Об одном методе автоматической классификации полнотекстовых описаний кернов

А.П. Антонов[1], С.А. Афонин[2], А.С. Козицын[2], В.М. Староверов[1]

[1] Механико-математический факультет МГУ им.М.В. Ломоносова
[2] НИИ механики МГУ им.М.В. Ломоносова

Аннотация. Использование методов автоматической обработки текстов, в том числе, методов классификации полнотекстовых описаний, позволяет достичь существенного снижения трудозатрат при обработке экспериментальных данных. В настоящей работе рассматривается применение метода автоматической классификации текстов в области обработки и классификации элементов керна и определения литофаций. Литофациями называют одновозрастные геологические тела (отложения), которые по своему составу или строению отличаются от соседних слоев. При проведении оценки нефтегазового потенциала месторождений требуется проводить построение карт и схем распространения литофаций. Для этого необходимо осуществить классификацию большого количества выполненных специалистами полнотекстовых описаний участков керна. Представленный в статье алгоритм позволяет на основе заданных правил и словарей провести классификацию с учетом порядка и значимости ключевых слов в предложениях. Преимуществом такого подхода являются: возможность различать близкие литофации, возможность использования архивных данных, простота настройки на новые классы, адаптация к русскоязычным описаниям кернов и возможность локального использования без необходимости передавать описания кернов сторонним приложениям.

Ключевые слова: классификация текстов, литофации, словари, информационные системы

On one method of automatic classification of full-text core descriptions

A.P. Antonov[1], S. A. Afonin[2], A.S. Kozitsyn[2], V.M. Staroverov[1]

[1] Faculty of Mechanics and Mathematics, Lomonosov Moscow State University
[2] Institute of Mechanics, Lomonosov Moscow State University

Abstract. The use of automatic text processing methods, including full-text description classification methods, allows achieving a significant reduction in

labor costs when processing experimental data. This paper discusses the use of the automatic text classification method in the field of processing and classifying core elements and determining lithofacies. Lithofacies are coeval geological bodies (deposits) that differ in composition or structure from adjacent layers. When assessing the oil and gas potential of fields, it is necessary to construct maps and diagrams of lithofacies distribution. This requires classifying a large number of full-text descriptions of core sections prepared by specialists. The algorithm presented in the article allows, based on specified rules and dictionaries, to conduct classification taking into account the order and significance of keywords in sentences. The advantages of this approach are: the ability to distinguish between close lithofacies (), the ability to use archival data, ease of adjustment to new classes, adaptation to Russian-language core descriptions and the possibility of local use without the need to transfer core descriptions to third-party applications.

Keywords: text classification, lithofacies, dictionaries, information systems

1. Введение

При проведении оценки нефтегазового потенциала месторождений требуется проводить построение карт и схем распространения литофаций [11]. Литофациями называют одновозрастные геологические тела (отложения), которые по своему составу или строению отличаются от соседних слоев. Классификация литофаций является двухуревневой. На первом уровне определяется название фации («фация морен», «фация аллювиальных конусов выноса», «фация речной поймы» и др.), на втором уровне определяется компонента лито- («алевро-глинистая», «углисто-алевролитовая», «битуминозно-кремнисто-глинистая» и др.). Одним из часто используемых методов построения карт распространения литофаций является анализ результатов бурения. Полученные в процессе бурения керны исследуются специалистами и описываются в свободном текстовом формате с разделением всей глубины керна на отдельные однородные участки. Результатом такого анализа является набор данных, включающий координаты и параметры скважины, глубину и полнотекстовое описание состава полученной в результате бурения пробы породы. На основе составленного текстового описания (литологического описания керна, содержащего, обычно, от одного до десяти предложений) необходимо сопоставить каждому участку керна один из заданных классов литофаций. Следует отметить, что в разных исследовательских группах и разных организациях существуют различные стандарты на классификаторы литофаций. В зависимости от предъявляемых требований и поставленных задач могут использоваться классификаторы, содержащих от 5-8 классов до сотни классов.

