ОБЗОРНАЯ СТАТЬЯ

DOI: https://doi.org/10.18599/grs.2023.4.13

УДК 553.9:004

Цифровая научная платформа «Агрегатор неструктурированных геолого-промысловых данных»: архитектура и базовые модели извлечения данных

О.А. Невзорова, Р.Р. Хакимуллин, И.И. Идрисов[*]

Казанский (Приволжский) федеральный университет, Казань, Россия

В статье описывается разрабатываемый проект цифровой научной платформы «Агрегатор неструктурированных геолого-промысловых данных», который потенциально может иметь важное значение для нефтегазовой отрасли. Применение новых интеллектуальных технологий в рамках этого проекта позволит существенно повысить эффективность процессов обработки, хранения и использования геолого-промысловой информации, содержащейся в различных текстовых источниках, в основном в отчетах о месторождениях.

Главной целью разработки цифровой научной платформы является интегрирование разнородной информации об объектах исследования недр, которая извлекается из отчетов о месторождениях Республики Татарстан. Это позволит создать сводную базу данных, которая станет основой для принятия обоснованных решений в нефтегазовой сфере. Проект цифровой научной платформы включает разработку архитектуры, алгоритмов и программных решений, основанных на современных методах обработки текстов и интеллектуальном анализе данных.

Ключевые слова: сбор и анализ данных, отчеты о месторождениях, база данных, автоматизация, большие данные, обработка текстовых данных, неструктурированные данные, извлечение информации

Для цитирования: Невзорова О.А., Хакимуллин Р.Р., Идрисов И.И. (2023). Цифровая научная платформа «Агрегатор неструктурированных геолого-промысловых данных»: архитектура и базовые модели извлечения данных. *Георесурсы,* 25(4), с. 149–156. https://doi.org/10.18599/grs.2023.4.13

Введение

Цифровая модернизация нефтегазовых компанийя вляется актуальной задачей. Основные пути развития цифровых технологий связаны с системами искусственного интеллекта (машинное обучение, глубокое обучение), ботосферой (роботизация, боты, дроны) и виртуальной реальностью (дополненная реальность, цифровой двойник, смешанная реальность), а также с использованием технологий больших данных (big data) для геолого-промысловых данных (Дежина и др., 2017; Abdelhamid et al., 2022; Choubey, Karmakar, 2021; Deloitte Analysis Report, 2019).

Ключевой проблемой методов машинного обучения является подготовка больших коллекций данных. Нефтегазовая область очень богата разнообразными эмпирическими данными. Отметим некоторые приложения больших данных в нефтегазовой отрасли, такие как обработка сейсмических данных для определения критических геологических особенностей, геофизические измерения во время бурения, данные каротажа скважин и др., для которых разработаны различные программные средства, в том числе на основе технологий искусственного интеллекта (Goodfellow et al., 2016; Technavio, 2015).

В настоящее время также существует большой объем неструктурированных геолого-промысловых данных, фиксированных в различных документах-источниках, таких как отчеты по месторождениям, акты выполненных работ различного назначения и т.п.

Обработка неструктурированных текстовых данных является сложной задачей, требующей привлечения автоматических методов обработки текстов, и программные решения в этой области недостаточно разработаны. Можно выделить ряд типовых задач, связанных с обработкой этих данных. Классической задачей является извлечение из геологических текстов ограниченного числа именованных сущностей. Так, в (Lucas P. Cinelli et al., 2021) рассматривается проблема автоматического извлечения событий из ежедневных отчетов о бурении. Предлагались два различных подхода: на основе правил экспертной системы и глубоких нейронных сетей. Событиями, извлекаемыми из текстов, являются различные сбои при бурении. Оба алгоритма разрабатывались на основе специально подготовленного набора данных на португальском языке и имели высокие значения средних истинно положительных результатов (для алгоритма на основе правил – 97,3%, для трансформеров – 85,61%).

В последние годы для извлечения именованных сущностей из геологических текстов на разных языках активно используются алгоритмы машинного обучения и нейросети. В работе (Nooralahzadeh et al., 2018) представлены модели Word Embeddings, специфичные для нефтегазовой отрасли. Показано, что специфичные эмбеддинги могут быть пригодными, даже если корпус текстов, используемый для их обучения, значительно меньше корпуса текстов общей направленности.

[*] Ответственный автор: Ильяс Ирекович Идрисов
e-mail: ilyas_irekovich@mail.ru

В (Qiu et al., 2019) рассматривается нейронная сеть BiLSTM-CRF с механизмом внимания для распознавания именованных сущностей на китайском языке в области геолого-геофизических исследований. В подходе использовались модели Word2vec и Glove Word Embeddings, обученные на китайской Википедии. Полученные результаты были сопоставимы с результатами других исследователей и оценивались метрикой F1, равной 91,47%.

В (Consoli et al., 2020) применяется подход на нейронных сетях двунаправленной долговременной краткосрочной памяти — условных случайных полей (BiLSTM-CRF), который достаточно широко используется в этой области исследований. В работе применялись три типа векторных и тензорных представлений (эмбеддингов) – Word Embeddings, Flair Embeddings и Stacked Embeddings. Две первые модели дали наилучший результат на специальном наборе данных GeoCorpus на португальском языке. Наилучший результат с метрикой F1 84,63% показала модель Flair с расширенными геологическими характеристиками и использованием слов общего домена.

