет адаптации технологий NLP, OCR и геокодирования данные методы позволяют достичь глубокого обогащения данных и выявления скрытых зависимостей, что недостижимо при стандартной обработке.

В рамках выполнения работ по договору от 27.12.2023г. № 70-2023-001309 на предоставление гранта в сфере искусственного интеллекта по направлению «Транспорт и логистика» лично Артамонов Алексей Анатольевич внес определяющий вклад в разработку аналитического фреймворка по сбору и насыщению данных из открытых информационных источников (Свидетельство о государственной регистрации программы для ЭВМ «Аналитический фреймворк обработки и представления научно-технической информации» №2024690979, 12 декабря 2024).

Члены комиссии:

Директор ИЦИИ НИЯУ МИФИ	А.Н. Петровский
Ведущий программист ИЦИИ НИЯУ МИФИ	Е.Н. Бажанова
Ведущий программист ИЦИИ НИЯУ МИФИ	А.И. Черкасский

Акционерное общество
«Научно-исследовательский центр
«Прикладная Логистика»
(АО НИЦ «Прикладная Логистика»)
ИНН 7725792300 КПП 770501001
115114, Москва, ул. Летниковская, д.10, стр. 4
тел.: +7 (495) 181-51-71
e-mail: info_pl@cals.ru

АКТ

практического использования в АО НИЦ «Прикладная логистика» результатов диссертационной работы Артамонова Алексея Анатольевича на соискание ученой степени доктора технических наук по теме «Модели, методы и технологии интеллектуального анализа информационных объектов в научно-технических и социально значимых задачах»
по специальности 2.3.1. «Системный анализ, управление и обработка информации, статистика»

Настоящим актом подтверждается, что результаты диссертационной работы Артамонова Алексея Анатольевича на тему «Модели, методы и технологии интеллектуального анализа информационных объектов в научно-технических и социально значимых задачах» были использованы в деятельности АО НИЦ «Прикладная логистика».

В рамках выполнения коммерческих договоров специалистами АО НИЦ «Прикладная логистика» были подготовлены отчёты о патентных исследованиях в области разработки средств подготовки, выпуска и сопровождения электронной конструкторской документации для машиностроительных изделий. Данная работа выполнялась с применением программного комплекса «СИА.Атташе».

Предложенные в диссертационной работе Артамонова А. А. модели, методы и механизмы визуализации результатов анализа данных определили состав, полноту и корректность сформированных отчётов о патентных исследованиях. Применение программного комплекса «СИА.Атташе» обеспечило горизонтальную масштабируемость процессов сбора, обработки и хранения разнотипных исходных данных по целевым научным и техническим направлениям.

Использование программного комплекса «СИА.Атташе» в деятельности АО НИЦ «Прикладная логистика» позволило проводить оперативный анализ патентной и публикационной информации из мировых информационных ресурсов (Lens.org, Google Patents, ScienceDirect, Springer) на различных языках и сократить затраты при выполнении коммерческих работ организации.

Генеральный директор Галин Илья Юрьевич

ПРИЛОЖЕНИЕ Б ПРИМЕР СТРУКТУРЫ JSON ФАЙЛА С ДАННЫМИ ПО НАУЧНОЙ ПУБЛИКАЦИИ «A SURVEY ON MULTIMODAL LARGE LANGUAGE MODELS»


```
{
    "title": "A survey on multimodal large language models",
    "published": {
      "journal": "National Science Review",
      "publisher": "Oxford University Press (OUP)",
      "volume": "11",
      "issue": "12",
      "pub_place": null,
      "page": {
        "start": null,
        "end": null
      },
      "eISSN": "2053-714X",
      "ISSN": "2095-5138",
      "year": 2024,
      "month": 11,
      "day": 12,
      "doi": "10.1093/nsr/nwae403",
      "submission_info": "Received 22 May 2024; Revised 11 October 2024;
Accepted 15 October 2024"
    },
    "authors": [
      {
        "name": "Shukang Yin",
        "email": null,
        "ids": {},
        "aff_keys": [1]
      },
      {
        "name": "Chaoyou Fu",
        "email": null,
        "ids": {},
        "aff_keys": [2,3]
      },
      {
        "name": "Sirui Zhao",
        "email": null,
        "ids": {},
        "aff_keys": [1]
      },
      {
        "name": "Ke Li",
        "email": null,
        "ids": {},
        "aff_keys": [4]
      },
      {
        "name": "Xing Sun",
        "email": null,
        "ids": {},
        "aff_keys": [4]
      },
      {
        "name": "Tong Xu",
        "email": null,
        "ids": {},
        "aff_keys": [1]
```

```
      },
      {
        "name": "Enhong Chen",
        "email": "cheneh@ustc.edu.cn",
        "ids": {},
        "aff_keys": [1]
      }
    ],
    "affiliations": [
      {
        "key": 1,
        "src": "School of Artificial Intelligence and  Data Science,
University of Science and Technology of China, Hefei 230026, China",
        "name": "University of Science and Technology of China",
        "department": "School of Artificial Intelligence and Data Science",
        "address": {
          "post_code": "230026",
          "settlement": "Hefei",
          "region": null,
          "country": "China"
        },
        "location": null
      },
      {
        "key": 2,
        "src": "State Key Laboratory for Novel Software Technology, Nanjing
University, Nanjing 210023, China",
        "name": "Nanjing University",
        "department": null,
        "address": {
          "post_code": "210023",
          "settlement": "Nanjing",
          "region": null,
          "country": "China"
        },
        "location": null
      },
      {
        "key": 3,
        "src": "School of Intelligence Science and Technology, Nanjing
University, Suzhou 215163, China",
        "name": "Nanjing University",
        "department": "School of Intelligence Science and Technology",
        "address": {
          "post_code": "215163",
          "settlement": "Suzhou",
          "region": null,
          "country": "China"
        },
        "location": null
      },
      {
        "key": 4,
        "src": "Tencent YouTu Lab, Shanghai 200233, China",
        "name": "Tencent YouTu Lab",
        "department": null,
        "address": {
          "post_code": "200233",
          "settlement": "Shanghai",
          "region": null,
          "country": "China"
        },
        "location": null
```

```
      },

    ],
    "countries": ["China"],
    "countries_code": ["CN"],
    "abstract": [
      "Recently, the multimodal large language model (MLLM) represented by
GPT-4V has been a new rising research hotspot, which uses powerful large
language models (LLMs) as a brain to perform multimodal tasks.",
      "The surprising emergent capabilities of the MLLM, such as writing
stories based on images and optical character recognition-free math
reasoning, are rare in traditional multimodal methods, suggesting a
potential path to artificial general intelligence.",
      "To this end, both academia and industry have endeavored to develop
MLLMs that can compete with or even outperform GPT-4V, pushing the limit of
research at a surprising speed.",
      "In this paper, we aim to trace and summarize the recent progress of
MLLMs.",
      "First, we present the basic formulation of the MLLM and delineate its
related concepts, including architecture, training strategy and data, as
well as evaluation.",
      "Then, we introduce research topics about how MLLMs can be extended to
support more granularity, modalities, languages and scenarios.",
      "We continue with multimodal hallucination and extended techniques,
including multimodal in-context learning, multimodal chain of thought and
LLM-aided visual reasoning.",
      "To conclude the paper, we discuss existing challenges and point out
promising research directions."
    ],
    "keywords": [
      "multimodal large language model",
      "vision language model",
      "large language model"
    ],
    "full_text": {
      "plain": [
        {
          "title": "INTRODUCTION",
          "content": [
            [
              "Recent years have seen remarkable progress in large language
models (LLMs) [1 ,2 ].",
              "By scaling up data size and model size, these LLMs raise
extraordinary emergent abilities, typically including instruction following
[3 ], in-context learning (ICL) [4 ] and chain of thought (CoT) [5 ].",
              "Although LLMs have demonstrated surprising zero/few-shot
reasoning performance on most natural language processing (NLP) tasks [6 ]
and even complex reallife applications [7 -9 ], they are inherently 'blind'
to vision since they can only understand discrete text.",
              "At the same time, large vision models (LVMs) can see clearly
[10 ,11 ], but commonly lag in reasoning."
            ],
            [
              "In light of this complementarity, an LLM and LVM run towards
each other, leading to the new field of the multimodal large language model
(MLLM).",
              "Formally, it refers to the LLM-based model with the ability
to receive, reason and output with multimodal information.",
              "Prior to the MLLM, there have been a lot of works devoted to
multimodality, which can be divided into discriminative [12 ,13 ] and
generative [14 ,15 ] paradigms.",
              "Contrastive language-image pretraining (CLIP) [12 ], as a
representative of the former, projects visual and textual information into a
```

```
unified representation space, building a bridge for downstream multimodal
tasks.",
              "In contrast, one for all (OFA) [14 ] is a representative of
the latter, which unifies multimodal tasks in a sequence-to-sequence
manner.",
              "The MLLM can be classified as the latter according to the
sequence operation, but it manifests two distinct traits compared with its
traditional counterparts.",
              "(i) The MLLM is based on an LLM with bi l lionscale
parameters, which is not available in previous models.",
              "(ii) The MLLM uses new training paradigms to unleash its full
potential, such as using multimodal instruction tuning [16 ] to encourage
the model to follow new instructions.",
              "Armed with the two traits, the MLLM exhibits new
capabilities, such as writing website code based on images [17 ],
understanding the deep meaning of a meme [18 ] and optical character
recognition-(OCR) free math reasoning [19 ].",
              "Ever since the release of GPT-4 [20 ], there has been a
research frenzy over MLLMs because of the amazing multimodal examples it
shows.",
              "Rapid development is fueled by efforts from both academia and
industry.",
              "Preliminary research on MLLMs focuses on text content
generation grounded in text prompts and image [16 ]/video [21 ,22 ]/audio
[23 ].",
              "Subsequent works have expanded the capabilities or the usage
scenarios, including:"
            ],
            [
              "(i) better granularity support-finer control on user prompts
is developed to support specifying regions through boxes [24 ] or a certain
object through a click [25 ]; (ii) enhanced support on input and output
modalities [26 ,27 ], such as image, video, audio and point cloud; (iii)
improved language suppor t-effor ts have been made to extend the success of
MLLMs to other languages (e.g.",
              "Chinese) with relatively limited training corpus [28 ]; (iv)
extension to more realms and usage scenarios-some studies transfer the
strong capabilities of MLLMs to other domains, such as medical image
understanding [29 ] and document parsing [30 ]."
            ],
            [
              "Moreover, multimodal agents are developed to assist in real-
world interaction, e.g.",
              "embodied agents [31 ] and graphical user interface (GUI)
agents [32 ].",
              "An MLLM timeline is i l lustrated in Fig. 1 ."
            ],
            [
              "In view of such rapid progress and the promising results of
this field, we have written this survey to provide researchers with a grasp
of the basic idea, main method and current progress in MLLMs.",
              "Note that we mainly focus on visual and language modalities,
but also include works involving other modalities like video and audio.",
              "Specifically, we cover the most important aspects of MLLMs
with corresponding summaries and have opened a GitHub page that wi l l be
updated in real time.",
              "To the best of our knowledge, this is the first survey on the
MLLM."
            ],
            [
              "The survey is structured as follows.",
              "We start with a comprehensive review of the essential aspects
of MLLMs, including the mainstream architecture, a full recipe for the
```

```
training strategy and data, and common practices for performance
evaluation.",
              "Then, we delve into a deeper discussion of some important
topics about MLLMs, each focusing on one of the following main problems.",
              "(i) What aspects can be further improved or extended?",
              "(ii) How can we relieve the multimodal hallucination issue?",
              "The survey continues with the introduction of three key
techniques, each specialized in a specific scenario.",
              "Multimodal in-context learning is an effective technique
commonly used at the inference stage to boost few-shot performance.",
              "Another important technique is multimodal chain of thought,
which is typically used in complex reasoning tasks.",
              "Afterward, we delineate general ideas for developing LLM-
based systems to solve composite reasoning tasks or to address common user
queries.",
              "We conclude our survey with a summary and potential research
directions.",
              "An illustration of typical MLLM architecture.",
              "It includes an encoder, a connector and an LLM.",
              "An optional generator can be attached to the LLM to generate
more modalities besides text.",
              "The encoder takes in images, audios or videos and outputs
features, which are processed by the connector so that the LLM can better
understand.",
              "There are broadly three types of connector: projection-based,
query-based and fusionbased connectors.",
              "The former two types adopt token-level fusion, processing
features into tokens to be sent along with text tokens, while the last type
enables a feature-level fusion inside the LLM."
            ]
          ]
        },
        {
          "title": "ARCHITECTURE",
          "content": [
            [
              "A typical MLLM can be abstracted into three modules: a pre-
trained modality encoder, a pre-trained LLM and a modality interface to
connect them.",
              "Drawing an analogy to humans, modality encoders such as
image/audio encoders are human eyes/ears that receive and pre-process
optical/acoustic signals, while LLMs are like human brains that understand
and reason with the processed signals.",
              "In between, the modality interface serves to align different
modalities.",
              "Some MLLMs also include a generator to output other
modalities apart from text.",
              "A diagram of the architecture is plotted in Fig. 2 .",
              "In this section, we introduce each module in sequence."
            ]
          ]
        },
        {
          "title": "Modality encoder",
          "content": [
            [
              "The encoders compress raw information, such as images or
audio, into a more compact representation.",
              "Rather than training from scratch, a common approach is to
use a pre-trained encoder that has been aligned to other modalities.",
              "For example, CLIP [12 ] incorporates a visual encoder
semantically aligned with the text through large-scale pre-training on
image-text pairs.",
```

```
              "Therefore, it is more practical to utilize such pre-aligned
encoders to align with LLMs through alignment pre-training."
            ],
            [
              "Commonly used image encoders are summarized in Table 1 .",
              "Apart from vani l la CLIP image encoders [12 ], some works
also explore using other variants.",
              "For example, MiniGPT-4 [17 ] adopts an EVA-CLIP [36 ] (ViT-
G/14) encoder, which is trained with improved training techniques.",
              "Osprey [25 ] introduces a convolution-based ConvNext-L
encoder [33 ] to utilize higher resolution and multi-level features.",
              "Some works also explore an encoder-free architecture.",
              "For instance, the image patches of Fuyu-8b [37 ] are directly
projected before sending to LLMs.",
              "With this design, the model naturally supports flexible input
image resolution."