Автоматизация процесса проведения такой классификации позволяет значительно упростить и унифицировать обработку полнотекстовой информации, полученной от различных специалистов по десяткам тысяч

кернов за последние десятилетия при анализе описаний кернов со скважин исследуемого региона. Возможные ошибки в работе впоследствии могут корректироваться при построении карт за счет сглаживания данных на соседних участках.

2. Обзор существующих подходов

Автоматизация процесса определения литофаций керна в настоящий момент развивается по трем основным направлениям. Наибольшее количество работ посвящено автоматизации распознавании изображений шлифов. Здесь следует отметить такие инструментальные средства, как ИС ABAI[1], Сервис DeepCore копании Digital Petrelium [2], комплекс DHD [3], Программный комплекс «Цифровой керн» [4]. Подобные решения позволяют существенно сократить трудозатраты при анализе данных бурения, поскольку не требуют работы специалистов по составлению описания. Однако, такая технология не применима для уже накопленного экспериментального материала. Кроме того, количество выделяемых классов фиксировано для каждой системы и оказывается существенно меньшим, чем при традиционных методах обработки текстовых данных.

В ряде работ предлагается проводить классификацию литофаций по числовым характеристикам и измеряемым физическим свойствам пород, например, в работе [5] сравниваются результаты классификации пород на 4 класса в соответствии со значением функций давления и внутреннего трения между частицами породы. Подобные методы также неприменимы для классификации по большим классификаторам, содержащим десятки классов.

В ряде работ для распознавания и классификации полнотекстовых описаний предлагается использовать современные системы с искусственным интеллектом. Например, в работе [6] описывается метод классификации описаний с использованием векторных моделей текстов и нейронных сетей. Для проведения анализа полнотекстовых описаний кернов авторы преобразуют текст в векторное описание в модели GeoVec[7]. Эта модель построена на основе модели GloVe[8], обученной на текстах 280 тысяч статей по геологии и страницах википедии. На основе построенной модели авторы вычисляют для каждого предложения усреднение вектора его слов в модели GeoVec, которые подают на вход обученной нейронной сети. Классификация текстовых описаний кернов производилась по 18-ти классам литофаций.

Основным недостатком подобного подхода является отсутствие учета порядка слов в предложении. Поскольку для обучения и анализа в качестве входных данных используется усредненный вектор слов, элементы описания «глины с вкраплениями песка» и «песок с вкраплением глин» становятся неразличимыми. Соответственно, такой подход не применим для проведения анализа с целью распределения текстов по детализированным

классификаторам литофаций, включающим такие классы как «глинисто-песчанные» и «песчанно-глинистые». Дополнительным ограничением применения машинного обучения на основе векторных моделей представления текстов является отсутствие предобученных моделей на русском языке, аналогичных GeoVec. Использование пословного перевода обученных моделей не дает удовлетворительного результата ввиду многозначности значительной части используемых в английском и русском языках слов.

В этой связи, для построения систем более точной классификации необходимо использование моделей и алгоритмов, учитывающих порядок слов в предложениях и адаптированных к русскому языку. Для этого можно использовать методы, которые используются в классических задачах тематического анализа по ключевым словам [9, 10].

3. Описание алгоритма

Разработанный авторами алгоритм опирается на использование составленных геологами словарей с описанием характеристик различных литофаций. На вход алгоритму поступает полнотекстовое описание участка керна, выполненное специалистом при анализе результатов бурения. Результатом работы алгоритма является ранжированный список возможных литофаций, которые соответствуют заданному текстовому описанию.

Настройка алгоритма на специфику предметной области производится при помощи формирования словаря описаний фаций. Словарь описания фаций формируется на основе описания возможных признаков фации, которые должны встречаться в описании (характерны для данной фации) или, наоборот, не могут встречаться в ее описании. Например, в описании фации русла рек не могут встречаться морские организмы. Каждый признак является словом или словосочетанием, и относится к определенному типу. В текущей программной реализации рассматривались следующие характеристики: название породы, ее цвет, структура, текстура, включения флоры и фауны, окатность, сортировка, границы. При необходимости список анализируемых характеристик может пополняться. В качестве словосочетаний рекомендуется использовать пары вида "существительное прилагательное" или "наречие причастие", например, "граница ровная" или "хорошо окатанные". Такой подход позволяет получить достаточно точные формализованные критерии, которыми пользуются геологи при решении аналогичной задачи в ручном режиме.