Использование онтологий – еще один подход в задаче извлечения именованных сущностей. В (Chengbin Wang et al., 2023) текстовые данные о месторождении используются для построения онтологии. Далее на основе этой онтологии разрабатываются схемы аннотации именованных объектов месторождения. В статье (Hoffimann Julio et al., 2018) описаны результаты разработки методологии автоматической классификации предложений, содержащихся в отчетах о бурении по трем меткам (событие, симптом и действие) для сотен скважин на реальном месторождении. Отмечается сложный характер исходных текстов-отчетов, такие как высокая частота использования технических символов, опечатки/сокращения технических терминов, а также наличие неполных предложений в отчетах о бурении.

Таким образом, можно утверждать, что для обработки неструктурированных геологических документов в настоящее время разработаны решения некоторых частных задач.

В настоящей статье представлены базовые решения по обработке неструктурированных геологических документов-отчетов на русском языке по реальным месторождениям нефти в Республике Татарстан, которые предназначены для цифровой научной платформы «Агрегатор неструктурированных геолого-промысловых данных». Эта платформа ставит целью формирование распределенной базы данных по нефтегазовым месторождениям на основе информации, извлекаемой из отчетов, подготовленных по месторождениям за многолетний период.

Задачи проекта цифровой научной платформы отличает сложный уровень информационных материалов (реальные отчеты о месторождениях), а также состав и уровень детализации извлекаемой информации (набор параметров объектов и их значений). Новизна предлагаемых базовых решений по обработке отчетов о месторождениях заключается в построении методов аннотирования основных именованных сущностей, выделяемых в геологических отчетах, а также извлечения из отчетов параметров выделенных сущностей, удовлетворяемых определенным ограничениям.

Создание интегрированной распределенной базы геолого-промысловых данных на основе неструктурированной информации является важной задачей, решение которой позволит в дальнейшем применять для анализа полученных данных методы машинного обучения и переходить на новый уровень автоматизации в нефтегазовой сфере. Информация из такой базы может быть эффективно использована в поисковых запросах, аналитических исследованиях по сравнению месторождений, рекомендательных системах и других интеллектуальных приложениях.

Архитектура цифровой научной платформы

Архитектура цифровой научной платформы «Агрегатор неструктурированных геолого-промысловых данных» разработана в виде набора микросервисов, что позволяет поддерживать набор независимых и слабосвязанных сервисов, которые можно создавать, используя различные языки программирования и технологии хранения данных (рис. 1).

К числу разработанных сервисов относятся:

1. программный сервис сбора, предназначенный для сбора текстов отчетов, представленных в различных форматах (PDF, PPT, Word, Excel, XML/JSON);
2. подсистема извлечения из текстов значимой информации по месторождениям;
3. подсистема хранения, предназначенная для интеграции обработанной информации и организации базы данных;
4. программный сервис доступа, предназначенный для предоставления доступа к сформированной базе данных пользователям (через интерфейс пользователя) и прикладным программам (через API).

На вход системы поступают сырые неструктурированные данные в виде текстовых документов в разных форматах (PDF, DOC, DOCX), из которых с помощью различных методов обработки текстов извлекаются структурированные геолого-промысловые данные о месторождениях и их заранее заданных характеристиках. Извлеченные данные сохраняются в базе данных и в текстовом документе в формате XML и могут быть использованы в разнообразных приложениях (генерация сводных таблиц, аналитика данных, обработка поисковых запросов).

Сравнение структурированных и неструктурированных данных представлено на рис. 2.

Далее опишем программные решения разработанных сервисов цифровой научной платформы.

Программный сервис сбора и предобработки исходных данных

В настоящее время собрана большая коллекция отчетов по месторождениям нефти в Республике Татарстан (РТ). Такая коллекция представлена в основном документами в формате PDF, поэтому на первом этапе стек программных решений включает задачу конвертации исходных отчетов в формате PDF в формат XML. Выбор формата XML определяется рядом преимуществ. XML является расширяемым видом языка разметки (markup language) и позволяет создавать структуру представления документа (XML-файл) в виде дерева элементов, которые удобно использовать в целях определения объектов текста