            ],
            [
              "When choosing encoders, one often considers factors such as
resolution, parameter size and pretraining corpus.",
              "Notably, many works have empirically verified that using
higher resolution can achieve remarkable performance gains [28 ,38 ].",
              "The approaches for scaling up input resolution can be
categorized into direct scaling and patch-division methods.",
              "The direct scaling method inputs images of higher resolutions
to the encoder, which often involves further tuning the encoder [28 ] or
replacing a pre-trained encoder with higher resolution [39 ].",
              "Similarly, CogAgent [32 ] uses a dual-encoder mechanism,
where two encoders process high-and lowresolution images, respectively.",
              "High-resolution features are injected into the low-resolution
branch through cross-attention.",
              "Patch-division methods cut a high-resolution image into
patches and reuse the low-resolution encoder.",
              "For example, Monkey [38 ] and SPHINX [40 ] divide a large
image into smaller patches and send sub-images together with a downsampled
high-resolution image to the image encoder, where the sub-images and the
low-resolution image capture local and global features, respectively.",
              "In contrast, parameter size and training data composition are
of less importance compared with input resolution, as found by empirical
studies [41 ]."
            ],
            [
              "Similar encoders are also available for other modalities.",
              "For example, Pengi [23 ] uses the CLAP [42 ] model as the
audio encoder.",
              "ImageBind-LLM [26 ] uses the ImageBind [43 ] encoder, which
supports encoding image, text, audio, depth, thermal and inertial
measurement unit data.",
              "Equipped with the strong encoder, the ImageBind-LLM can
respond to the input of various modalities."
            ]
          ]
        },
        {
          "title": "Pre-trained LLM",
          "content": [
            [
              "Instead of training an LLM from scratch, it is more efficient
and practical to start with a pre-trained one.",
              "Through tremendous pre-training on the web corpus, LLMs have
been embedded with rich world knowledge, and demonstrate strong
generalization and reasoning capabilities.",
              "[12 ] OpenAI's WIT 224/336 13 304.0",
```

```
              "EVA-CLIP-ViT-G/14 [34 ] LAION-2B,COYO-700M 224 11 10 0 0.0
OpenCLIP-ViT-G/14 [33 ] LAION-2B 224 34 1012.7 OpenCLIP-ViT-bigG/14 [33 ]
LAION-2B 224 34 1844.9",
              "InternViT-6B [35 ] Multiple datasets 448 -5540.0",
              "We summarize the commonly used and publicly available LLMs in
Table 2 .",
              "Notably, most LLMs fall in the causal decoder category,
following GPT-3 [4 ].",
              "Among them, Flan-T5 [44 ] series are relatively early LLMs
used in works like BLIP-2 [50 ] and Instruct-BLIP [51 ].",
              "LLaMA series [45 ] and the Vicuna family [46 ] are
representative open-sourced LLMs that have attracted much academic
attention.",
              "Since the two LLMs are mainly pre-trained on the English
corpus, they are limited in multi-language support, such as Chinese.",
              "In contrast, Qwen [48 ] is a bilingual LLM with Chinese and
English support."
            ],
            [
              "It should be noted that scaling up the parameter size of LLMs
also brings additional gains, similar to the case of increasing input
resolution.",
              "Specifically, Liu et al. [39 ,52 ] found that simply scaling
up the LLM from 7B to 13B brings comprehensive improvement on various
benchmarks.",
              "Furthermore, when using a 34B LLM, the model shows emergent
zero-shot Chinese capability, given that only English multimodal data are
used during training.",
              "Lu et al. [53 ] observed a similar phenomenon by scaling up
LLMs from 13B to 35B and 65B/70B, where the larger model size brings
consistent gains on benchmarks specifically designed for MLLMs.",
              "Some works instead use smal ler LLMs to faci litate
deployment on mobile devices.",
              "For example, MobileVLM series [54 ] use downscaled LLaMA [45
] to enable efficient inference on mobile processors."
            ],
            [
              "Recently, explorations of the mixture-of-experts (MoE)
architecture for LLMs have garnered rising attention [55 ].",
              "Compared with dense models, the sparse architecture enables
scaling up the total parameter size without increasing the computational
cost, by selective activation of the parameters.",
              "Empirically, MM1 [41 ] and MoE-LLaVA [56 ] find that MoE
implementation achieves better performance than the dense counterpart on
almost all the benchmarks."
            ]
          ]
        },
        {
          "title": "Modality interface",
          "content": [
            [
              "Since LLMs can only perceive text, bridging the gap between
natural language and other modalities is necessary.",
              "Nevertheless, it would be costly to train from scratch a
large multimodal model in an end-toend manner.",
              "A more practical way is to introduce a learnable connector
between the pre-trained visual encoder and LLM.",
              "The other approach is to translate images into languages with
the help of expert models, and then send the language to the LLM."
            ]
          ]
        },
```

```
        {
          "title": "Learnable connector",
          "content": [
            [
              "The learnable connector is responsible for bridging the gap
between different modalities.",
              "Specifically, the module projects information into the space
that the LLM can understand efficiently.",
              "Based on how multimodal information is fused, there are
broadly two ways to implement such interfaces: token-level and feature-level
fusion for different modalities."
            ],
            [
              "For token-level fusion, features output from encoders are
transformed into tokens and concatenated with text tokens before being sent
into LLMs.",
              "A common solution is to leverage a group of learnable query
tokens to extract information in a query-based manner [57 ], which was first
implemented in BLIP-2 [50 ], and subsequently inherited by a variety of
works [22 ,51 ].",
              "Such Q-Former-style approaches compress visual tokens into a
smaller number of representation vectors.",
              "In contrast, some methods simply use an MLP-based interface
to bridge the modality gap [16 ].",
              "For example, LLaVA series adopt an MLP [16 ,39 ] to project
visual tokens and align the feature dimension with word embeddings.",
              "BLIVA [58 ] adopts an ensemble of MLPbased and Q-Former-based
connectors to enhance performance in text-rich scenarios."
            ],
            [
              "As another line, feature-level fusion inserts extra modules
that enable deep interaction and fusion between text features and visual
features.",
              "For example, Flamingo [59 ] inserts extra cross-attention
layers between frozen transformer layers of LLMs, thereby augmenting
language features with external visual cues.",
              "Similarly, CogVLM [60 ] plugs in a visual expert module in
each transformer layer to enable dual interaction and fusion between vision
and language features.",
              "For better performance, the QKV weight matrix of the
introduced module is initialized from the pre-trained LLM.",
              "Likewise, LLaMA-Adapter [61 ] introduces learnable prompts
into transformer layers.",
              "These prompts are first embedded w ith v isual knowledge and
then concatenated with text features as prefixes."
            ],
            [
              "On a related note, MM1 [41 ] has conducted ablation studies
on the design choices of the connector and found that, for token-level
fusion, the type of modality adapter is far less important than the number
of visual tokens and input resolution.",
              "Nevertheless, Zeng et al. [62 ] compared the performance of
token-and feature-level fusion, and empirically revealed that the token-
level fusion variant performs better in terms of VQA benchmarks.",
              "Regarding the performance gap, the authors suggested that
cross-attention models might require a more complicated hyper-parameter
searching process to achieve comparable performance."
            ],
            [
              "In terms of parameter size, learnable interfaces generally
comprise a small portion compared with encoders and LLMs.",
              "Take Qwen-VL [28 ] as an example; the parameter size of the
Q-Former is about 0.08B, accounting for less than 1% of the whole
```

```
parameters, while the encoder and the LLM account for about 19.8% (1.9B) and
80.2% (7.7B), respectively."
            ]
          ]
        },
        {
          "title": "Expert model",
          "content": [
            [
              "Apart from the learnable interface, using expert models, such
as an image captioning model, is also a feasible way to bridge the modality
gap [63 ].",
              "The basic idea is to convert multimodal inputs into languages
Scheme 1.",
              "A simplified template to structure the caption data.",
              "{ < image > } is the placeholder for the visual tokens, and
{caption} is the caption for the image.",
              "Note that only the part marked in red is used for loss
calculation."
            ],
            [
              "without training.",
              "In this way, LLMs can understand multimodality by the
converted languages.",
              "For example, VideoChat-Text [21 ] uses pre-trained vision
models to extract visual information such as actions and enriches the
descriptions using a speech recognition model.",
              "Though using expert models is straightforward, it may not be
as flexible as adopting a learnable interface.",
              "The conversion of foreign modalities into text would cause
information loss.",
              "For example, transforming videos into textual descriptions
distorts spatial-temporal relationships [21 ]."
            ]
          ]
        },
        {
          "title": "TRAINING STRATEGY AND DATA",
          "content": [
            [
              "A full-fledged MLLM undergoes three stages of training: pre-
training, instruction tuning and alignment tuning.",
              "Each phase of training requires different types of data and
fulfil ls different objectives.",
              "In this section, we discuss training objectives, as well as
data collection and characteristics for each training stage."
            ]
          ]
        },
        {
          "title": "Pre-training",
          "content": []
        },
        {
          "title": "Training detail",
          "content": [
            [
              "As the first training stage, pre-training mainly aims to
align different modalities and learn multimodal world knowledge.",
              "The pre-training stage generally entails large-scale text-
paired data, e.g.",
              "caption data.",
```

```
              "Typically, the caption pairs describe images/audio/videos in
natural language."
            ],
            [
              "Here, we consider a common scenario where MLLMs are trained
to align vision with text.",
              "As illustrated in Scheme 1 , given an image, the model is
trained to autoregressively predict the caption of the image, following a
standard cross-entropy loss.",
              "A common approach for pre-training is to freeze pretrained
modules (e.g.",
              "visual encoders and LLMs) and train a learnable interface [16
].",
              "The idea is to align different modalities without losing pre-
trained knowledge.",
              "Some methods [28 ] also unfreeze more modules (e.g. the
visual encoder) to enable more trainable parameters for alignment.",
              "It should be noted that the training scheme is closely
related to data quality.",
              "For short and noisy caption data, using lower resolution
(e.g.",
              "224 pixels) can speed up the training process, while for
longer and cleaner data, it is better to utilize higher resolutions (e.g.",
              "448 pixels or higher) to mitigate hallucinations.",
              "Besides, ShareGPT4V [64 ] finds that, with high-quality
caption data in the pre-training stage, unlocking the vision encoder
promotes better alignment."
            ]
          ]
        },
        {
          "title": "Data",
          "content": [
            [
              "Pre-training data mainly serve two purposes: aligning
different modalities and providing world knowledge.",
              "The pre-training corpora can be divided into coarse-grained
and fine-grained data according to granularities, which we wi l l introduce
sequentially.",
              "We summarize commonly used pre-training datasets in Table 3
."
            ],
            [
              "Coarse-grained caption data share some typical traits in
common.",
              "(i) The data volume is large since samples are generally
sourced from the internet.",
              "(ii) Because of the web-scrawled nature, the captions are
usually short and noisy since they originate from the alt-text of the web
images.",
              "These data can be cleaned and filtered via automatic tools,
for example using the CLIP [12 ] model to filter out image-text pairs whose
similarities are lower than a predefined threshold.",
              "In what follows, we introduce some representative coarse-
grained datasets."
            ],
            [
              "CC Series.",
              "CC-3M [65 ] is a web-scale caption dataset of 3.3M image-
caption pairs, where the raw descriptions are derived from alt-text
associated with images.",
              "The authors designed a complicated pipeline to clean data.",
```

```
              "For images, those with inappropriate content or aspect ratio
are filtered.",
              "For text, NLP tools are used to obtain text annotations, with
samples filtered according to the designed heuristics.",
              "For image-text pairs, images are assigned labels via
classifiers.",
              "If text annotations do not overlap with image labels, the
corresponding samples are dropped.",
              "CC-12M [66 ] is a following work of CC-3M and contains 12.4M
image-caption pairs.",
              "Compared w ith the prev ious work, CC-12M relaxes and
simplifies the data-collection pipeline, thus collecting more data."
            ],
            [
              "SBU Captions [67 ].",
              "This is a captioned photo dataset containing 1M image-text
pairs, with images and descriptions sourced from Flickr.",
              "Specifically, an initial set of images is acquired by
querying the Flickr website with a large number of query terms.",
              "The descriptions attached to the images thus serve as
captions.",
              "Then, to ensure that descriptions are relevant to the images,
the retained images fulfil l the following requirements:"
            ],
            [
              "(i) descriptions of the images are of satisfactory length,
decided by observation; (ii) captions should contain at least two words in
the predefined term lists and a propositional word (e.g.",
              "'on' , 'under') that suggests spatial relationships."
            ],
            [
              "LAION.",
              "These series are large web-scale datasets, with images
scrawled from the internet and associated alt-text as captions.",
              "To filter the image-text pairs, the following steps are
performed:"
            ],
            [
              "(i) text with short lengths or images with too small or too
big sizes are dropped; (ii) image deduplication is performed based on the
URL; (iii) CLIP [12 ] embeddings for images and text are extracted, and the
embeddings are used to drop possibly i l legal content and image-text pairs
with low cosine similarity between embeddings."
            ],
            [
              "Here we offer a brief summary of some typical variants.",
              "r LAION-5B [68 ].",
              "This variant is a researchpurpose dataset of 5.85B image-text
pairs.",
              "The dataset is multilingual with a 2B English subset.",
              "r LAION-COCO [69 ].",
              "This variant contains 600M images extracted from the English
subset of LAION-5B.",
              "The captions are synthetic, using BLIP [70 ] to generate
various image captions and using CLIP [12 ] to pick the best fit.",
              "[71 ].",
              "This dataset contains 747M image-text pairs, which are
extracted from Common-Crawl.",
              "In terms of data filtering, the authors designed the
following strategies to filter out data samples.",
              "For images, those with inappropriate size, content, format or
aspect ratio are filtered.",
```

```
              "Moreover, the images are filtered based on the pHash value to
remove images overlapped with public datasets such as Im-ageNet and MS-
COCO.",
              "For text, only English text with satisfactory length, noun
forms and appropriate words are saved.",
              "Whitespace before and after the sentence wi l l be removed,
and consecutive whitespace characters wi l l be replaced with a single
whitespace.",
              "Moreover, text appearing more than 10 times (e.g.",
              "'image for') wi l l be dropped.",
              "For image-text pairs, duplicated samples are removed based on
the (pHash, text) tuple."
            ]
          ]
        },
        {
          "title": "COYO-700M",
          "content": [
            [
              "Recently, more works [64 ,73 ] have explored generating high-
quality fine-grained data through prompting strong MLLMs (e.g.",
              "GPT-4V).",
              "Compared with coarse-grained data, these data generally
contain longer and more accurate descriptions of the images, thus enabling
finer-grained alignment between image and text modalities.",
              "However, since the approach general ly requires cal ling
commercial-use MLLMs, the cost is higher, and the data volume is smaller.",
              "Notably, ShareGPT4V [64 ] strikes a balance by first training
a captioner with GPT-4V-generated 100K data, then scaling up the data volume
to 1.2M using the pre-trained captioner."