Для учета специфики формирования полнотекстовых описаний используются вспомогательные словари, которые не несут в себе информации о предметной области, но позволяют правильно расставлять акценты и определять значимость ключевых элементов описания. Словарь ослабляющих слов и выражений с глаголами (например, "изредка встречаются") определяет слова и выражения, после которых значимость

всех ключевых терминов уменьшается до конца предложения или глагола. Словарь усиливающих слов и выражений с глаголами (например, "превалирует") позволяет задавать усиление значимости всех ключевых терминов до конца предложения или глагола. Также используется словарь глаголов-исключений, который не прерывают действие ослабляющих и повышающих слов, словари синонимов и гиперонимов. Следует отметить разницу в обработке синонимов и гиперонимов. При сравнениях все синонимы считаются совпадающими терминами с учетом транзитивных зависимостей. Гипонимы и гиперонимы при обработке текстовых описаний считаются совпадающими терминами, но без учета транзитивных зависимостей.

Схема алгоритма представлена на рисунке 1.

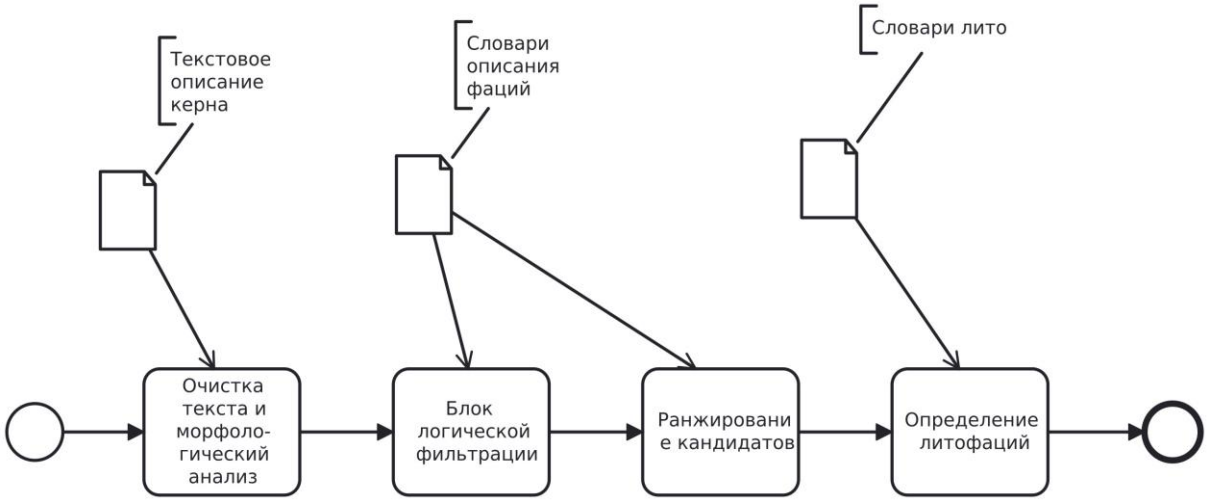

Рис. 1. Схема работы алгоритма классификации

На первом этапе работы алгоритма производится подготовка текста для анализа, включая очистку текста от лишних символов, разметка предложений, разметка областей, заключенных в скобочки, проводится морфологический анализ. Для каждого найденного термина (слова или словосочетания) указывается его относительная позиция в тексте, принадлежность словарям характеристик фаций, наличие повышения или понижения его значимости в соответствии с наличием повышающих ("преобладает", "преимущественно", :) и понижающих ("иногда бывать", "изредка", "иногда", :) слов. После повышающего или ослабляющего выражения до конца предложения, другого повышающего слова или глагола не из списка глаголов исключений все термины помечаются, соответственно, повышенными или ослабленными. Результатом работы этого этапа является структурированной описание анализируемого участка керна.