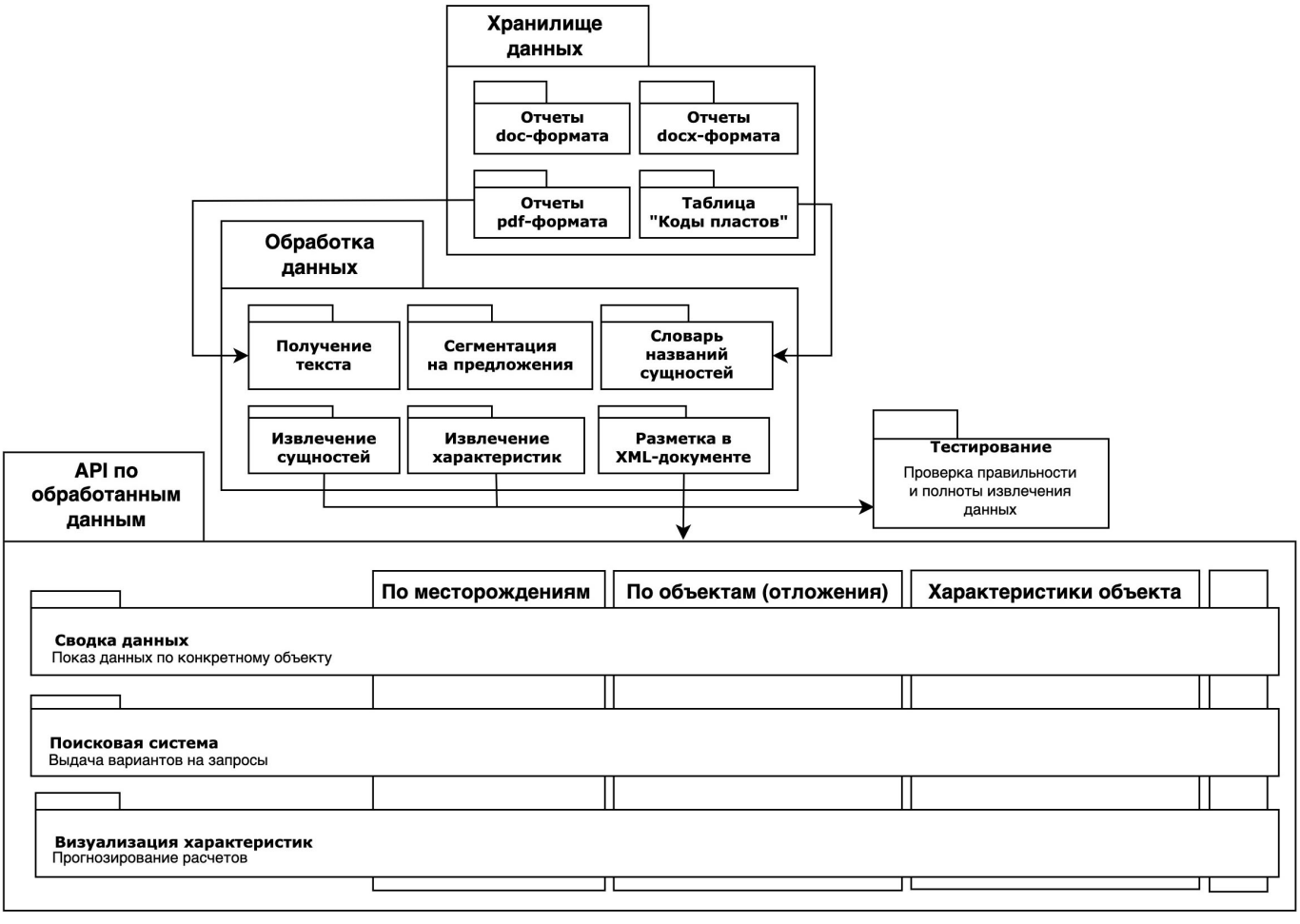

Рис. 1. Стек программных решений цифровой научной платформы

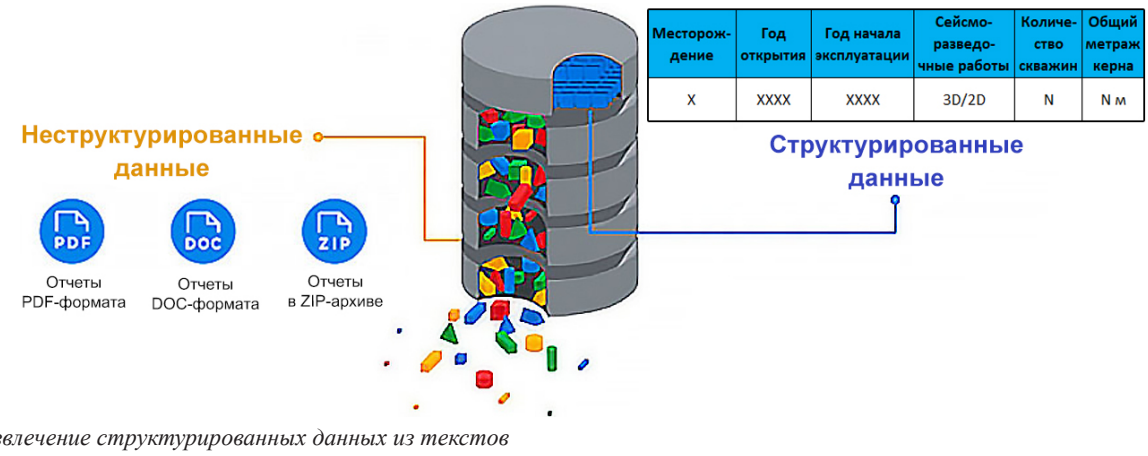

Рис. 2. Извлечение структурированных данных из текстов

и их атрибутов. Данный формат файла позволяет хранить самые разнообразные виды информации, предназначен для обмена данными между программными комплексами на различных платформах и имеет поддержку языка запросов. Для решения задачи конвертации разработан специальный алгоритм, реализованный с помощью библиотеки pdfplumber языка программирования Python.

Алгоритм конвертации использует набор специальных XML-тегов для кодирования извлекаемой из текстов отчетов информации. Введены структурные теги для выделения структурных элементов документа (глава, раздел, предложение), а также семантические теги для разметки названий месторождений и всех извлекаемых атрибутов месторождений (всего 24 тега). На рис. 3 представлено типовое XML-дерево документа отчета о месторождении.

Предобработка текстов отчетов дополнительно включала задачи замены сокращений на полные именования (например, замены «скв.» на «скважина», «мест.» на «месторождение» и др.), перевод словесной формы записи чисел в цифровую, сегментацию предложений текста с учетом их синтаксической структуры.