            ]
          ]
        },
        {
          "title": "Instruction tuning",
          "content": []
        },
        {
          "title": "Introduction",
          "content": [
            [
              "Instruction refers to the description of tasks.",
              "Intuitively, instruction tuning aims to teach models to
better understand the instructions from users and fulfill the demanded
tasks.",
              "Tuning in this way, LLMs can generalize to unseen tasks by
following new instructions, thus boosting zero-shot performance.",
              "This simple yet effective idea has sparked the success of
subsequent NLP works, such as ChatGPT [77 ] and InstructGPT [78 ]."
            ],
            [
              "The comparisons between instruction tuning and related
typical learning paradigms are i l lustrated in Fig. 3 .",
              "The supervised fine-tuning approach usually requires a large
amount of task-specific data to train a task-specific model.",
              "The prompting approach reduces the reliance on large-scale
data and can fulfil l a specialized task via prompt engineering.",
              "In such Scheme 2. A simplified template to structure the
multimodal instruction data.",
              "< instruction > is a textual description of the task.",
              "{ < image > , < text > } and < output > are the input and
output from the data sample.",
```

```
              "Note that < text > in the input may be missed for some
datasets; for example, image caption datasets merely have < image > .",
              "The example is adapted from [81 ]."
            ],
            [
              "a case, though the few-shot performance has been improved,
the zero-shot performance is sti l l quite average [4 ].",
              "Differently, instruction tuning learns how to generalize to
unseen tasks rather than fitting specific tasks like the two counterparts.",
              "Moreover, instruction tuning is highly related to multi-task
prompting [79 ] and learning [80 ]."
            ],
            [
              "In this section, we delineate the format of instruction
samples, training objectives, typical ways to gather instruction data and
corresponding commonly used datasets."
            ]
          ]
        },
        {
          "title": "Training detail",
          "content": [
            [
              "A multimodal instruction sample often includes an optional
instruction and an input-output pair.",
              "The instruction is typically a natural language sentence
describing the task, such as 'Describe the image in detail.'",
              "The input can be an image-text pair like the VQA task [82 ]
or only an image like the image caption task [83 ].",
              "The output is the answer to the instruction conditioned on
the input.",
              "The instruction template is flexible and subject to manual
designs [21 ], as exemplified in Scheme 2 .",
              "Note that the instruction template can also be generalized to
the case of multi-round human-agent conversations [16 ,81 ]."
            ],
            [
              "Formally, a multimodal instruction sample can be denoted in a
triplet form, i.e. (I, M , R ) , where I, M , R represent the instruction,
the multimodal input and the ground-truth response, respectively.",
              "The MLLM predicts an answer given the instruction and the
multimodal input:"
            ],
            [
              "Here, A denotes the predicted answer, and θ are the
parameters of the model.",
              "The training objective is typically the original auto-
regressive objective used to train LLMs [16 ], based on which the MLLM is
encouraged to predict the next token of the response Scheme 3. Instruction
templates for VQA datasets, cited from [51 ]. < Image > and {Question} are
the image and the question in the original VQA datasets, respectively."
            ],
            [
              "sequentially:"
            ],
            [
              "with N the length of the ground truth."
            ]
          ]
        },
        {
          "title": "Data collection",
          "content": [
```

```
            [
              "Since instruction data are more flexible in formats and
varied in task formulations, it is usually trickier and more costly to
collect data samples.",
              "In this section, we summarize three typical ways to harvest
instruction data at scale: data adaptation, selfinstruction and data
mixture.",
              "Data adaptation.",
              "Task-specific datasets are rich sources of high-quality
data.",
              "Hence, abundant works [51 ,84 ] have utilized existing high-
quality datasets to construct instruction-formatted datasets.",
              "Take the transformation of VQA datasets as an example; the
original sample is an input-out pair where the input comprises an image and
a natural language question, and the output is the textual answer to the
question conditioned on the image.",
              "The input-output pairs of these datasets could naturally
comprise the multimodal input and response of the instruction sample.",
              "The instructions, i.e. the descriptions of the tasks, can
either derive from manual design or from semi-automatic generation aided by
GPT.",
              "Specifically, some works [17 ] handcraft a pool of candidate
instructions and sample one of them during training.",
              "We offer an example of instruction templates for the VQA
datasets in Scheme 3 .",
              "The other works manually design some seed instructions and
use these to prompt GPT to generate more [21 ]."
            ],
            [
              "Note that, since the answers of existing VQA and caption
datasets are usually concise, directly using these datasets for instruction
tuning may limit the output length of MLLMs.",
              "There are two common strategies to tackle this problem.",
              "The first strategy is to specify the corresponding
requirements explicitly in the instructions.",
              "For example, ChatBridge [85 ] explicitly declares short and
brief for short-answer data.",
              "The second strategy is to extend the length of existing
answers [86 ].",
              "For example, M 3 IT [86 ] proposes to rephrase the original
answer by prompting Chat-GPT with the original question, answer and
contextual information of the image (e.g.",
              "caption and text extracted through OCR)."
            ],
            [
              "Self-instruction.",
              "Although existing multi-task datasets can contribute a rich
source of data, they usually do not meet human needs well in real-world
scenarios, such as multiple-round conversations."
            ],
            [
              "To tackle this issue, some works collect samples through
self-instruction [89 ], which utilizes LLMs to generate textual instruction-
following data using a few hand-annotated samples.",
              "Specifically, some instruction-following samples are
handcrafted as demonstrations, after which ChatGP T/GP T-4 is prompted to
generate more instruction samples with the demonstrations as guidance.",
              "LLaVA [16 ] extends the approach to the multimodal field by
translating images into text of captions and bounding boxes, and prompting
text-only GPT-4 to generate new data with the guidance of requirements and
demonstrations.",
              "In this way, a multimodal instruction dataset is constructed,
called LLaVA-Instruct-150k.",
```

```
              "Following this idea, subsequent works such as MiniGPT-4 [17 ]
and GPT4Tools [90 ] develop different datasets catering to different
needs.",
              "Recently, with the release of the more powerful multimodal
model GPT-4V, many works have adopted GPT-4V to generate data of higher
quality, as exemplified by LVIS-Instruct4V [72 ] and ALLaVA [73 ].",
              "We summarize the popular datasets generated through self-
instruction in Table 4 .",
              "It should be noted that Data mixture.",
              "Apart from the multimodal instruction data, language-only
user-assistant conversation data can also be used to improve conversational
proficiencies and instruction-fol lowing abi lities [91 ].",
              "LaVIN [91 ] directly constructs a minibatch by randomly
sampling from both language-only and multimodal data.",
              "MultiInstruct [84 ] probes different strategies for training
with a fusion of single-modal and multimodal data, including mixed
instruction tuning (combine both types of data and randomly shuffle) and
sequential instruction tuning (text data followed by multimodal data)."
            ]
          ]
        },
        {
          "title": "Data quality",
          "content": [
            [
              "Recent research has revealed that the data quality of
instruction-tuning samples is no less important than quantity.",
              "Lynx [62 ] finds that models pre-trained on large-scale but
noisy image-text pairs do not perform as well as models pre-trained with
smaller but cleaner datasets.",
              "Similarly, Wei et al. [92 ] found that less instruction-
tuning data with higher quality can achieve better performance.",
              "For data filtering, the work proposes some metrics to
evaluate data quality and, correspondingly, a method to automatically filter
out inferior vision-language data.",
              "Here we discuss two important aspects of data quality."
            ],
            [
              "Prompt diversity.",
              "The diversity of instructions has been found to be critical
for model performance.",
              "Lynx [62 ] empirically verifies that diverse prompts help
improve model performance and generalization ability."
            ],
            [
              "Task coverage.",
              "In terms of tasks involved in training data, Du et al. [93 ]
performed an empirical study and found that the visual reasoning task is
superior to captioning and QA tasks for boosting model performance.",
              "Moreover, the study suggests that more complex instructions
are better than increasing task diversity and incorporating fine-grained
spatial annotations."
            ]
          ]
        },
        {
          "title": "Alignment tuning",
          "content": []
        },
        {
          "title": "Introduction",
          "content": [
            [
```

```
              "Alignment tuning is more often used in scenarios where models
need to be aligned with specific human preferences, e.g.",
              "response with fewer hallucinations.",
              "Currently, reinforcement learning with human feedback (RLHF)
and direct preference optimization (DPO) are two main techniques for
alignment tuning.",
              "In this section, we introduce the main ideas of the two
techniques in sequence, offer some examples of how they are utilized in
addressing practical problems and, finally, give a compilation of the
related datasets."
            ]
          ]
        },
        {
          "title": "Training detail",
          "content": [
            [
              "RLHF [94 ,95 ].",
              "This technique aims to utilize reinforcement learning
algorithms to align LLMs with human preferences, with human annotations as
supervision in the training loop.",
              "As exemplified in In-structGPT [78 ], RLHF incorporates three
key steps."
            ],
            [
              "(i) Supervised fine-tuning.",
              "This step aims to finetune a pre-trained model to present the
preliminary desired output behavior.",
              "The fine-tuned model in the RLHF setting is called a policy
model .",
              "Note that this step might be skipped since the supervised
policy model π SFT can be initialized from an instruction-tuned model.",
              "(ii) Reward modeling.",
              "A reward model is trained using preference pairs in this
step.",
              "Given a multimodal prompt (e.g.",
              "image and text) x and a response pair (y w , y l ) , the
reward model r θ learns to give a higher reward to the preferred response y
w , and vice versa for y l , with the objective"
            ],
            [
              "where D = { (x, y w , y l ) } is the comparison dataset
labeled by human annotators.",
              "In practice, the reward model r θ shares a similar structure
with the policy model.",
              "(iii) Reinforcement learning.",
              "In this step, the proximal policy optimization (PPO)
algorithm is adopted to optimize the RL policy model π RL φ .",
              "A per-token KL penalty is often added to the training
objective to avoid deviating too far from the original policy [78 ],
resulting in the objective"
            ],
            [
              "where β is the coefficient for the KL penalty term.",
              "Typically, both the RL policy π RL φ and the reference model
π REF are initialized from the supervised model π SFT .",
              "The obtained RL policy model is expected to align with human
preferences through this tuning process."
            ],
            [
              "Researchers have explored using the RLHF techniques for
better multimodal alignment.",
```

```
              "For example, LLaVA-RLHF [96 ] collects human preference data
and tunes a model with fewer hallucinations based on LLaVA [16 ]."
            ],
            [
              "DPO [97 ].",
              "This technique learns from human preference labels, utilizing
a simple binary classification loss.",
              "Compared with the PPO-based RLHF algorithm, DPO is exempt
from learning an explicit reward model, thus simplifying the whole pipeline
to two steps: human preference data collection and preference learning.",
              "The learning objective for the algorithm is"
            ],
            [
              "RLHF-V [98 ] collects fine-grained (segment-level) preference
data pairs by correcting hallucinations in the model response and uses the
obtained data to perform dense DPO.",
              "Si l kie [99 ] instead collects preference data via prompting
GPT-4V and disti l ls the preference supervision into an instruction-tuned
model through DPO."
            ]
          ]
        },
        {
          "title": "Data",
          "content": [
            [
              "The gist of data collection for alignment tuning is to
collect feedback for model responses, i.e. to decide which response is
better.",
              "It is generally more expensive to collect such data, and the
amount of data used for this phase is typically even less than that used in
previous stages.",
              "In this part, we introduce some datasets and summarize them
in Table 5 ."
            ],
            [
              "LLaVA-RLHF [96 ].",
              "This dataset contains 10K preference pairs collected from
human feedback in terms of honesty and helpfulness.",
              "It mainly serves to reduce hallucinations."
            ],
            [
              "RLHF-V [98 ].",
              "This dataset has 5.7K fine-grained human feedback data
collected by performing segment-level hallucination corrections."
            ],
            [
              "VLFeedback [99 ].",
              "This dataset utilizes AI to provide feedback on model
responses.",
              "It contains more than 380K comparison pairs scored by GPT-4V
in terms of helpfulness, faithfulness and ethical concerns."
            ]
          ]
        },
        {
          "title": "EVALUATION",
          "content": [
            [
              "Evaluation is an essential part of developing MLLMs since it
provides feedback for model optimization and helps to compare the
performance of different models.",
```

```
              "Compared with evaluation methods of traditional multimodal
models, the evaluation of MLLMs exhibits several new traits.",
              "(1) Since MLLMs are generally versatile, it is important to
evaluate MLLMs comprehensively.",
              "(2) MLLMs exhibit many emergent capabilities that require
special attention (e.g.",
              "OCR-free math reasoning) and thus require new evaluation
schemes.",
              "The evaluation of MLLMs can be broadly categorized into two
types according to the question genres: closed-set and open-set
evaluation.",
              "Closed-set evaluation often involves task-specific benchmarks
and more comprehensive benchmarks specifically designed for the MLLM, where
answers are limited to predefined sets.",
              "Open-set evaluation typically includes manual scoring, GPT
scoring and case study."
            ]
          ]
        },
        {
          "title": "Closed set",
          "content": [
            [
              "Closed-set questions refer to a type of question where the
possible answer options are predefined and limited to a finite set.",
              "The evaluation is usually performed on task-specific
datasets.",
              "In this case, the responses can be naturally judged by
benchmark metrics.",
              "For example, InstructBLIP [51 ] reports the accuracy on
ScienceQA [100 ], as well as the CIDEr score [101 ] on NoCaps [102 ].",
              "The evaluation setting is typically zero shot [51 ,84 ] or
finetuning [29 ,51 ].",
              "The first setting often selects a wide range of datasets
covering different general tasks and splits them into held-in and held-out
datasets.",
              "After tuning on the former, zero-shot performance is
evaluated on the latter with unseen datasets or even unseen tasks.",
              "In contrast, the second setting is often observed in the
evaluation of domain-specific tasks.",
              "For example, LLaVA [16 ]"
            ]
          ]
        },
        {
          "title": "Open set",
          "content": [
            [
              "In contrast to the closed-set questions, the responses to
open-set questions can be more flexible, where MLLMs usually play a chatbot
role.",
              "Because the content of the chat can be arbitrary, it would be
trickier to judge than the closed-ended output.",
              "The criterion can be classified into manual scoring, GPT
scoring and case study approaches.",
              "Manual scoring requires humans to assess the generated
responses.",
              "This kind of approach often involves handcrafted questions
that are designed to assess specific dimensions.",
              "For example, mPLUG-Owl [107 ] collects a visually related
evaluation set to judge capabilities like natural image, diagram and
flowchart understanding.",
```

```
              "Similarly, GPT4Tools [90 ] builds two sets for the fine-
tuning and zero-shot performance, respectively, and evaluates the responses
in terms of thought, action, arguments and the whole."