В блоке логической фильтрации осуществляется фильтрация фаций-кандидатов по запрещающим правилам. Для каждой фации в словарях

указывается, каких терминов из различных типов описаний в них не может встречаться. В случае, если хотя бы один из запрещенных для фации терминов встретился в анализируемом описании керна, указанная фация исключается из списка возможных кандидатов. Кроме того, фация удаляется из списка кандидатов, если в описании керна не встретилось ни одного термина из какого-либо значимого (не пустого и с коэффициентом значимости «1») описания данной фации. При анализе встречаемости терминов в описании фаций используются словари синонимов ("песок, песчаник") и гипонимов ("красный: светло-красный, темно-красный, алый").

В блоке ранжирования производится анализ списка всех найденных в описании керна терминов отдельно для каждого типа характеристики. Итоговый ранг фации вычисляется как сумма баллов, набранных фацией за соответствие каждого типа характеристики рассматриваемому описанию керна по следующей формуле.

$$css \cdot \sum_{i \in H} \sum_{w \in W_i} \begin{cases} dsk_i \cdot ss_{p_i} \cdot t_{iw} & w \in D \\ -k_2 \cdot dsk_i \cdot ssn_{p_i} \cdot t_{iw} & w \notin D \end{cases},$$

где

css - нормирующий коэффициент фации, зависящий от количества описанных типов характеристик для рассматриваемой фации,

dsk_i - коэффициент значимости типа признака i в расчете ранга,

ss_j - коэффициент учета в ранге найденных в описании слов из признака со значимостью j.

ssn_j - коэффициент учета в ранге не найденных в описании слов из признака со значимостью j.

p_i - заданная экспертом значимость признака i для рассматриваемой фации. Принимает значение 1,2,3 (1-наиболее важное, 3 –наименее важное).

k_2 - размер штрафа за ненайденное слово

$t_{iw} = clf_i \cdot twi(w)$ - вес термина w_i

clf_i -способ учет длины описания типа признака i в рассматриваемой фации, в зависимости от значения параметра обучения либо 1, либо $1/n_i$,

n_i -количество терминов в описании типа признака i для фации,

$twi(w)$ - функция учета idf для слова w.

Конкретные значения для коэффициентов определяются на этапе обучения модели.

На последнем этапе для каждой найденной фации выбирается название лито из заданного экспертом в словаре списка возможных значений. Основным принципом построения названия лито является повышение приоритета слова от начала к концу термина. Например, название лито "углисто-алевро-глинистая" означает, что его основу составляют глины, в меньшей степени встречаются алевролитовые слои и

совсем небольшом количестве в образце присутствуют угольные вкрапления. Для поиска подходящего названия лито из описания керна выделяются все термины, сортируются с учетом позиции текста и значимости (значимые передвигаются вперед, незначимые - назад), после этого все термины с дефисом переворачиваются. Для каждого возможного лито определяется порядок встречаемости его терминов в получившемся списке. Описания лито, термины которых не встретились в списке, или встретились не в том порядке исключаются из рассмотрения. Среди оставшихся названий лито выделяется название, термины которого оказались ближе к началу списка. Результатом работы данного шага алгоритма является наиболее вероятное название лито для каждой рассматриваемой фации. Дополнительно в случае нахождения единственного возможного названия лито, или нескольких возможных вариантов лито к рангу фации, вычисленному на предыдущем шаге, добавляется дополнительный коэффициент.

Результатом алгоритма является ранжированный список фаций с указанием лито для каждой фации. Данный список может использоваться для автоматической (выбирается фация с наиболее высоким рангом) или автоматизированной (пользователю предлагается выбор из нескольких фаций, если их ранги близки) классификации.

3. Программная реализация и результаты

Программная реализация алгоритма выполнена на языке Python 3.11. Морфологический анализ осуществляется с использованием библиотеки pymorphy3. Также для работы используются библиотеки dataclass и numpy. Для описания словарей использовались файлы в текстовом формате. Для ускорения работы по исходным словарям строятся дополнительные индексные файлы.