Подсистема извлечения значимой информации из текста отчета

Основной подсистемой платформы является подсистема извлечения значимой информации из текстов отчетов

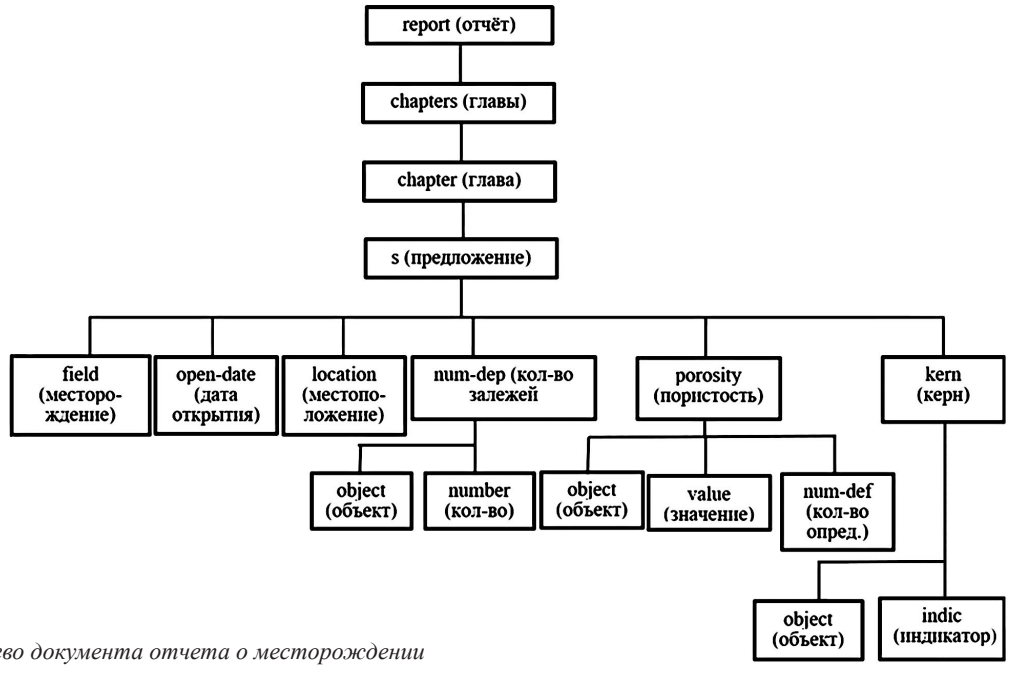

Рис. 3. XML-дерево документа отчета о месторождении

по месторождениям нефти для последующей передачи извлеченных данных в подсистему хранения (базу данных). Для решения поставленной задачи рассматривались различные технологии обработки текстов (современные технологии NLP – Natural Language Processing), включая семантико-лингвистические технологии и методы машинного обучения (в том числе нейросетевые модели). На данном этапе разработки при отсутствии размеченных наборов геологических данных наиболее эффективными являются семантико-лингвистические технологии обработки текстов. На последующих этапах при накоплении данных эффективно использовать нейросетевые модели в задачах классификации, кластеризации и прогнозирования.

Алгоритмы подсистемы извлечения значимой информации из текстов отчетов используют ряд внешних ресурсов, среди которых первостепенное значение имеет таблица стратиграфических кодов пластов. Применение этой таблицы в алгоритмах обработки потребовало внесение ряда уточнений, необходимых для машинной обработки данных, был выполнен парсинг таблицы для создания машиночитаемого классификатора кодов геологических объектов. Данный классификатор позволяет однозначно определять любой объект месторождения (горизонт, ярус, пласт) в тексте отчета не только по наименованию, но и по региональному и общероссийскому коду. Классификатор позволяет эффективно устанавливать идентичность объектов месторождения, в случае упоминания их в одном фрагменте текста как по названию, так и по коду.

Комплекс основных алгоритмов опирается на разработанную общую структурную схему месторождения, в которой выделены как общие параметры (название, дата открытия, дата начала эксплуатация и др.), так и группа параметров, характеризующих объекты месторождения (количество залежей, опробование, керн, пористость, неф-тенасыщенность и др.). В текущей версии программной системы извлекаются 7 целевых параметров по месторождению и 10 целевых параметров по каждому объекту месторождения.

Пример результатов извлеченных объектов месторождений и их целевых параметров представлен в табл. 1.

Извлечение именованных сущностей и значений их характеристик для месторождения в целом и всех его объектов выполнено на основе современных методов обработки текстов, обеспечивающих максимальную точность результата извлечения данных. Разработанные методы относятся к группе семантико-лингвистических методов и используют в своей основе разработанные шаблоны нечетких правил, позволяющих извлекать и анализировать контексты параметров месторождения и объектов. Найденные результаты сохраняются в виде атрибутов специальных XML-тегов, что позволяет в дальнейшем применять поисковые запросы к XML-документу отчета для генерации таблицы результатов обработки отчета.

Разработанные алгоритмы используют сложный комплекс моделей и методов, таких как модели автоматического разрешения кореферентности выделенных описаний (разрешение ссылок к одним и тем же объектам, заданным различными текстовыми метками), методы преобразования фрагментов текста к числовой форме, распознавание в тексте таблиц целевых параметров и их табличных значений, методы анализа контекста целевых параметров и методы расширения контекста при частичном его задании в тексте.

Обобщенная схема основного алгоритма выделения именованных сущностей (далее целевой показатель – ЦП) представлена на рис. 4.