            ],
            [
              "Since manual assessment is labor intensive, some researchers
have explored rating with GPT, namely, GPT scoring.",
              "This approach is often used to evaluate performance on
multimodal dialogue.",
              "LLaVA [16 ] proposes to score the responses via text-only
GPT-4 in terms of different aspects, such as helpfulness and accuracy.",
              "Specifically, 30 images are sampled from the COCO [108 ]
validation set, each associated with a short question, a detailed question
and a complex reasoning question via self-instruction on GPT-4.",
              "The answers generated by both the model and GPT-4 are sent to
GPT-4 for comparison.",
              "Subsequent works follow this idea and prompt ChatGPT or GPT-4
to rate results [29 ] or judge which one is better [109 ]."
            ],
            [
              "A main issue of applying tex t-only GPT-4 for evaluation is
that the judge is only based on translated text content, such as captions or
bounding box coordinates, without accessing the image [29 ].",
              "Thus, it may be questionable to set GPT-4 as the performance
upper bound in this case.",
              "With the release of the vision inter face of GP T, some works
exploit the more advanced GPT-4V model to assess the performance of MLLMs.",
              "For example, Woodpecker [63 ] adopts the GPT-4V model to
judge the response quality of model answers.",
              "The evaluation is expected to be more accurate than using
textonly GPT-4 since GPT-4V has direct access to the image."
            ],
            [
              "Since the benchmark evaluation is not comprehensive enough, a
supplementary approach is to compare the different capabilities of MLLMs
through case studies.",
              "For instance, some studies evaluate two typical advanced
commercial-use models, GPT-4V and Gemini.",
              "Yang et al. [110 ] performed in-depth qualitative analysis on
GPT-4V by crafting a series of samples across various domains and tasks,
spanning from preliminary ski l ls, such as caption and object counting, to
complex tasks that require world knowledge and reasoning, such as joke
understanding and indoor navigation as an embodied agent.",
              "Wen et al. [111 ] made a more focused evaluation of GPT-4V by
designing samples targeting automatic driving scenarios.",
              "Fu et al. [112 ] carried out a comprehensive evaluation on
Gemini-Pro by comparing the model against GPT-4V.",
              "The results suggest that GPT-4V and Gemini exhibit comparable
visual reasoning abilities in spite of different response styles."
            ]
          ]
        },
        {
          "title": "EXTENSIONS",
          "content": [
            [
              "Recent studies have made significant strides in extending the
capabilities of MLLMs, spanning from more potent foundational abilities to
broader coverage of scenarios.",
              "We trace the principal development of MLLMs in this regard."
            ]
          ]
        },
```

```
        {
          "title": "Granularity support",
          "content": [
            [
              "To facilitate better interaction between agents and users,
researchers have developed MLLMs with finer support of granularities in
terms of model inputs and outputs.",
              "On the input side, models that support finer control from
user prompts are developed progressively, evolving from image to region [24
] and even pixels [25 ].",
              "Specifical ly, Shi kra [24 ] supports region-level input and
understanding.",
              "Users may interact with the assistant more flexibly by
referring to specific regions, which are represented in bounding boxes of
natural language forms.",
              "Ferret [113 ] takes a step further and supports more flexible
referring by devising a hybrid representation scheme.",
              "The model supports different forms of prompts, including
point, box and sketch.",
              "Similarly, Osprey [25 ] supports point input by utilizing a
segmentation model [10 ].",
              "Aided by the exceptional capabilities of the pre-trained
segmentation model, Osprey enables specifying a single entity or part of it
with a single click.",
              "On the output side, grounding capabilities are improved in
line with the development of input support.",
              "Shikra [24 ] supports response grounded in the image with box
annotations, resulting in higher precision and finer referring experience.",
              "LISA [114 ] fur ther suppor ts mask-level understanding and
reasoning, which makes pixel-level grounding possible."
            ]
          ]
        },
        {
          "title": "Modality support",
          "content": [
            [
              "Increased support for modalities is a tendency for MLLM
studies.",
              "On the one hand, researchers have explored adapting MLLMs to
support the input of more multimodal content, such as the threedimensional
po int cloud [115 ].",
              "On the other hand, MLLMs are also extended to generate
responses of more modalities, such as image [116 ], audio [117 ] and video
[118 ].",
              "For example, NE xT-GP T [119 ] proposes a framework that
supports inputs and outputs of mixed modalities, specifically, combinations
of text, image, audio and video, with the help of diffusion models [120 ]
attached to the MLLM.",
              "The framework applies an encoder-decoder architecture and
puts the LLM as a pivot for understanding and reasoning."
            ]
          ]
        },
        {
          "title": "Language support",
          "content": [
            [
              "Current models are predominantly unilingual, probably due to
the fact that a high-quality non-English training corpus is scarce.",
              "Some works have been devoted to developing multilingual
models so that a broader range of users can be covered.",
```

```
              "Vis-CPM [121 ] transfers model capabilities to the
multilingual setting by designing a multi-stage training scheme.",
              "Specifically, the scheme takes English as a pivotal language,
with an abundant training corpus.",
              "Utilizing a pre-trained bilingual LLM, the multimodal
capabilities are transferred to Chinese by adding some translated samples
during instruction tuning.",
              "Taking a similar approach, Qwen-VL [28 ] is developed from
the bilingual LLM Qwen [48 ] and supports both Chinese and English.",
              "During pre-training, Chinese data are mixed into the training
corpus to preserve the bilingual capabilities of the model, taking up 22.7%
of the whole data volume."
            ]
          ]
        },
        {
          "title": "Scenario/task extension",
          "content": [
            [
              "Apart from developing common general-purpose assistants, some
studies have focused on more specific scenarios where practical conditions
should be considered, while others extend MLLMs to downstream tasks with
specific expertise."
            ],
            [
              "A typical tendency is to adapt MLLMs to more specific real-
life scenarios.",
              "For example, some works develop agents that interact with the
real world, e.g.",
              "user-friendly assistants specially designed for GUI, as
exemplified by CogAgent [32 ], AppAgent [122 ] and Mobile-Agent [123 ].",
              "Researchers also develop embodied agents [19 ,31 ] that can
perform reasoning, navigation and manipulation in the real world,
facilitating the development of automatic agents that can execute tasks for
humans.",
              "In general, these assistants excel in planning and performing
each step to fulfill tasks specified by users, acting as helpful agents for
humans."
            ],
            [
              "Another line is to augment MLLMs with specific ski l ls for
solving tasks in different domains, e.g.",
              "document understanding [30 ] and medical domains [29 ].",
              "For document understanding, mPLUG-DocOwl [124 ] utilizes
various forms of documentlevel data for tuning, resulting in an enhanced
model in OCR-free document understanding.",
              "TextMonkey [30 ] incorporates multiple tasks related to
document understanding to improve model performance.",
              "Similarly, MLLMs can also be trained to accommodate
traditional vision tasks such as visual grounding [125 ,126 ].",
              "Compared with traditional methods [13 ,127 ], MLLMs unify the
I/O format and streamline the whole learning and inference process.",
              "Specifically, it is feasible to recast the grounding task
into a conditioned box coordinate prediction task under a unified language
modeling objective [24 ,28 ,52 ].",
              "The model is trained to predict the coordinates of specified
objects in the form of natural language.",
              "MLLMs can also be extended to medical domains by insti l ling
specialized knowledge.",
              "For example, LLaVA-Med [29 ] develops assistants specialized
in medical image understanding and question answering by injecting domain
knowledge."
            ]
```

```
          ]
        },
        {
          "title": "Efficient MLLMs",
          "content": [
            [
              "Recently, using lightweight MLLMs for efficient deployment
has gained increased popularity [128 -130 ).",
              "These models are meticulously designed and optimized for more
economical utilization or resource-limited scenarios without compromising
too much on model performance."
            ],
            [
              "From a model perspective, various techniques have been
explored to facilitate efficient training and inference.",
              "For instance, MobileVLM [54 ] explores developing small-size
variants of MLLMs for resource-limited scenarios.",
              "Some designs and techniques are utilized for deployment on
mobile devices, such as LLMs of smaller size and quantization techniques to
speed up computation.",
              "Similarly, MiniCPM-V [129 ] builds efficient MLLMs for
endside computation.",
              "A Q-Former [28 ] is adopted to cut down the number of visual
tokens for each patch of the image."
            ],
            [
              "From a data perspective, Bunny [130 ] comprehensively
investigates efficient data selection and combination schemes for model
training.",
              "The obtained models achieve performance on par with MLLMs of
larger parameter sizes."
            ]
          ]
        },
        {
          "title": "MULTIMODAL HALLUCINATION",
          "content": [
            [
              "Multimodal hallucination refers to the phenomenon of
responses generated by MLLMs being inconsistent with the image content [63
].",
              "The fundamental problem has received increased attention.",
              "In this section, we briefly introduce related concepts and
research development.",
              "In what follows, we first introduce evaluation methods, which
are useful to gauge the performance of methods for mitigating
hallucinations.",
              "Then, we discuss mitigation methods of different kinds of
approaches."
            ]
          ]
        },
        {
          "title": "Preliminaries",
          "content": []
        },
        {
          "title": "Evaluation methods",
          "content": [
            [
              "CHAIR [132 ] is an early metric that evaluates hallucination
levels in open-ended captions.",
```

```
            "The metric measures the proportion of sentences with
hallucinated objects or the proportion of hallucinated objects in all the
objects mentioned.",
            "In contrast, POPE [133 ] is a method that evaluates closed-
set choices.",
            "Specifically, multiple prompts with binary choices are
formulated, each querying if a specific object exists in the image.",
            "With a similar evaluation approach, MME [104 ] provides a
more comprehensive evaluation, covering aspects of existence, count,
position and color, as exemplified in [63 ]."
          ],
          [
            "Different from previous approaches that use matching
mechanisms to detect hallucinations, some works explore automatic evaluation
of text responses via models.",
            "For example, HaELM [134 ] proposes using LLMs as a judge to
decide whether MLLMs' captions are correct against reference captions.",
            "In view of the fact that text-only LLMs can only access
limited image context and require reference annotations, Woodpecker [63 ]
uses GPT-4V to directly assess model responses grounded in the image."
          ]
        ]
      },
      {
        "title": "Mitigation methods",
        "content": [
          [
            "According to high-level ideas for mitigating hallucinations,
current methods can be roughly divided into three categories: pre-
correction, in-process correction and post-correction."
          ],
          [
            "Pre-correction.",
            "An intuitive solution for hallucination is to collect
specialized data (e.g.",
            "negative data) and use the data for fine-tuning, thus
achieving models with fewer hallucinations."
          ],
          [
            "LRV-Instruction [135 ] introduces a visual instruction-tuning
dataset to encourage faithful generation.",
            "Similarly, LLaVA-RLHF [96 ] collects human-preference pairs
and fine-tunes models with reinforcement learning techniques."
          ],
          [
            "In-process correction.",
            "Another line is to make improvements in architectural design
or feature representation.",
            "These works try to explore the reasons for hallucinations and
design remedies to mitigate them in the generation process.",
            "For example, HallE-Switch [131 ] introduces a continuous
controlling factor to control the extent of imagination in model output
during inference."
          ],
          [
            "Post-correction.",
            "Different from previous paradigms, post-correction mitigates
hallucinations in a post-remedy way.",
            "For example, Woodpecker [63 ] is a training-free framework
for hallucination correction.",
            "Specifically, the method incorporates expert models to
supplement Scheme 4. A simplified example of the template to structure an M-
ICL query, adapted from [81 ].",
```

```
              "For illustration, we list two in-context examples and a query
divided by a dashed line.",
              "{instruction} and {response} are texts from the data
sample.",
              "< image > is a placeholder to represent the multimodal input
(an image in this case).",
              "< BOS > and < EOS > are tokens denoting the start and the end
of the input to the LLM, respectively.",
              "contextual information of the image and crafts a pipeline to
correct hallucinations step by step."
            ]
          ]
        },
        {
          "title": "EXTENDED TECHNIQUES",
          "content": []
        },
        {
          "title": "Multimodal in-context learning",
          "content": [
            [
              "ICL is one of the important emergent abilities of LLMs.",
              "The essence of the technique is prompting the model with a
few examples as guidance to make it easier for the model to answer the
query.",
              "There are two good traits of ICL.",
              "(i) The crux of ICL is to learn from analogy [136 ], thus
largely reducing the requirement of data samples.",
              "(ii) ICL is usually implemented in a training-free way [136 ]
and can be flexibly integrated into various frameworks at inference time."
            ],
            [
              "In the context of the MLLM, ICL has been extended to more
modalities, leading to multimodal incontext learning (M-ICL).",
              "At inference time, M-ICL can be implemented by adding a
demonstration set, i.e. a set of in-context samples, to the original
sample.",
              "In this case, the template can be extended, as i l lustrated
in Scheme 4 ."
            ]
          ]
        },
        {
          "title": "Improvement on ICL capabilities",
          "content": [
            [
              "Recently, a growing amount of work has focused on enhancing
ICL performance under various scenarios.",
              "In this section, we trace the development of this field and
summarize relevant works."
            ],
            [
              "MIMIC-IT [137 ] combines in-context learning with instruction
tuning by building an instruction dataset formatted with multimodal
context.",
              "Some other works explore improving few-shot learning
performance under specific settings.",
              "For example, Link-context learning [138 ] focuses on the
causal relationships between demonstrations and queries, and casts a
contrast training scheme by formulating positive and negative image-
description pairs.",
```

```
            "Similarly, Yang et al. [139 ,140 ] explored different
strategies to optimize demonstration configurations (selections or orderings
of in-context samples) to achieve better few-shot performance."
          ]
        ]
      },
      {
        "title": "Applications",
        "content": [
          [
            "In terms of applications in multimodality, M-ICL is mainly
used in two scenarios: solving various visual reasoning tasks [141 ]"
          ]
        ]
      },
      {
        "title": "Multimodal chain of thought",
        "content": [
          [
            "CoT is 'a series of intermediate reasoning steps' [5 ].",
            "The technique has been proven to be effective in complex
reasoning tasks.",
            "The main idea is to prompt LLMs to output not only the final
answer, but also the reasoning process that leads to the answer, resembling
the cognitive process of humans."
          ],
          [
            "Inspired by the success in NLP realms, multiple works [144
,145 ] have proposed to extend the technique to multimodal CoT (M-CoT).",
            "We first introduce different paradigms for acquiring the M-
CoT ability.",
            "Then, we delineate more specific aspects of M-CoT, including
the chain configuration and the pattern."
          ]
        ]
      },
      {
        "title": "Learning paradigms",
        "content": [
          [
            "There are broadly three ways to acquire the M-CoT ability:
through fine-tuning and training-free few-or zero-shot learning."
          ],
          [
            "Intuitively, the fine-tuning approach often involves curating
specific datasets for M-CoT learning.",
            "For example, Lu et al."