Процесс обучения заключался в подборе значений следующих параметров (таб. 1). Обучение разработанной программной реализации алгоритма проводилось на основе обучающей выборки из 66 примеров. Для обучения использовался метод градиентного спуска. Метрика качества для проведения обучения рассчитывается по следующей формуле $L = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{k_i}$, где N – количество тестовых примеров, k_i – порядковый номер правильного ответа в полученном ранжированном списке для примера i. В случае наличия в ранжированном списке нескольких примеров с одинаковым рангом порядковый номер для них усредняется. Например для ранжированного (с указанием ранга) списка A(5.2), B(4.7), C(4.7), D(4.7)),E(1.6) при правильном ответе B значение k_i будет равняться 3 (средняя позиция B, C и D). В результате обучения на обучающей выборке было достигнуто значение метрики качества L=0.681.

Таблица 1. Параметры обучения алгоритма.

Параметр	Возможные значения	Оптимальное значение
Учитывать ли длину описания фации.	0 - нет, 1 - да.	0
Нормирующий коэффициент фации, зависящий от количества описанных типов характеристик для рассматриваемой фации.	0 - не учитывается, 1 - учитывается линейно; иначе с заданным основанием логарифма (например,)	10
Коэффициент учета в ранге найденных в описании слов признака со значимостью 1.	Число	1
Коэффициент учета в ранге найденных в описании слов признака со значимостью 2.	Число	0.9
Коэффициент учета в ранге найденных в описании слов признака со значимостью 3.	Число	0.8
Коэффициент учета в ранге не найденных в описании слов со значимостью 1.	Число	1
Коэффициент учета в ранге не найденных в описании слов со значимостью 2.	Число	2/3
Коэффициент учета в ранге не найденных в описании слов со значимостью 3.	Число	4/9
Коэффициент понижения значимости для "незначимых" слов.	Число	0
Размер штрафа за ненайденное слово	Число	1
Тип учета idf	Целое число: 0- не учитывается, 1- учитывается линейно, иначе учитывается как логарифм с указанным основанием	0
Коэффициент значимости типа признака в расчете ранга	Список значений коэффициентов для каждого типа признака	{"rock":2, "Color":0.6, "structures":1,

		"texture":1, "inclusion":1, "inclusionorg":1, "roundness":1, "sort":1, "border":1}
Добавочный коэффициент за единственное найденное название лито	число	3
Добавочный коэффициент за несколько возможных найденное названий лито	число	0.5

Подобранные значения параметров использовались для проведения тестирования результатов работы на тестовой выборке описаний кернов.

Тестирование проводилось на 58 примерах. Результаты тестирования: L=0.576, на первом месте нужная фация при выборе из 42 фаций оказалась в 25 примерах, на втором в 7 примерах, на третьем в 8 примерах. Компонента лито была определена правильно в 46 примерах.

4. Заключение

Представленный алгоритм классификации полнотекстовых описаний кернов может использоваться для автоматизации процесса определения классов литофаций при построении литофационных карт, в том числе в разрабатываемых в настоящее время системах, которые должны заменить Petromod в национальных корпорациях. При обработке специалистом описаний кернов алгоритм подбирает наиболее вероятные классы, сокращая время разметки исходного материала. Преимуществом алгоритма являются: возможность обработки архивных данных и данных сторонних исследований, адаптация к русскому языку, возможность локального использования, возможность учета порядка слов в описаниях.

Литература

1. Информационная система ABAI — https://kmge.kz/abai/
2. Е. Е. Барабошкин, Е. А. Панченко, А. Е. Демидов и др., Система автоматического описания керна в производственном процессе. Опыт применения // Пути реализации нефтегазового потенциала Западной Сибири : Материалы XXV научно-практической конференции, Ханты-Мансийск, 23-26 ноября 2021 года / Под редакцией Э.А. Вторушиной, Е.Е. Оксенойд, С.А. Алёшина, Н.Н. Захарченко, Е.В. Олейник, Т.Н. Печёрина. - Ханты-Мансийск: Автономное учреждение Ханты-Мансийского автономного округа - Югры "Научно-аналитический