Обобщенная схема основного алгоритма выделения целевого показателя включает следующие этапы.

1. Выбор целевого показателя для распознавания в тексте отчета его значения. В качестве целевых показателей выступают все выделяемые характеристики из списка именованных сущностей, приведенного выше.

2. Настройка шаблона распознавания целевого показателя, посредством которого осуществляется извлечение значения целевого показателя из отчета.

Месторождение	Год открытия	Год начала эксплуатации	Местоположение	Сейсморазведочные работы
Архангельское месторождение	1974 год	1978 год	Территория Новошешминского района Республика Татарстан	**2D**

Источники	Архангельское_месторождение_Пересчет_запасов_КГ
	Кн.2_Отчёт Архангельское_ПЕЧАТЬ

Объекты	Кол-во залежей	Пористость, %	Керн	Нефте-насыщенность, %	Кин, д. Ед.	$К_{выт}$, д. Ед.	Кол-во опробованных скважин	Нефть	Нефть с водой	Вода	Приток отсутствует	Дебит нефти, т\сут
Шешминский горизонт	4	29,5	да	66,1	0,36	0,682	3	0	3	0	0	0,03-0,9
Каширский горизонт	8	18,7	да	81,6	0,25	0,35	2	0	0	2	0	0
Верейский горизонт	12	15,9	да	73,4	0,25	0,403	201	190	11	0	1	0,2-8,9
Башкирский ярус	9	13,5	да	79,7	0,291 долей ед. 0,210 долей ед. 0,250 долей ед.	0,464 доли ед. СВН – 0,300 доли ед .	268	201	63	3	1	0,009-21,3
Алексинский горизонт	20	12,4	да	70,1	0,27	0,4	36	34	2	0	0	0,1-22,7
Тульский горизонт	7	24,1	да	86,5	0,465	0,536	369	316	33	19	1	0,06-54,0
Бобриковский горизонт	16	20,6	да	84,6	0,401	0,473	19	16	0	2	1	0,8-8,7
Турнейский ярус	13	12,6	да	68,9	0,25	0,4	36	27	7	2	0	0,2-13,1
Кыновско-пашийский	1		да		0,316	0,535	2	1	0	1	0	4

Табл. 1. Набор выделяемых именованных сущностей

3. Поиск раздела отчета по поисковому индексу. Алгоритм автоматического создания поискового индекса по оглавлению документа отчета формирует инвертированный индекс, связывая целевой показатель и главу отчета, в названии которой используется название целевого показателя, или основывается на полученных ранее данных по другим обработанным отчетам. Таким образом, поисковый индекс позволяет найти наиболее вероятный локальный фрагмент отчета, в котором содержится требуемая информация по целевому показателю. Поскольку построение поискового индекса опирается на оглавление отчета, то чем выше структурированность изложения информации в отчете, тем выше эффективность применения построенного инвертированного поискового индекса.

4. Распознавание целевого показателя по шаблону. Этот этап является ключевым этапом алгоритма и подробно рассмотрен на рис. 5. Найденные результаты сохраняются в виде атрибутов специальных XML-тегов, что позволяет в дальнейшем использовать поисковые запросы к XML-документу отчета для генерации итоговой таблицы результатов обработки отчета. На этапе 4 формируется множество потенциальных кандидатов-предложений, удовлетворяющих условиям шаблона распознавания целевого показателя.

5. Выбор наилучшего результата. На данном этапе осуществляется выбор наилучшего результата на основании критерия, установленного в шаблоне.

6. Занесение наилучшего результата по целевому показателю в итоговую таблицу. Результаты обработки отчета по месторождению в подсистеме извлечения именованных сущностей выдаются в структурированном виде в формате xls-таблицы.

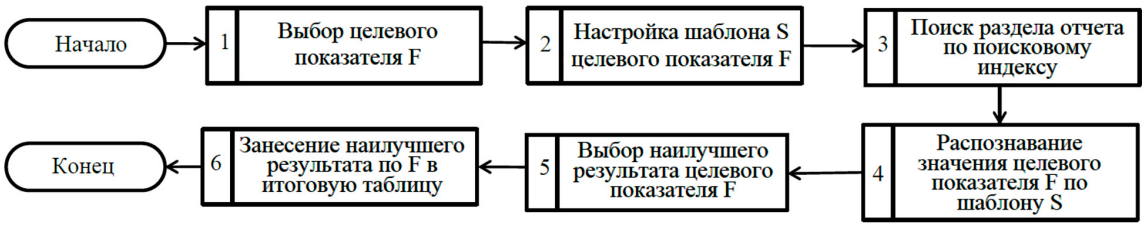

Рис. 4. Обобщенная схема основного алгоритма выделения целевого показателя. Использованные обозначения: S – шаблон, F – целевой показатель, ID – уникальный идентификатор номера предложения в отчете, k – счетчик цикла, P[k] – предложение с номером ID = k, OBJ – показатель «Объект» в шаблоне S