          ],
          [
            "[100 ] constructed a scientific question-answering dataset
ScienceQA with lectures and explanations, which can serve as sources of
learning CoT reasoning."
          ],
          [
            "Compared with fine-tuning, few/zero-shot learning is more
computationally efficient.",
            "The fewshot learning approach typically requires handcrafted
in-context examples to teach reasoning step by step.",
            "In contrast, the zero-shot learning approach directly prompts
with designed instructions [144 ]."
          ]
        ]
      },
```

```
        {
          "title": "Chain configuration",
          "content": [
            [
              "Structure and length are two critical aspects of the
reasoning chains.",
              "In terms of structure, current methods can be divided into
single-chain [100 ] and tree-shape methods [146 ].",
              "Chain length can be categorized into adaptive and predefined
formations.",
              "The former configuration requires LLMs to decide when to halt
the reasoning chains [100 ], while the latter setting stops the chains with
a predefined length [147 ]."
            ]
          ]
        },
        {
          "title": "Generation patterns",
          "content": [
            [
              "We summarize the relevant works into an infillingbased
pattern and a predicting-based pattern.",
              "Specifical ly, the infil ling-based pattern demands deducing
steps between surrounding context (previous and following steps) to fill the
logical gaps [144 ].",
              "In contrast, the predicting-based pattern requires extending
the reasoning chains given conditions such as instructions and previous
reasoning history [142 ]."
            ]
          ]
        },
        {
          "title": "LLM-aided visual reasoning",
          "content": []
        },
        {
          "title": "Introduction",
          "content": [
            [
              "Inspired by the success of tool-augmented LLMs [148 ], some
researchers have explored the possibilities of invoking external tools or
vision foundation models for visual reasoning tasks.",
              "Taking LLMs as helpers with different roles, these works
build task-specific or general-purpose visual reasoning systems."
            ],
            [
              "Compared with conventional visual reasoning models, these
works manifest several good traits.",
              "For this part, we start by introducing different training
paradigms employed in the construction of LLM-aided visual reasoning
systems.",
              "Then, we delve into the primary roles that LLMs play within
these systems."
            ]
          ]
        },
        {
          "title": "Training paradigms",
          "content": [
            [
              "According to training paradigms, LLM-aided visual reasoning
systems can be divided into two types: training-free and fine-tuning."
            ],
```

```
            [
              "Training-free.",
              "With abundant prior knowledge stored in pre-trained LLMs, an
intuitive and simple way is to freeze pre-trained models and directly prompt
LLMs to fulfil l various needs.",
              "According to the setting, the reasoning systems can be
further categorized into few-shot models [142 ] and zero-shot models [150
]."
            ],
            [
              "Fine-tuning.",
              "Some works adopt further finetuning to improve the planning
abilities with respect to tool usage [90 ] or to improve localization
capabilities [114 ] of the system.",
              "For example, GPT4Tools [90 ] collects a tool-related
instruction dataset to fine-tune the model."
            ]
          ]
        },
        {
          "title": "Functions",
          "content": [
            [
              "Regarding what roles LLMs exactly play in LLMaided visual
reasoning systems, existing related works are divided into three types:"
            ],
            [
              "(i) the LLM as a controller; (ii) the LLM as a decision
maker; (iii) the LLM as a semantics refiner."
            ],
            [
              "We delineate how LLMs serve these roles in the following."
            ],
            [
              "The LLM as a controller.",
              "In this case, LLMs act as a central controller that breaks
down a complex task into simpler sub-tasks/steps and assigns these tasks to
appropriate tools/modules.",
              "Specifically, LLMs are prompted explicitly to output task
planning [151 ] or, more directly, the modules to call [90 ,142 ,143 ].",
              "For example, VisProg [143 ] prompts GPT-3 to output a visual
program, where each program line invokes a module to perform a sub-task."
            ],
            [
              "The LLM as a decision maker.",
              "In this case, complex tasks are solved in a multi-round
manner, often in an iterative way [152 ].",
              "Decision-makers often summarize the context to decide whether
to finish the task and organize the answer in a user-friendly way."
            ],
            [
              "The LLM as a semantics refiner.",
              "When the LLM is used as a semantics refiner, researchers
mainly utilize its rich linguistic and semantic knowledge.",
              "Specifically, LLMs are often instructed to integrate
information into fluent natural language sentences [153 ] or generate texts
according to different specific needs [149 ,150 ,154 ]."
            ]
          ]
        },
        {
          "title": "CHALLENGES AND FUTURE DIRECTIONS",
          "content": [
```

```
            [
              "The development of MLLMs is sti l l in a rudimentary stage
and thus leaves much room for improvement, which we summarize below."
            ],
            [
              "r Current MLLMs are limited in processing multimodal
information of long context.",
              "This restricts the development of advanced models with more
multimodal tokens, e.g.",
              "long-video understanding and long documents interleaved with
images and text.",
              "r MLLMs should be upgraded to follow more complicated
instructions.",
              "For example, a mainstream approach to generating high-quality
questionanswer pair data is sti l l prompting closed-source GPT-4V because
of its advanced instructionfol lowing capabi lities, whi le other models
general ly fai l to achieve such goals.",
              "r There is sti l l a large space for improvement in
techniques like M-ICL and M-CoT.",
              "Current research on the two techniques is sti l l
rudimentary, and the related capabilities of MLLMs are sti l l weak.",
              "Therefore, explorations on the underlying mechanisms and
potential improvements are promising.",
              "r Developing embodied agents based on MLLMs is a heated
topic.",
              "It would be meaningful to develop such agents that can
interact with the real world.",
              "Such endeavors require models with critical capabilities,
including perception, reasoning, planning and execution.",
              "r Safety issues: similar to LLMs, MLLMs can be vulnerable to
crafted attacks.",
              "In other words, MLLMs can be misled to output biased or
undesirable responses.",
              "Thus, improving model safety wi l l be an important research
topic.",
              "r Interdisciplinary research: given the strong generalization
capabilities and abundant pre-trained knowledge of MLLMs, a promising
research direction could be utilizing MLLMs to boost research fields of
natural sciences, e.g.",
              "leveraging MLLMs for analysis of medical images or remote
sensing images.",
              "To achieve this goal, injecting domain-specific multimodal
knowledge into MLLMs might be necessary."
            ]
          ]
        },
        {
          "title": "CONCLUSION",
          "content": [
            [
              "In this paper, we review the existing MLLM literature and
offer a broad view of its main directions, including the basic recipe and
related extensions.",
              "Moreover, we underscore the current research gaps that need
to be fil led and point out some promising research directions.",
              "We hope that this survey can offer readers a clear picture of
the current progress of the MLLM and inspire more relevant works.",
              "In light of the fact that the era of the MLLM has only just
begun, we wi l l keep updating this survey and hope that it can inspire more
research.",
              "An associated GitHub link collecting the latest papers is
available at: https://github.com/BradyFU/",
              "Awesome-Multimodal-Large-Language-Models."
```

```
            ]
          ]
        }
      ],
      "keywords": [
        "mllms",
        "models",
        "images",
        "training",
        "works",
        "instruction tuning",
        "multimodal instruction",
        "multimodal data",
        "caption data",
        "model performance"
      ],
      "measurement_units": [
        {
          "Symbol": "K",
          "Value": [
            "380"
          ],
          "SI_converted": "380.0, K"
        },
        {
          "Symbol": "K",
          "Value": [
            "5.7"
          ],
          "SI_converted": "5.7, K"
        },
        {
          "Symbol": "V",
          "Value": [
            "4"
          ],
          "SI_converted": "4.0, kg*m^2*s^-3*A^-1"
        },
        {
          "Symbol": "K",
          "Value": [
            "100"
          ],
          "SI_converted": "100.0, K"
        },
        {
          "Symbol": "K",
          "Value": [
            "10"
          ],
          "SI_converted": "10.0, K"
        }
      ]
    },
    "acknowledgement": null,
    "funding": "This work was supported in part by the National Natural
Science Foundation of China (62222213, 62406264, U22B2059, U23A20319,
62072423 and 61727809) and the Young Scientists Fund of the Natural Science
Foundation of Sichuan Province (2023NSFSC1402).",
    "references": [
      {
        "src": "Zhao WX, Zhou K, Li J et al. A survey of large language
models. arXiv: 2303.18223.",
```

```
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Xu B and Poo Mm. Large language models and brain-inspired
general intelligence. Natl Sci Rev 2023; 10 : nwad267.",
        "title": "Large language models and brain-inspired general
intelligence",
        "year": 2023,
        "journal": "National Science Review",
        "ISSN": "2095-5138",
        "ISSNe": "2053-714X"
      },
      {
        "src": "Peng B, Li C, He P et al. Instruction tuning with GPT-4.
arXiv: 2304.03277.",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Brown T, Mann B, Ryder N et al . Language models are few-
shot learners. In: Proceedings of the 34th International Conference on
Neural Information Processing Systems . Red Hook, NY: Cur- ran Associates,
2020, 1877-1901.",
        "title": "Language models are few-shot learners",
        "year": 2020,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Wei J, Wang X, Schuurmans D et al . Chain-of-thought prompt-
ing elicits reasoning in large language models. In: Proceedings of the 36th
International Conference on Neural Information Processing Systems . Red
Hook, NY: Curran Associates, 2024, 24824-37.",
        "title": "Chain-of-thought prompting elicits reasoning in large
language models",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Li H. Deep learning for natural language processing: advan-
tages and challenges. Natl Sci Rev 2018; 5 : 24-6.",
        "title": "Deep learning for natural language processing: advantages
and challenges",
        "year": 2018,
        "journal": "National Science Review",
        "ISSN": "2095-5138",
        "ISSNe": "2053-714X"
      },
      {
        "src": "Zhao W. A panel discussion on AI for science: the opportuni-
ties, challenges and reflections. Natl Sci Rev 2024; nwae119.",
        "title": "A panel discussion on AI for science: the opportunities,
challenges and reflections",
```

```
        "year": 2024,
        "journal": "National Science Review",
        "ISSN": "2095-5138",
        "ISSNe": "2053-714X"
      },
      {
        "src": "Xie WJ and Warshel A. Harnessing generative AI to decode
enzyme catalysis and evolution for enhanced engineering. Natl Sci Rev 2023;
10 : nwad331.",
        "title": "Harnessing generative AI to decode enzyme catalysis and
evolution for enhanced engineering",
        "year": 2023,
        "journal": "National Science Review",
        "ISSN": "2095-5138",
        "ISSNe": "2053-714X"
      },
      {
        "src": "Gong P, Guo H, Chen B et al. iEarth: an interdisciplinary
frame- work in the era of big data and AI for sustainable development. Natl
Sci Rev 2023; 10 : nwad178.",
        "title": "iEarth: an interdisciplinary framework in the era of big
data and AI for sustainable development",
        "year": 2023,
        "journal": "National Science Review",
        "ISSN": "2095-5138",
        "ISSNe": "2053-714X"
      },
      {
        "src": "Kirillov A, Mintun E, Ravi N et al. Segment anything. In:
2023 IEEE/CVF International Conference on Computer Vision (ICCV) .
Piscataway, NJ: IEEE Press, 2023, 3992-4003.",
        "title": "Segment Anything",
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Shen Y, Fu C, Chen P et al. Aligning and prompting every-
thing all at once for universal visual perception. In: 2024 IEEE/CVF
Conference on Computer Vision and Pattern Recog- nition (CVPR) . Los
Alamitos, CA: IEEE Computer Society, 2024, 13193-203.",
        "title": "Aligning and Prompting Everything All at Once for
Universal Visual Perception",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Radford A, Kim JW, Hallacy C et al. Learning transferable
vi- sual models from natural language supervision. In: Proceed- ings of the
38th International Conference on Machine Learn- ing . PMLR, 2021, 8748-63.
Natl Sci Rev, 2024, Vol. 11, nwae403",
        "title": "Learning transferable visual models from natural language
supervision",
        "year": 2021,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
```

```
        "src": "Li J, Selvaraju R, Gotmare A et al. Align before fuse:
Vision and language representation learning with momentum distillation. In:
Proceedings of the 35th International Conference on Neural Information
Processing Systems . Red Hook, NY: Curran Associates, 2024, 9694-705.",
        "title": "Align before fuse: Vision and language representation
learning with momentum distillation",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Wang P, Yang A, Men R et al. OFA: unifying architectures,
tasks, and modali- ties through a simple sequence-to-sequence learning
framework. In: Proceed- ings of the 39th International Conference on Machine
Learning . PMLR, 2022, 23318-40.",
        "title": "OFA: unifying architectures, tasks, and modalities through
a simple sequence-to-sequence learning framework",
        "year": 2022,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Cho J, Lei J, Tan H et al. Unifying vision-and-language
tasks via text genera- tion. In: Proceedings of the 38th International
Conference on Machine Learn- ing . PMLR, 2021, 1931-42.",
        "title": "Unifying vision-and-language tasks via text generation",
        "year": 2021,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Liu H, Li C, Wu Q et al. Visual instruction tuning. In:
Proceedings of the 37th In- ternational Conference on Neural Information
Processing Systems . Red Hook, NY: Curran Associates, 2023, 34892-916.",
        "title": "Visual instruction tuning",
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Zhu D, Chen J, Shen X et al. MiniGPT-4: enhancing vision-
language under- standing with advanced large language models. In: 12th
International Confer- ence on Learning Representations . Vienna, Austria, 7-
11 May 2024.",
        "title": "Enhancing Interactive Image Retrieval With Query Rewriting
Using Large Language Models and Vision Language Models",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Yang Z, Li L, Wang J et al. MM-REACT: prompting ChatGPT for
multimodal reasoning and action . arXiv: 2303.11381.",
        "title": "MM-InstructEval: Zero-shot evaluation of (Multimodal)
Large Language Models on multimodal reasoning tasks",
        "year": 2025,
        "journal": "Information Fusion",
        "ISSN": "1566-2535",
```

```
        "ISSNe": null
      },
      {
        "src": "Driess D, Xia F, Sajjadi MS et al. PaLM-E: an embodied
multimodal lan- guage model. In: Proceedings of the 40th International
Conference on Ma- chine Learning . JMLR, 2023, 8469-88.",
        "title": "PaLM-E: an embodied multimodal language model",
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "OpenAI, Achiam J and Steven Adler S et al. GPT-4 technical
report. arXiv: 2303.08774.",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Li K, He Y, Wang Y et al. VideoChat: chat-centric video
understanding. arXiv: 2305.06355.",
        "title": "UniFormerV2: Unlocking the Potential of Image ViTs for
Video Understanding",
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Zhang H, Li X, Bing L. Video-LLaMA: an instruction-tuned
audio-visual lan- guage model for video understanding. In: Proceedings of
the 2023 Conference on Empirical Methods in Natural Language Processing.
Association for Com- putational Linguistics, 2023, 543-53.",
        "title": "Video-LLaMA: An Instruction-tuned Audio-Visual Language
Model for Video Understanding",
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Deshmukh S, Elizalde B, Singh R et al. Pengi: an audio
language model for audio tasks. In: Proceedings of the 37th International
Conference on Neu- ral Information Processing Systems . Red Hook, NY: Curran
Associates, 2023, 18090-108.",
        "title": "PAM: Prompting Audio-Language Models for Audio Quality
Assessment",
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Chen K, Zhang Z, Zeng W et al. Shikra: unleashing multimodal
LLM's referen- tial dialogue magic. arXiv: 2306.15195.",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
```

```
      },
      {
        "src": "Yuan Y, Li W, Liu J et al. Osprey: pixel understanding with visual instruction tuning. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recog- nition (CVPR) . Los Alamitos, CA: IEEE Computer Society, 2024, 28202-11.",
        "title": "Osprey: Pixel Understanding with Visual Instruction Tuning",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Han J, Zhang R, Shao W et al. ImageBind-LLM: multi-modality instruction tun- ing. arXiv: 2309.03905.",
        "title": "Figure 7: Prompt for multi-agent LLM for best instruction choice.",
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Moon S, Madotto A, Lin Z et al. AnyMAL: an efficient and scalable any- modality augmented language model. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (Industry Track), Asso- ciation for Computational Linguistics , 2024, 1314-32.",
        "title": "AnyMAL: An Efficient and Scalable Any-Modality Augmented Language Model",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Bai J, Bai S, Yang S et al. Qwen-VL: a versatile vision-language model for understanding, localization, text reading, and beyond. arXiv: 2308.12966.",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Li C, Wong C, Zhang S et al. LLaVA-med: training a large language-and-vision assistant for biomedicine in one day. In: Proceedings of the 37th International Conference on Neural Information Processing Systems . Red Hook, NY: Curran Associates, 2023, 28541-64.",
        "title": "LLaVA-MR: Large Language-and-Vision Assistant for Video Moment Retrieval",
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Liu Y, Yang B, Liu Q et al. TextMonkey: an OCR-free large multimodal model for understanding document. arXiv: 2403.04473.",
        "title": null,
        "year": 2025,
        "journal": null,
```

```
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Huang J, Yong S, Ma X et al. An embodied generalist agent in
3D world. International Conference on Machine Learning, Vienna, Austria, 21-
27 July 2024 .",
        "title": "Vienna (July 27-28)",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Hong W, Wang W, Lv Q et al. CogAgent: a visual language
model for gui agents. In: 2024 IEEE/CVF Conference on Computer Vision and
Pattern Recog- nition (CVPR) . Los Alamitos, CA: IEEE Computer Society,
2024, 14281-90.",
        "title": "CogAgent: A Visual Language Model for GUI Agents",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Cherti M, Beaumont R, Wightman R et al. Reproducible scaling
laws for con- trastive language-image learning. In: 2023 IEEE/CVF Conference
on Computer Vision and Pattern Recognition (CVPR) . Los Alamitos, CA: IEEE
Computer So- ciety, 2023, 2818-29.",
        "title": "Reproducible Scaling Laws for Contrastive Language-Image
Learning",
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Sun Q, Fang Y, Wu L et al. EVA-CLIP: improved training
techniques for CLIP at scale. arXiv: 2303.15389.",
        "title": "Alpha-CLIP: A CLIP Model Focusing on Wherever you Want",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Chen Z, Wang W, Tian H et al. How far are we to GPT-4V?
closing the gap to commercial multimodal models with open-source suites.
arXiv: 2404.16821.",
        "title": "How far are we to GPT-4V? Closing the gap to commercial
multimodal models with open-source suites",
        "year": 2024,
        "journal": "Science China Information Sciences",
        "ISSN": "1674-733X",
        "ISSNe": "1869-1919"
      },
      {
        "src": "Fang Y, Wang W, Xie B et al. EVA: exploring the limits of
masked visual repre- sentation learning at scale. In: 2023 IEEE/CVF
Conference on Computer Vision and Pattern Recognition (CVPR) . Los Alamitos,
CA: IEEE Computer Society, 2023, 19358-69.",
        "title": "EVA: Exploring the Limits of Masked Visual Representation
Learning at Scale",
```

```
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Bavishi R, Elsen E, Hawthorne C et al. Introducing our
multimodal models . https://www.adept.ai/blog/fuyu-8b (17 October 2024, date
last accessed).",
        "title": null,
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Li Z, Yang B, Liu Q et al. Monkey: image resolution and text
label are important things for large multi-modal models. In: 2024 IEEE/CVF
Conference on Com- puter Vision and Pattern Recognition (CVPR) . Los
Alamitos, CA: IEEE Computer Society, 2024, 26753-63.",
        "title": "Monkey: Image Resolution and Text Label are Important
Things for Large Multi-Modal Models",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Liu H, Li C, Li Y et al. Improved baselines with visual
instruction tuning. In: 2024 IEEE/CVF Conference on Computer Vision and
Pattern Recognition (CVPR) . Los Alamitos, CA: IEEE Computer Society, 2024,
26286-96.",
        "title": "Improved Baselines with Visual Instruction Tuning",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Lin Z, Liu C, Zhang R et al. SPHINX: the joint mixing of
weights, tasks, and visual embeddings for multi-modal large language models.
arXiv:2311.07575.",
        "title": "SPHINX: A Mixer of Weights, Visual Embeddings and Image
Scales for Multi-modal Large Language Models",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "McKinzie B, Gan Z, Fauconnier JP et al. MM1: methods,
analysis & insights from multimodal LLM pre-training. In: Computer Vision-
ECCV 2024 . Berlin: Springer, 2020, 304-23.",
        "title": "MM1: Methods, Analysis and Insights from Multimodal LLM
Pre-training",
        "year": 2020,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Elizalde B, Deshmukh S, Al Ismail M et al. CLAP learning
audio concepts from natural language supervision. In: 2023 IEEE
```

```
International Conference on Acoustics, Speech and Signal Processing (ICASSP)
. Piscataway, NJ: IEEE Press, 2023, 1-5.",
        "title": "CLAP Learning Audio Concepts from Natural Language
Supervision",
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Girdhar R, El-Nouby A, Liu Z et al. ImageBind: one embedding
space to bind them all. In: 2023 IEEE/CVF Conference on Computer Vision and
Pattern Recognition (CVPR) . Los Alamitos, CA: IEEE Computer Society, 2023,
15180- 90.",
        "title": "ImageBind One Embedding Space to Bind Them All",
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Chung HW, Hou L, Longpre S et al. Scaling instruction-
finetuned language models. J Mach Learn Res 2024; 25 : 70.",
        "title": "Scaling instruction-finetuned language models",
        "year": 2024,
        "journal": "J Mach Learn Res",
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Touvron H, Lavril T, Izacard G et al. LLaMA: open and
efficient foundation language models. arXiv: 2302.13971.",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Chiang WL, Li Z, Lin Z et al. Vicuna: an open-source chatbot
impressing GPT-4 with 90% ChatGPT quality . https://vicuna.lmsys.org (17
October 2024, date last accessed).",
        "title": null,
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Touvron H, Martin L, Stone K et al. Llama 2: open foundation
and fine-tuned chat models. arXiv:2307.09288.",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Bai J, Bai S, Chu Y et al. Qwen technical report. arXiv:
2309.16609.",
        "title": null,
        "year": 2021,
        "journal": null,
```

```
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Meta. Introducing Meta Llama 3: the most capable openly
available LLM to date . https://ai.meta.com/blog/meta-llama-3 (17 October
2024, date last ac- cessed).",
        "title": null,
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Li J, Li D, Savarese S et al. BLIP-2: bootstrapping
language-image pre-training with frozen image encoders and large language
models. In: Proceedings of the 40th International Conference on Machine
Learning . JMLR, 2023, 19730- 42.",
        "title": "BLIP-2: bootstrapping language-image pre-training with
frozen image encoders and large language models",
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Dai W, Li J, Li D et al. InstructBLIP: towards general-
purpose vision-language models with instruction tuning. In: Proceedings of
the 37th International Natl Sci Rev, 2024, Vol. 11, nwae403 Conference on
Neural Information Processing Systems . Red Hook, NY: Cur- ran Associates,
2023, 49250-67.",
        "title": "InstructBLIP: towards general-purpose vision-language
models with instruction tuning",
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Liu H, Li C, Li Y et al. LLaVA-NeXT: improved reasoning,
OCR, and world knowledge . https://llava-vl.github.io/blog/2024-01-30-llava-
next (17 Octo- ber 2024, date last accessed).",
        "title": "LLaVA-MR: Large Language-and-Vision Assistant for Video
Moment Retrieval",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Lu Y, Li C, Liu H et al. An empirical study of scaling
instruct-tuned large mul- timodal models. arXiv: 2309.09958.",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Chu X, Qiao L, Lin X et al. MobileVLM: a fast, reproducible
and strong vision language assistant for mobile devices. arXiv:
2312.16886.",
```

```
        "title": "Retracted: FastQR: Fast Pose Estimation of Objects Based
on Multiple QR Codes and Monocular Vision in Mobile Embedded Devices",
        "year": 2023,
        "journal": "Wireless Communications and Mobile Computing",
        "ISSN": "1530-8669",
        "ISSNe": "1530-8677"
      },
      {
        "src": "Shen S, Hou L, Zhou Y et al. Mixture-of-experts meets
instruction tuning: a winning combination for large language models. In:
12th International Con- ference on Learning Representations . Vienna,
Austria, 7-11 May 2024.",
        "title": "Mixture-of-experts meets instruction tuning: a winning
combination for large language models",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Lin B, Tang Z, Ye Y et al. MoE-LLaVA: mixture of experts for
large vision- language models. arXiv: 2401.15947.",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Carion N, Massa F, Synnaeve G et al. End-to-end object
detection with trans- formers. In: Computer Vision-ECCV 2020 . Berlin:
Springer, 2020, 213-29.",
        "title": "End-to-End Object Detection with Transformers",
        "year": 2020,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Hu W, Xu Y, Li Y et al. BLIVA: a simple multimodal LLM for
better handling of text-rich visual questions. In: 38th AAAI Conference on
Artificial Intelligence . Washington, DC: Association for the Advancement of
Artificial Intelligence, 2024, 2256-64.",
        "title": "BLIVA: A Simple Multimodal LLM for Better Handling of
Text-Rich Visual Questions",
        "year": 2024,
        "journal": "Proceedings of the AAAI Conference on Artificial
Intelligence",
        "ISSN": "2159-5399",
        "ISSNe": "2374-3468"
      },
      {
        "src": "Alayrac J-B, Donahue J, Luc P et al. Flamingo: a visual
language model for few-shot learning. In: Proceedings of the 36th
International Conference on Neural Information Processing Systems . Red
Hook, NY: Curran Associates, 2024, 23716-36.",
        "title": "Flamingo: a visual language model for few-shot learning",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
```

```
        "src": "Wang W, Lv Q, Yu W et al. CogVLM: visual expert for
pretrained language models. In: 38th International Conference on Neural
Information Processing Systems , Vancouver, Canada, 10-15 Dec 2024.",
        "title": "CogVLM: visual expert for pretrained language models",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Zhang R, Han J, Zhou A et al. LLaMA-Adapter: efficient fine-
tuning of language models with zero-init attention. 12th International
Conference on Learning Representations , Vienna, Austria, 7-11 May 2024.",
        "title": "LLaMA-Adapter: efficient fine-tuning of language models
with zero-init attention",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Zeng Y, Zhang H, Zheng J et al. What matters in training a
GPT4-style lan- guage model with multimodal inputs?In: Proceedings of the
2024 Conference of the North American Chapter of the Association for
Computational Linguis- tics . Kerrville, TX: Association for Computational
Linguistics, 2024, 7937-64.",
        "title": "What Matters in Training a GPT4-Style Language Model with
Multimodal Inputs?",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Yin S, Fu C, Zhao S et al. Woodpecker: hallucination
correction for multimodal large language models. arXiv: 2310.16045.",
        "title": "Woodpecker: hallucination correction for multimodal large
language models",
        "year": 2024,
        "journal": "Science China Information Sciences",
        "ISSN": "1674-733X",
        "ISSNe": "1869-1919"
      },
      {
        "src": "Chen L, Li J, Dong X et al. ShareGPT4V: improving large
multi-modal models with better captions. arXiv: 2311.12793.",
        "title": "ShareGPT4V: Improving Large Multi-modal Models with Better
Captions",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Sharma P, Ding N, Goodman S et al. Conceptual captions: a
cleaned, hyper- nymed, image alt-text dataset for automatic image
captioning. In: Proceedings of the 56th Annual Meeting of the Association
for Computational Linguistics . Kerrville, TX: Association for Computational
Linguistics, 2018, 2556-65.",
        "title": "Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text
Dataset For Automatic Image Captioning",
        "year": 2018,
        "journal": null,
```

```
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Changpinyo S, Sharma P, Ding N et al. Conceptual 12M:
pushing web- scale image-text pre-training to recognize long-tail visual
concepts. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern
Recognition (CVPR) . Los Alamitos, CA: IEEE Computer Society, 2021, 3557-
67.",
        "title": "Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training
To Recognize Long-Tail Visual Concepts",
        "year": 2021,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Ordonez V, Kulkarni G, Berg T. Im2Text: describing images
using 1 million cap- tioned photographs. In: Proceedings of the 24th
International Conference on Neural Information Processing Systems . Red
Hook, NY: Curran Associates, 2011, 1143-51.",
        "title": "Im2Text: describing images using 1 million captioned
photographs",
        "year": 2011,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Schuhmann C, Beaumont R, Vencu R et al. LAION-5B: an open
large-scale dataset for training next generation image-text models. In:
Proceedings of the 36th International Conference on Neural Information
Processing Systems . Red Hook, NY: Curran Associates, 2024, 25278-94.",
        "title": "LAION-5B: an open large-scale dataset for training next
generation image-text models",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Schuhmann C, Köpf A, Vencu R et al. Laion coco: 600M
synthetic captions from Laion2B-en . https://laion.ai/blog/laion-coco (17
October 2024, date last accessed).",
        "title": null,
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Li J, Li D, Xiong C et al. BLIP: bootstrapping language-
image pre-training for unified vision-language understanding and generation.
In: Proceedings of the 39th International Conference on Machine Learning .