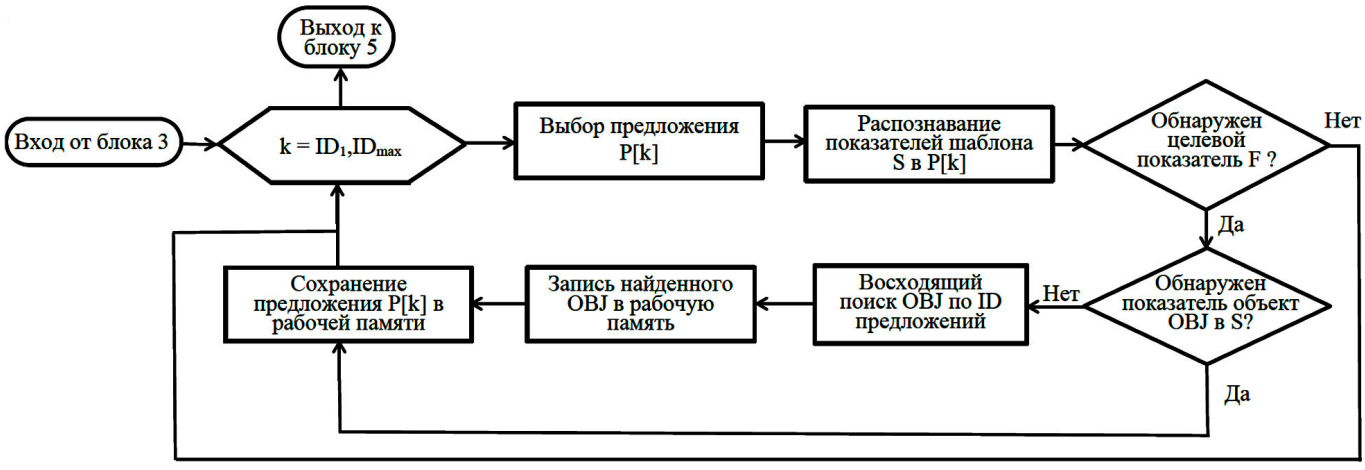

Рис. 5. 4 этап. Алгоритм распознавания целевого показателя объекта по шаблону. Обозначения см. на рис. 4

Модель шаблона для алгоритма распознавания целевого показателя

Алгоритм распознавания целевого показателя использует для распознавания модель шаблона в виде фрейма, слоты которого содержат распознаваемые в тексте отчета характеристики целевого показателя. Шаблон определяет структуру, задающую все основные элементы описания контекста целевого показателя. При этом для конкретного алгоритма требуется специальная настройка общего шаблона, при которой слоты получают конкретные значения, характерные для настраиваемого целевого показателя. Общая схема фрейма шаблона (фрейма-прототипа) приведена в табл. 2.

Слоты фрейма шаблон-прототип содержат характеристики, которые алгоритму требуется выделить в тексте при распознавании значения целевого показателя. В качестве целевых показателей выступают любые именованные сущности или их характеристики, представленные в табл. 2. Для распознавания значения в тексте целевого показателя алгоритм автоматически устанавливает раздел отчета по поисковому индексу, где вероятнее всего содержится требуемая информация, отнесенность целевого показателя к объекту месторождения, значение целевого показателя в заданных единицах измерения с указанием функции оценки и, возможно, критерия оценивания.

Имя слота	Обязательность	Значение слота
Целевой показатель	*	Название
Раздел отчета	*	Название
Ярус/Горизонт	*	Название/Код
Объект	*	Название/Код
Функция оценки целевого показателя	*	Список функций (средневзвешенная, средняя, …)
Единицы измерения ЦП	*	%/Доли ед./Название
Критерий отбора	-	Название
Значение ЦП	*	Число
Контекст ЦП	*	Лексемы в предложении с ЦП

*Табл. 2. Общая схема фрейма шаблона распознавания целевого показателя. * обязателен; – необязателен*

Дополнительно могут быть установлены специальные лексемы, наличие которых в предложении позволяет более точно распознавать требуемый контекст целевого показателя.

В качестве результата работы алгоритма распознавания в тексте целевого показателя «пористость» приведем фрейм-экземпляр данного целевого показателя, заполненный значениями из приведенного ниже фрагмента текста (табл. 3).

Имя слота	Обязательность	Значение слота
Целевой показатель	*	пористость
Раздел отчета	*	ГИС
Ярус/Горизонт	*	горизонт
Объект	*	верейский
Функция оценки целевого показателя	*	среднее
Единицы измерения ЦП	*	%
Критерий отбора	-	75 определений
Значение ЦП	*	16,4
Контекст ЦП	*	«значение»

*Табл. 3. Фрейм-экземпляр целевого показателя «пористость». * обязателен; – необязателен*

Тестирование и оценивание разработанных алгоритмов

Тестирование разработанных алгоритмов проводилось на 8 отчетах различных месторождений Республики Татарстан. Средний размер отчета составляет 100–500 страниц.

В тестировании оценивалась точность алгоритмов извлечения значений целевых показателей по микроусредненным (P_1) и макроусредненным (P_2) значениям точности:

$$P_1 = \frac{\sum_{i=1}^{m}\sum_{j=1}^{n_i} k_{ij}}{\sum_{i=1}^{m} n_i},\ P_2 = \frac{\sum_{i=1}^{m}\frac{\sum_{j=1}^{n_i} k_{ij}}{n_i}}{m},$$

где m – количество месторождений; n_i – количество объектов, имеющих целевой показатель в i месторождении; k_{ij} – число правильно извлеченных значений целевого показателя (1 – true, 0 – false).

	Микроусредненная точность	Макроусредненная точность
Месторождение	1	1
Год открытия	1	1
Год эксплуатации	1	1
Местоположение	1	1
Количество залежей нефти	0,85	0,85
Пористость	0,89	0,9
Керн	1	1
Нефтенасыщенность	0,85	0,85
Кин	0,96	0,98
$К_{выт}$	0,96	0,98

Табл. 4. Результаты тестирования по целевым показателям

Рассчитанные метрики по целевым показателям представлены в табл. 4.