PMLR, 2022, 12888-900.",
        "title": "BLIP: bootstrapping language-image pre-training for
unified vision-language understanding and generation",
        "year": 2022,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
```

```
        "src": "Byeon M, Park B, Kim H et al. COYO-700M: image-text pair
dataset . https: //github.com/kakaobrain/coyo-dataset (17 October 2024, date
last accessed).",
        "title": null,
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Wang J, Meng L, Weng Z et al. To see is to believe:
prompting GPT-4V for better visual instruction tuning. arXiv: 2311.07574.",
        "title": null,
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Chen GH, Chen S, Zhang R et al. ALLaVA: harnessing GPT4V-
synthesized data for a lite vision-language model. arXiv: 2402.11684.",
        "title": "Vision Language Model Helps Private Information De-
Identification in Vision Data",
        "year": 2025,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Xu J, Mei T, Yao T et al. MSR-VTT: a large video description
dataset for bridg- ing video and language. In: 2016 IEEE Conference on
Computer Vision and Pattern Recognition (CVPR) . Los Alamitos, CA: IEEE
Computer Society, 2016, 5288-96.",
        "title": "MSR-VTT: A Large Video Description Dataset for Bridging
Video and Language",
        "year": 2016,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Mei X, Meng C, Liu H et al. WavCaps: a ChatGPT-assisted
weakly-labelled au- dio captioning dataset for audio-language multimodal
research. In: IEEE/ACM Transactions on Audio, Speech, and Language
Processing . Piscataway, NJ: IEEE Press, 2024, 3339-54.",
        "title": "WavCaps: A ChatGPT-Assisted Weakly-Labelled Audio
Captioning Dataset for Audio-Language Multimodal Research",
        "year": 2024,
        "journal": "IEEE/ACM Transactions on Audio, Speech, and Language
Processing",
        "ISSN": "2329-9290",
        "ISSNe": "2329-9304"
      },
      {
        "src": "Wei J, Bosma M, Zhao VY et al. Finetuned language models are
zero-shot learners. International Conference on Learning Representations ,
Virtual, 25- 29 April 2022.",
        "title": "Finetuned language models are zero-shot learners",
        "year": 2022,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
```

```
      {
        "src": "OpenAI. Introducing ChatGPT .
https://www.openai.com/research/chatgpt (17 October 2024, date last
accessed).",
        "title": null,
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Ouyang L, Wu J, Jiang X et al. Training language models to
follow instructions with human feedback. In: Proceedings of the 36th
International Conference on Neural Information Processing Systems . Red
Hook, NY: Curran Associates, 2024, 27730-44.",
        "title": "Training language models to follow instructions with human
feedback",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Sanh V, Webson A, Raffel C et al. Multitask prompted
training enables zero- shot task generalization. International Conference on
Learning Representa- tions , Virtual, 25-29 April 2022.",
        "title": "Multitask prompted training enables zeroshot task
generalization",
        "year": 2022,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Zhang Y and Yang Q. An overview of multi-task learning. Natl
Sci Rev 2018; 5 : 30-43.",
        "title": "An overview of multi-task learning",
        "year": 2018,
        "journal": "National Science Review",
        "ISSN": "2095-5138",
        "ISSNe": "2053-714X"
      },
      {
        "src": "Gong T, Lyu C, Zhang S et al. MultiModal-GPT: a vision and
language model for dialogue with humans. arXiv: 2305.04790.",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Antol S, Agrawal A, Lu J et al. VQA: visual question
answering. In: 2015 IEEE International Conference on Computer Vision (ICCV)
. Los Alamitos, CA: IEEE Computer Society, 2015, 2425-33.",
        "title": "VQA: Visual Question Answering",
        "year": 2015,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
```

```
        "src": "Karpathy A and Fei-Fei L. Deep visual-semantic alignments
for generating im- age descriptions. In: 2015 IEEE Conference on Computer
Vision and Pattern Recognition (CVPR) . Los Alamitos, CA: IEEE Computer
Society, 2015, 3128- 37.",
        "title": "Deep visual-semantic alignments for generating image
descriptions",
        "year": 2015,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Xu Z, Shen Y, Huang L. MultiInstruct: improving multi-modal
zero-shot learn- ing via instruction tuning. In: Proceedings of the 61st
Annual Meeting of the Association for Computational Linguistics . Kerrville,
TX: Association for Com- putational Linguistics, 2023, 11445-65.",
        "title": "MultiInstruct: Improving Multi-Modal Zero-Shot Learning
via Instruction Tuning",
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Zhao Z, Guo L, Yue T et al. ChatBridge: bridging modalities
with large language model as a language catalyst. arXiv: 2305.16103.",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Li L, Yin Y, Li S et al. M 3 IT: a large-scale dataset
towards multi-modal multi- lingual instruction tuning. arXiv: 2306.04387.",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Maaz M, Rasheed H, Khan S et al. Video-ChatGPT: towards
detailed video understanding via large vision and language models. In:
Proceed- ings of the 62nd Annual Meeting of the Association for
Computational Linguistics . Kerrville, TX: Association for Computational
Linguistics, 2024, 12585-602.",
        "title": "Video-ChatGPT: Towards Detailed Video Understanding via
Large Vision and Language Models",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Drossos K, Lipping S, Virtanen T. Clotho: an audio
captioning dataset. In: 2020 IEEE International Conference on Acoustics,
Speech and Signal Processing (ICASSP) . Piscataway, NJ: IEEE Press, 2020,
736-40. Natl Sci Rev, 2024, Vol. 11, nwae403",
        "title": "Clotho: an Audio Captioning Dataset",
        "year": 2020,
        "journal": null,
        "ISSN": null,
```

```
        "ISSNe": null
      },
      {
        "src": "Wang Y, Kordi Y, Mishra S et al. Self-instruct: aligning
language model with self generated instructions. In: Proceedings of the 61st
Annual Meeting of the Association for Computational Linguistics . Kerrville,
TX: Association for Computational Linguistics, 2023, 13484-508.",
        "title": "Self-Instruct: Aligning Language Models with Self-
Generated Instructions",
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Yang R, Song L, Li Y et al. GPT4Tools: teaching large
language model to use tools via self-instruction. In: Proceedings of the
37th International Conference on Neural Information Processing Systems . Red
Hook, NY: Curran Associates, 2023, 71995-2007.",
        "title": "InstOptima: Evolutionary Multi-objective Instruction
Optimization via Large Language Model-based Instruction Operators",
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Luo G, Zhou Y, Ren T et al. Cheap and quick: efficient
vision-language instruc- tion tuning for large language models . In:
Proceedings of the 37th Interna- tional Conference on Neural Information
Processing Systems . Red Hook, NY: Curran Associates, 2023, 29615-27.",
        "title": "Cheap and quick: efficient vision-language instruction
tuning for large language models",
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Wei L, Jiang Z, Huang W et al. InstructionGPT-4: a 200-
instruction paradigm for fine-tuning MiniGPT-4. arXiv: 2308.12067.",
        "title": "Table 4: Llama-3.1 fine tuning instruction format.",
        "year": null,
        "journal": "MiniGPT",
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Du Y, Guo H, Zhou K et al. What makes for good visual
instructions? Syn- thesizing complex visual reasoning instructions for
visual instruction tuning. arXiv: 2311.01487.",
        "title": "INSTRUCTIONS FOR AUTHORS",
        "year": 1996,
        "journal": "Visual Resources",
        "ISSN": "0197-3762",
        "ISSNe": "1477-2809"
      },
      {
        "src": "Ziegler DM, Stiennon N, Wu J et al. Fine-tuning language
models from human preferences. arXiv: 1909.08593.",
        "title": null,
        "year": null,
        "journal": null,
```

```
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Stiennon N, Ouyang L, Wu J et al. Learning to summarize with
human feedback. In: Proceedings of the 34th International Conference on
Neural Information Processing Systems . Red Hook, NY: Curran Associates,
2020, 3008-21.",
        "title": "Learning to summarize with human feedback",
        "year": 2020,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Sun Z, Shen S, Cao S et al. Aligning large multimodal models
with factually augmented RLHF. In: Findings of the Association for
Computational Linguis- tics: ACL 2024 . Kerrville, TX: Association for
Computational Linguistics, 2024, 13088-110.",
        "title": "Aligning Large Multimodal Models with Factually Augmented
RLHF",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Rafailov R, Sharma A, Mitchell E et al. Direct preference
optimization: your language model is secretly a reward model. In:
Proceedings of the 37th International Conference on Neural Information
Processing Systems . Red Hook, NY: Curran Associates, 2023, 53728-41.",
        "title": "Retracted: College English Teaching Evaluation Model Using
Natural Language Processing Technology and Neural Networks",
        "year": 2023,
        "journal": "Mobile Information Systems",
        "ISSN": "1574-017X",
        "ISSNe": "1875-905X"
      },
      {
        "src": "Yu T, Yao Y, Zhang H et al. RLHF-V: towards trustworthy
MLLMs via behavior alignment from fine-grained correctional human feedback.
In: 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition
(CVPR) . Los Alamitos, CA: IEEE Computer Society, 2024, 13807-16.",
        "title": "RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment
from Fine-Grained Correctional Human Feedback",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Li L, Xie Z, Li M et al. Silkie: preference distillation for
large visual language models. arXiv: 2312.10665. 100",
        "title": "ZVQAF: Zero-shot visual question answering with feedback
from large language models",
        "year": 2024,
        "journal": "Neurocomputing",
        "ISSN": "0925-2312",
        "ISSNe": null
      },
      {
        "src": "Lu P, Mishra S, Xia T et al. Learn to explain: multimodal
reasoning via thought chains for science question answering. In: Proceedings
```

```
of the 36th Interna- tional Conference on Neural Information Processing
Systems . Red Hook, NY: Curran Associates, 2024, 2507-21. 101",
        "title": "Learn to explain: multimodal reasoning via thought chains
for science question answering",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Vedantam R, Lawrence Zitnick C, Parikh D. CIDEr: consensus-
based image de- scription evaluation. In: 2015 IEEE Conference on Computer
Vision and Pattern Recognition (CVPR) . Los Alamitos, CA: IEEE Computer
Society, 2015, 4566-75. 102",
        "title": "CIDEr: Consensus-based image description evaluation",
        "year": 2015,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Agrawal H, Desai K, Wang Y et al. nocaps: novel object
captioning at scale. In: 2019 IEEE/CVF International Conference on Computer
Vision (ICCV) . Los Alamitos, CA: IEEE Computer Society, 2019, 8947-56.
103",
        "title": "nocaps: novel object captioning at scale",
        "year": 2019,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "He X, Zhang Y, Mou L et al. PathVQA: 30000+ questions for
medical visual question answering. arXiv: 2003.10286. 104",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Fu C, Chen P, Shen Y et al. MME: a comprehensive evaluation
benchmark for multimodal large language models. arXiv: 2306.13394. 105",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Liu Y, Duan H, Zhang Y et al. MMBench: is your multi-modal
model an all- around player? In: Leonardis A, Ricci E, Roth S et al. (eds)
Computer Vision- ECCV 2024 . Cham: Springer, 2024, 216-33. 106",
        "title": "MMBench: Is Your Multi-modal Model an All-Around Player?",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Ning M, Zhu B, Xie Y et al. Video-Bench: a comprehensive
benchmark and toolkit for evaluating video-based large language models.
```

```
arXiv: 2311.16103. 107. Ye Q, Xu H, Xu G et al. mPLUG-Owl: modularization
empowers large language models with multimodality. arXiv: 2304.14178. 108",
        "title": "Video-Bench: a comprehensive benchmark and toolkit for
evaluating video-based large language models",
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Lin TY, Maire M, Belongie S et al. Microsoft COCO: common
objects in con- text. In: Fleet D, Pajdla T, Schiele B et al. (eds) Computer
Vision-ECCV 2014 . Cham: Springer, 2014, 740-55. 109",
        "title": "Microsoft COCO: Common Objects in Context",
        "year": 2014,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Gao P, Han J, Zhang R et al. LLaMA-Adapter V2: parameter-
efficient visual instruction model. arXiv: 2304.15010. 110",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Yang Z, Li L, Lin K et al. The dawn of LMMs: preliminary
explorations with GPT-4V(ision). arXiv: 2309.17421. 111",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Wen L, Yang X, Fu D et al. On the road with GPT-4V(ision):
early explorations of visual-language model on autonomous driving. arXiv:
2311.05332. 112",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Fu C, Zhang R, Lin H et al. A challenger to GPT-4V? Early
explorations of Gemini in visual expertise. arXiv: 2312.12436. 113",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "You H, Zhang H, Gan Z et al. Ferret: refer and ground
anything anywhere at any granularity. 12th International Conference on
Learning Representations , Vienna, Austria, 7-11 May 2024. 114",
        "title": "Ferret: refer and ground anything anywhere at any
granularity",
        "year": 2024,
```

```
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Lai X, Tian Z, Chen Y et al. LISA: reasoning segmentation
via large lan- guage model. In: 2024 IEEE/CVF Conference on Computer Vision
and Pat- tern Recognition (CVPR) . Los Alamitos, CA: IEEE Computer Society,
2024, 9579-89. 115",
        "title": "LISA: Reasoning Segmentation via Large Language Model",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Xu R, Wang X, Wang T et al. PointLLM: empowering large
language models to understand point clouds . In: Leonardis A, Ricci E, Roth
S et al. (eds) Computer Vision-ECCV 2024 . Cham: Springer, 2024, 131-47.
116",
        "title": "PointLLM: Empowering Large Language Models to Understand
Point Clouds",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Sun Q, Yu Q, Cui Y et al. Generative pretraining in
multimodality. 12th Inter- national Conference on Learning Representations ,
Vienna, Austria, 7-11 May 2024. 117",
        "title": "Generative pretraining in multimodality",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Zhang D, Li S, Zhang X et al. SpeechGPT: empowering large
language models with intrinsic cross-modal conversational abilities. In:
Findings of the 2023 Conference on Empirical Methods in Natural Language
Processing . Kerrville, TX: Association for Computational Linguistics, 2023,
15757-73. 118",
        "title": "SpeechGPT: Empowering Large Language Models with Intrinsic
Cross-Modal Conversational Abilities",
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Wang X, Zhuang B, Wu Q. ModaVerse: efficiently transforming
modalities with llms . In: 2024 IEEE/CVF Conference on Computer Vision and
Pattern Recognition (CVPR) . Los Alamitos, CA: IEEE Computer Society, 2024,
26596- 606. 119",
        "title": "ModaVerse: Efficiently Transforming Modalities with LLMs",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
```

```
        "src": "Wu S, Fei H, Qu L et al. NExT-GPT: any-to-any multimodal
LLM. 12th Interna- tional Conference on Learning Representations , Vienna,
Austria, 7-11 May 2024. 120",
        "title": "NExT-GPT: any-to-any multimodal LLM",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic
models. In: Proceed- ings of the 34th International Conference on Neural
Information Processing Systems . Red Hook, NY: Curran Associates, 2020,
6840-51. 121",
        "title": "Denoising diffusion probabilistic models",
        "year": 2020,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Hu J, Yao Y, Wang C et al. Large multilingual models pivot
zero-shot multi- modal learning across languages. 12th International
Conference on Learning Representations , Vienna, Austria, 7-11 May 2024.
122",
        "title": "Large multilingual models pivot zero-shot multimodal
learning across languages",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Yang Z, Liu J, Han Y et al. AppAgent: multimodal agents as
smartphone users. arXiv: 2312.13771. 123",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Wang J, Xu H, Ye J et al. Mobile-Agent: autonomous multi-
modal mobile de- vice agent with visual perception. arXiv: 2401.16158. 124",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Ye J, Hu A, Xu H et al. mPLUG-DocOwl: modularized multimodal
large lan- guage model for document understanding. arXiv: 2307.02499. 125",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Yu L, Poirson P, Yang S et al. Modeling context in referring
expressions. In: Leibe B, Matas J, Sebe N et al. (eds) Computer Vision-ECCV
2016 . Cham: Springer, 2016, 69-85. 126",
```

```
        "title": "Modeling Context in Referring Expressions",
        "year": 2016,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Mao J, Huang J, Toshev A et al. Generation and comprehension
of unambigu- ous object descriptions. In: 2016 IEEE Conference on Computer
Vision and Pattern Recognition (CVPR) . Los Alamitos, CA: IEEE Computer
Society, 2016, 11-20. 127",
        "title": "Generation and Comprehension of Unambiguous Object
Descriptions",
        "year": 2016,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Zeng Y, Zhang X, Li H. Multi-grained vision language pre-
training: Aligning texts with visual concepts. In: Proceedings of the 39th
International Confer- ence on Machine Learning . PMLR, 2022, 25994-6009.
128. OpenAI. GPT-4o mini: advancing cost-efficient intelligence .
https://openai. com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/
(17 October 2024, date last accessed). 129",
        "title": "Multi-grained vision language pre-training: Aligning texts
with visual concepts",
        "year": 2022,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Yao Y, Yu T, Zhang A et al. MiniCPM-V: a GPT-4V level MLLM
on your phone. arXiv: 2408.01800. 130",
        "title": null,
        "year": 2025,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "He M, Liu Y, Wu B et al. Efficient multimodal learning from
data-centric per- spective. arXiv: 2402.11530. 131",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Zhai B, Yang S, Zhao X et al. HallE-Control: controlling
object hallucination in large multimodal models. arXiv: 2310.01779. 132",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Rohrbach A, Hendricks LA, Burns K et al. Object
hallucination in image cap- tioning. In: Proceedings of the 2018 Conference
```

```
on Empirical Methods in Nat- ural Language Processing . Kerrville, TX:
Association for Computational Lin- guistics, 2018, 4035-45. 133",
        "title": "Object Hallucination in Image Captioning",
        "year": 2018,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Li Y, Du Y, Zhou K et al. Evaluating object hallucination in
large vision- language models. In: Proceedings of the 2023 Conference on
Empirical Meth- ods in Natural Language Processing . Kerrville, TX:
Association for Computa- tional Linguistics, 2023, 292-305. 134",
        "title": "Evaluating Object Hallucination in Large Vision-Language
Models",
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Wang J, Zhou Y, Xu G et al. Evaluation and analysis of
hallucination in large vision-language models. arXiv: 2308.15126. 135",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Liu F, Lin K, Li L et al. Mitigating hallucination in large
multi-modal models via robust instruction tuning. 12th International
Conference on Learning Rep- resentations , Vienna, Austria, 7-11 May 2024.
136",
        "title": "Mitigating hallucination in large multi-modal models via
robust instruction tuning",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Dong Q, Li L, Dai D et al. A survey for in-context learning.
arXiv: 2301.00234. 137",
        "title": "A Survey on In-context Learning",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Li B, Zhang Y, Chen L et al. MIMIC-IT: multi-modal in-
context instruction tun- ing. arXiv: 2306.05425. 138",
        "title": "Otter: a Multi-Modal Model with in-Context Instruction
Tuning",
        "year": 2025,
        "journal": "IEEE Transactions on Pattern Analysis and Machine
Intelligence",
        "ISSN": "0162-8828",
        "ISSNe": "1939-3539"
      },
      {
```

```
        "src": "Tai Y, Fan W, Zhang Z et al. Link-context learning for
multimodal LLMs. In: 2024 IEEE/CVF Conference on Computer Vision and Pattern
Recognition (CVPR) . Los Alamitos, CA: IEEE Computer Society, 2024, 27166-
75. 139",
        "title": "Link-Context Learning for Multimodal LLMs",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Yang X, Wu Y, Yang M et al. Exploring diverse in-context
configurations for im- age captioning. In: Proceedings of the 37th
International Conference on Neu- ral Information Processing Systems . Red
Hook, NY: Curran Associates, 2023, 40924-43. 140",
        "title": "Exploring diverse in-context configurations for image
captioning",
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Yang X, Peng Y, Ma H et al. Lever LM: configuring in-context
sequence to lever large vision language models. In: 38th International
Conference on Neural Information Processing Systems , Vancouver, Canada, 10-
15 Dec 2024. 141",
        "title": "Empowering Vision-Language Models for Reasoning Ability
through Large Language Models",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Yang Z, Gan Z, Wang J et al. An empirical study of GPT-3 for
few-shot knowledge-based VQA. In: 36th AAAI Conference on Artificial
Intelligence (AAAI-22 . Washington, DC: Association for the Advancement of
Artificial In- telligence, 2022, 3081-89. 142",
        "title": "An Empirical Study of GPT-3 for Few-Shot Knowledge-Based
VQA",
        "year": 2022,
        "journal": "Proceedings of the AAAI Conference on Artificial
Intelligence",
        "ISSN": "2159-5399",
        "ISSNe": "2374-3468"
      },
      {
        "src": "Lu P, Peng B, Cheng H et al. Chameleon: plug-and-play
compositional rea- soning with large language models. In: Proceedings of the
37th International Conference on Neural Information Processing Systems . Red
Hook, NY: Curran Associates, 2023, 43447-78. 143",
        "title": "Chameleon: plug-and-play compositional reasoning with
large language models",
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Gupta T and Kembhavi A. Visual programming: compositional
visual reason- ing without training. In: 2023 IEEE/CVF Conference on
```

```
Computer Vision and Pattern Recognition (CVPR) . Los Alamitos, CA: IEEE
Computer Society, 2023, 14953-62. 144",
        "title": "Visual Programming: Compositional visual reasoning without
training",
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Rose D, Himakunthala V, Ouyang A et al. Visual chain of
thought: bridging logical gaps with multimodal infillings. arXiv:
2305.02317. 145",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Zhang Z, Zhang A, Li M et al. Multimodal chain-of-thought
reasoning in lan- guage models. In: Transactions on Machine Learning
Research . Brookline, MA: Microtome Publishing, 2024. 146",
        "title": "Multimodal Chain-of-Thought Reasoning in Language Models",
        "year": 2024,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Zheng G, Yang B, Tang J et al. DDCoT: duty-distinct chain-
of-thought prompt- ing for multimodal reasoning in language models. In:
Proceedings of the 37th International Conference on Neural Information
Processing Systems . Red Hook, NY: Curran Associates, 2023, 5168-91. 147",
        "title": "Multimodal Chain-of-Thought Reasoning in Language Models",
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Ge J, Luo H, Qian S et al. Chain of thought prompt tuning in
vision language models. arXiv: 2304.07919. 148",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Parisi A, Zhao Y, Fiedel N. TALM: tool augmented language
models. arXiv: 2205.12255. 149",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Zhu X, Zhang R, He B et al. PointCLIP V2: prompting CLIP and
GPT for pow- erful 3D open-world learning. In: 2023 IEEE/CVF International
Conference on Computer Vision (ICCV) . Los Alamitos, CA: IEEE Computer
Society, 2023, 2639-50. 150",
```

```
        "title": "PointCLIP V2: Prompting CLIP and GPT for Powerful 3D Open-
world Learning",
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Wang T, Zhang J, Fei J et al. Caption anything: interactive
image description with diverse multimodal controls. arXiv: 2305.02677. 151",
        "title": null,
        "year": null,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
      {
        "src": "Shen Y, Song K, Tan X et al. , HuggingGPT: solving AI tasks
with chatgpt and its friends in hugging face. Conference on Neural
Information Processing Sys- tems 2024; In: Proceedings of the 37th
International Conference on Neural Information Processing Systems . Red
Hook, NY: Curran Associates, 2023, 38154-80. 152",
        "title": "LatinX in AI at Neural Information Processing Systems
Conference 2023",
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      },
…
      {
        "src": "Zhang R, Hu X, Li B et al. Prompt, generate, then cache:
cascade of founda- tion models makes strong few-shot learners. In: 2023
IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) . Los
Alamitos, CA: IEEE Computer Society, 2023, 15211-22.",
        "title": "Prompt, Generate, Then Cache: Cascade of Foundation Models
Makes Strong Few-Shot Learners",
        "year": 2023,
        "journal": null,
        "ISSN": null,
        "ISSNe": null
      }
    ],
    "alliances": ["UN", "BRICS", "G20", "SCO"],
    "tables": [
      {
        "name": "Table 1",
        "number": "1",
        "caption": "A summary of commonly used image encoders.",
        "src": "85a20e29-6e93-4a9e-a39f-0f47cc382a11/tables/Table 1
IoMd3gJV.html"
      },
      {
        "name": "Table 2",
        "number": "2",
        "caption": "A summary of commonly used open-sourced LLMs.En, Zh, Fr
and De stand for English, Chinese, French and German, respectively.",
        "src": "85a20e29-6e93-4a9e-a39f-0f47cc382a11/tables/Table 2
ezpkwxIC.html"
      },
      {
        "name": "Table 3",
        "number": "3",
```

```
        "caption": "Common datasets used for pre-training.",
        "src": "85a20e29-6e93-4a9e-a39f-0f47cc382a11/tables/Table 3
3u8dEQlt.html"
      },
      {
        "name": "Table 4",
        "number": "4",
        "caption": "A summary of popular datasets generated by self-
instruction.For input/output modalities, I denotes image, T denotes text, V
denotes video, A denotes audio.For data composition, M-T and S-T denote
multi-turn and single-turn, respectively.",
        "src": "85a20e29-6e93-4a9e-a39f-0f47cc382a11/tables/Table 4
igtw6lkV.html"
      },
      {
        "name": "Table 5",
        "number": "5",
        "caption": "A summary of datasets for alignment tuning.For
input/output modalities, I denotes image and T denotes text.",
        "src": "85a20e29-6e93-4a9e-a39f-0f47cc382a11/tables/Table 5
d82NMpb1.html"
      }
    ],
    "images": [
      {
        "name": "Figure 1",
        "number": "1",
        "caption": "A time line of representative MLLMs. We are witnessing
rapid growth in this field. More works can be found on our released GitHub
page, which is updated daily",
        "src": "85a20e29-6e93-4a9e-a39f-0f47cc382a11/images/Figure 1
t51NVpc3.jpeg"
      },
      {
        "name": "Figure 2",
        "number": "2",
        "caption": "An illustration of typical MLLM architecture",
        "src": "85a20e29-6e93-4a9e-a39f-0f47cc382a11/images/Figure 2
f32VVcqz.jpeg"
      },
      {
        "name": "Figure 3",
        "number": "3",
        "caption": "Comparison of three typical learning paradigms, adapted
from [76].",
        "src": "85a20e29-6e93-4a9e-a39f-0f47cc382a11/images/Figure 3
er2w1lkC.jpeg"
      }
    ],
    "file_paths": [
      "4e845a0f-501c-4954-9cf3-c6cc53f3ff19/85a20e29-6e93-4a9e-a39f-
0f47cc382a11.pdf"
    ],
    "tasks_uuid": [
      "4e845a0f-501c-4954-9cf3-c6cc53f3ff19"
    ],
    "collections": [
      "test65"
    ],
    "created_at": "05/08/2025",
    "updated_at": "05/08/2025"
  }
```

ПРИЛОЖЕНИЕ В ПРИМЕР СТРУКТУРЫ JSON ФАЙЛА С ДАННЫМИ ЛС, СОБРАННОГО ИЗ MEDSCAPE


```
{
 ...
  {
    "name": "enasidenib",
    "comment": "Rx",
    "other_names": [
        "Idhifa"
    ],
    "classes": [
        "IDH2 Inhibitors"
    ],
    "source": "https://reference.medscape.com/drug/idhifa-enasidenib-
1000160",
    "pregnancy": {
        "common": [
            "Based on animal embryofetal toxicity studies, fetal harm may
occur when administered to pregnant females",
            "No data available on use in pregnant females to inform a drug-
associated risk of major birth defects and miscarriage"
        ],
        "specific": [
            {
                "type": "Animal data",
                "description": [
                    "Oral administration of enasidenib to pregnant rats and
rabbits during organogenesis was associated with embryofetal mortality and
alterations to growth starting at 0.1 times the steady state clinical
exposure based on the AUC at the recommended human dosage",
                    "If used during pregnancy, or if patient becomes
pregnant while taking this drug, advise of the potential risk to a fetus"
                ]
            },
...
        ]
    },
    "lactation": {
        "common": [
            "There are no data on the presence of enasidenib or its
metabolites in human milk, effects on the breastfed infants, or effects on
milk production",
            "Advise women not to breastfeed during treatment with enasidenib
and for at least 2 months after last dose"
        ]
    },
    "warnings": {
        "black_box_warning": {
            "common": [],
            "specific": [
                {
                    "type": "Differentiation syndrome",
                    "description": [
                        "In the clinical trial, 14% of patients treated with
enasidenib experienced symptoms of differentiation syndrome, which can be
fatal if not treated",
                        "Symptoms may include fever, dyspnea, acute
respiratory distress, pulmonary infiltrates, pleural or pericardial
effusions, rapid weight gain or peripheral edema, lymphadenopathy, bone
pain, and hepatic, renal, or multiorgan dysfunction",
```

```
                        "Differentiation syndrome has been observed with and
without concomitant hyperleukocytosis, as early as 1 day and at up to 5
months after initiating enasidenib", ...
                    ]
                }
            ]
        },
    },
...
    "interactions": [
        {
            "classification_type": "Minor",
            "interaction_with": "dienogest/estradiol valerate",
            "description": {
                "common": "enasidenib, dienogest/estradiol valerate. unknown
mechanism. Minor/Significance Unknown. Coadministration of enasidenib may
increase or decrease the concentrations of combined hormonal contraceptives.
Clinical significance of this interaction is unknown."
            }
        },
...
    ],
    "adverse effects": [
        {
            "name": "Total bilirubin increased",
            "percent": "81"
        },
        {
            "name": "Calcium decreased",
            "percent": "74"
        },
...
    ]
}
  ...
}
```