Заключение

В статье представлены результаты по разработке архитектуры цифровой научной платформы «Агрегатор неструктурированных геолого-промысловых данных» и базовых моделей извлечения данных из неструктурированных текстов отчетов по месторождениям Республики Татарстан.

Разработанные модели реализованы в программной системе и протестированы на реальных отчетах по различным месторождениям. Проведенное тестирование показало высокие (85–100%) оценки точности извлекаемых целевых показателей (объектах месторождений и их характеристиках), что соответствует результатам, полученным для других языков (Lucas P. Cinelli et al., 2021; Nooralahzadeh et al., 2018; Qiu et al., 2019; Consoli et al., 2020). На основе разработанной программной системы формируется сводная база данных по месторождениям, результаты которой в дальнейшем могут быть использованы в различной аналитике с применением методов машинного обучения и нейросетевого анализа больших данных. Интегрирование и обобщение разнородной информации об объектах исследования недр позволят на основании полученных данных принимать обоснованные решения и переходить на новый уровень автоматизации в нефтегазовой сфере.

Финансирование

Работа выполнена при поддержке Министерства науки и высшего образования Российской Федерации по соглашению № 075-15-2022-299 в рамках программы создания и развития НЦМУ «Рациональное освоение запасов жидких углеводородов планеты».

Литература

Дежина И.Г., Мясников А.В., Коротеев Д.А. и др. (2017). Актуальные технологические направления в разработке и добыче нефти и газа: публичный аналитический доклад. М.: БиТуБи, 220 с.

Abdelhamid K., Ammar T.B., Laid K. (2022). Artificial Intelligent in Upstream Oil and Gas Industry: A Review of Applications, Challenges and Perspectives. *Artificial Intelligence and Its Applications. AIAP 2021. Lecture Notes in Networks and Systems*, vol. 413. Lejdel B., Clementini E., Alarabi L. (eds). Springer, Cham. https://doi.org/10.1007/978-3-030-96311-8_2424

Consoli B., Santos J., Gomes D., Cordeiro F., Vieira R., Moreira V. (2020). Embeddings for Named Entity Recognition in Geoscience Portuguese Literature. *Proc. 12th Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, 11–16 May 2020. https://aclanthology.org/2020.lrec-1.568.pdf

Chengbin Wang, Yuanjun Li, Jianguo Chen, Xiaogang Ma (2023). Named entity annotation schema for geological literature mining in the domain of porphyry copper deposits. *Ore Geology Reviews*, 152, 105243. https://doi.org/10.1016/j.oregeorev.2022.105243

Choubey S., Karmakar G.P. (2021). Artificial intelligence techniques and their application in oil and gas industry. *Artif Intell Rev,* 54, pp. 3665–3683. https://doi.org/10.1007/s10462-020-09935-1

Deloitte Analysis Report (2019). Digital transformation of oil and gas sector. https://www.petrotech.in/static/pdf/Theme-Session-Deloitte.pdf

Goodfellow I., Bengio Y., Courville A. (2016). Deep learning. Cambridge, MA: MIT Press.

Hoffimann Julio, Mao Youli, Wesley Avinash, Aimee Taylor (2018). Sequence Mining and Pattern Analysis in Drilling Reports with Deep Natural Language Processing. *SPE Annual Technical Conference and Exhibition*, Dallas, Texas, USA, September 2018. https://doi.org/10.2118/191505-MS

Lucas P. Cinelli, José F.L. de Oliveira, Vinicius M. de Pinho et al. (2021). Automatic event identification and extraction from daily drilling reports using an expert system and artificial intelligence. *Journal of Petroleum Science and Engineering*, 205. https://doi.org/10.1016/j.petrol.2021.108939

Nooralahzadeh F., Øvrelid L., Lønning J.T. (2018). Evaluation of domain-specific word embeddings using knowledge resources. *Proc. 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 1438–1445. http://www.lrec-conf.org/proceedings/lrec2018/pdf/268.pdf

Qiu Q., Xie Z., Wu L., Tao L., and Li W. (2019). Bilstm-crf for geological named entity recognition from the geoscience literature. *Earth Science Informatics*, 12, pp. 565–579. https://doi.org/10.1007/s12145-019-00390-3

Technavio (2015). How oil and gas is using Big Data for better operations. http://www.technavio.com/blog/how-oil-and-gas-using-big-data-better-operations

Сведения об авторах

Ольга Авенировна Невзорова – доцент, кандидат тех. наук, старший научный сотрудник, Казанский (Приволжский) федеральный университет

Россия, 420008, Казань, ул. Кремлевская, д. 18, корп.1

Рустем Рафаилевич Хакимуллин – лаборант, Казанский (Приволжский) федеральный университет

Россия, 420008, Казань, ул. Кремлевская, д. 18, корп.1

Ильяс Ирекович Идрисов – научный сотрудник, Казанский (Приволжский) федеральный университет

Россия, 420008, Казань, ул. Кремлевская, д. 18, корп.1

e-mail: ilyas_irekovich@mail.ru

Статья поступила в редакцию 13.09.2023;
Принята к публикации 29.09.2023; Опубликована 30.12.2023

IN ENGLISH

REVIEW ARTICLE

Digital scientific platform “Aggregator of unstructured geological and field data”: architecture and basic models of data extraction

*O.A. Nevzorova, R.R. Khakimullin, I.I. Idrisov**

Kazan Federal University, Kazan, Russian Federation

**Corresponding author: Ilyas I. Idrisov, e-mail: ilyas_irekovich@mail.ru*

Abstract. The article describes the project being developed for the digital scientific platform "Aggregator of unstructured geological and field data", which could potentially be important for the oil and gas industry. The use of new intelligent technologies within the framework of this project will significantly improve the efficiency of processing, storage and use of geological and field information contained in various text sources, mainly in field reports.

The main goal of developing a digital scientific platform is to integrate heterogeneous information about the objects of subsurface exploration, which is extracted from reports on deposits of the Republic of Tatarstan. This will create a consolidated database that will become the basis for making informed decisions in the oil and gas sector. The project of the digital scientific platform includes the development of architecture, algorithms and software solutions based on modern methods of text processing and data mining.

Keywords: data collection and analysis, field reports, database, automation, big data, text data processing, unstructured data, information extraction

Recommended citation: Nevzorova O.A., Khakimullin R.R., Idrisov I.I. (2023). Digital scientific platform "Aggregator of unstructured geological and field data": architecture and basic models of data extraction. *Georesursy = Georesources*, 25(4), pp. 149–156. https://doi.org/10.18599/grs.2023.4.13

Acknowledgements

This work was supported by the Ministry of Science and Higher Education of the Russian Federation under agreement No. 075-15-2022-299 as part of the program for the creation and development of the NCMU "Rational development of liquid hydrocarbon reserves of the planet".

References

Abdelhamid K., Ammar T.B., Laid K. (2022). Artificial Intelligent in Upstream Oil and Gas Industry: A Review of Applications, Challenges and Perspectives. *Artificial Intelligence and Its Applications. AIAP 2021. Lecture Notes in Networks and Systems*, vol. 413. Lejdel B., Clementini E., Alarabi L. (eds). Springer, Cham. https://doi.org/10.1007/978-3-030-96311-8_2424

Consoli B., Santos J., Gomes D., Cordeiro F., Vieira R., Moreira V. (2020). Embeddings for Named Entity Recognition in Geoscience Portuguese Literature. *Proc. 12th Conference on Language Resources and Evaluation (LREC 2020)*, Marseille, 11–16 May 2020. https://aclanthology.org/2020.lrec-1.568.pdf

Chengbin Wang, Yuanjun Li, Jianguo Chen, Xiaogang Ma (2023). Named entity annotation schema for geological literature mining in the domain of porphyry copper deposits. *Ore Geology Reviews*, 152, 105243. https://doi.org/10.1016/j.oregeorev.2022.105243

Choubey S., Karmakar G.P. (2021). Artificial intelligence techniques and their application in oil and gas industry. *Artif Intell Rev,* 54, pp. 3665–3683. https://doi.org/10.1007/s10462-020-09935-1

Deloitte Analysis Report (2019). Digital transformation of oil and gas sector. https://www.petrotech.in/static/pdf/Theme-Session-Deloitte.pdf

Dezhina I.G., Myasnikov A.V., Koroteev D.A. et al. (2017). Current technological trends in the development and production of oil and gas: public analytical report. Moscow: BiTuBi, 220 p. (In Russ.)

Goodfellow I., Bengio Y., Courville A. (2016). Deep learning. Cambridge, MA: MIT Press.

Hoffimann Julio, Mao Youli, Wesley Avinash, Aimee Taylor (2018). Sequence Mining and Pattern Analysis in Drilling Reports with Deep Natural Language Processing. *SPE Annual Technical Conference and Exhibition*, Dallas, Texas, USA, September 2018. https://doi.org/10.2118/191505-MS

Lucas P. Cinelli, José F.L. de Oliveira, Vinicius M. de Pinho et al. (2021). Automatic event identification and extraction from daily drilling reports using an expert system and artificial intelligence. *Journal of Petroleum Science and Engineering*, 205. https://doi.org/10.1016/j.petrol.2021.108939

Nooralahzadeh F., Øvrelid L., Lønning J.T. (2018). Evaluation of domain-specific word embeddings using knowledge resources. *Proc. 11th International Conference on Language Resources and Evaluation (LREC 2018)*, pp. 1438–1445. http://www.lrec-conf.org/proceedings/lrec2018/pdf/268.pdf

Qiu Q., Xie Z., Wu L., Tao L., and Li W. (2019). Bilstm-crf for geological named entity recognition from the geoscience literature. *Earth Science Informatics*, 12, pp. 565–579. https://doi.org/10.1007/s12145-019-00390-3

Technavio (2015). How oil and gas is using Big Data for better operations. http://www.technavio.com/blog/how-oil-and-gas-using-big-data-better-operations

About the Authors

Olga A. Nevzorova – Associate Professor, Cand. Sci. (Engineering), Senior Researcher, Kazan Federal University
18 Kremlevskaya st., Kazan, 420008, Russian Federation

Rustem R. Khakimullin – Laboratory Assistant, Kazan Federal University
18 Kremlevskaya st., Kazan, 420008, Russian Federation

Ilyas I. Idrisov – Researcher, Kazan Federal University
18 Kremlevskaya st., Kazan, 420008, Russian Federation
e-mail: ilyas_irekovich@mail.ru

Manuscript received 13 September 2023;
Accepted 29 September 2023; Published 30 December 2023