центр рационального недропользования им.В.И.Шпильмана", 2022. - С. 293-299.
3. Комплекс DHD — https://magazine.neftegaz.ru/articles/tsifrovizatsiya/682038-tsifrovoy-analiz-kerna-v-zadachakh-proektirovaniya-razrabotki-neftyanykh-i-gazovykh-mestorozhdeniy-/
4. Программный комплекс «Цифровой керн» — https://globalcio.ru/projects/10448/
5. Li H, Wan B, Chu D, Wang R, Ma G, Fu J, Xiao Z. Progressive Geological Modeling and Uncertainty Analysis Using Machine Learning. ISPRS International Journal of Geo-Information. 2023; 12(3):97. https://doi.org/10.3390/ijgi12030097
6. Ignacio Fuentes, José Padarian, Takuya Iwanaga, R. Willem Vervoort, 3D lithological mapping of borehole descriptions using word embeddings//Computers & Geosciences,Volume 141, 2020, 104516, DOI:https://doi.org/10.1016/j.cageo.2020.104516. URL: https://www.sciencedirect.com/science/article/pii/S0098300419306533
7. Padarian, J. and Fuentes, I.: Word embeddings for application in geosciences: development, evaluation, and examples of soil-related concepts, SOIL, 5, 2019 — P. 177-187, https://doi.org/10.5194/soil-5-177-2019, 2019. URL: https://soil.copernicus.org/articles/5/177/2019/
8. Pennington, J., Socher, R., and Manning, C.: Glove: Global vectors for word representation// Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014 — P. 1532-1543
9. Денисов, Д. В. Анализ методов машинного обучения для тематической классификации текстов// Международный журнал информационных технологий и энергоэффективности. - 2024. - Т. 9, № 4(42). - С. 5-11.
10. Козицын А. С. Алгоритмы тематического поиска данных в наукометрических системах // Программная инженерия. — 2022. — Т. 13, № 6. — С. 291–300.
11. Автоматизированное построение реалистичных литофациальных карт методами комбинаторной оптимизации / А. П. Антонов, С. А. Афонин, А. С. Козицын и др. // Интеллектуальные системы. Теория и приложения. - 2024. - Т. 28, № 4. - С. 5-20.

References

1. Informatsionnaia sistema ABAI — https://kmge.kz/abai/
2. E. E. Baraboshkin, E. A. Panchenko, A. E. Demidov i dr., Sistema avtomaticheskogo opisaniia kerna v proizvodstvennom protsesse. Opyt primeneniia // Puti realizatsii neftegazovogo potentsiala Zapadnoi Sibiri : Materialy XXV nauchno-prakticheskoi konferentsii, Khanty-Mansiisk, 23-26 noiabria 2021 goda / Pod redaktsiei E.A. Vtorushinoi, E.E. Oksenoid, S.A. Aleshina, N.N. Zakharchenko, E.V. Oleinik, T.N. Pecherina. - Khanty-

Mansiisk: Avtonomnoe uchrezhdenie Khanty-Mansiiskogo avtonomnogo okruga - Iugry "Nauchno-analiticheskii tsentr ratsionalnogo nedropolzovaniia im.V.I.Shpilmana", 2022. - P. 293-299.
3. Kompleks DHD — https://magazine.neftegaz.ru/articles/tsifrovizatsiya/682038-tsifrovoy-analiz-kerna-v-zadachakh-proektirovaniya-razrabotki-neftyanykh-i-gazovykh-mestorozhdeniy-/
4. Programmnyi kompleks «Tsifrovoi kern» — https://globalcio.ru/projects/10448/
5. Li H, Wan B, Chu D, Wang R, Ma G, Fu J, Xiao Z. Progressive Geological Modeling and Uncertainty Analysis Using Machine Learning. ISPRS International Journal of Geo-Information. 2023; 12(3):97. https://doi.org/10.3390/ijgi12030097
6. Ignacio Fuentes, José Padarian, Takuya Iwanaga, R. Willem Vervoort, 3D lithological mapping of borehole descriptions using word embeddings//Computers & Geosciences,Volume 141, 2020, 104516, DOI:https://doi.org/10.1016/j.cageo.2020.104516. URL: https://www.sciencedirect.com/science/article/pii/S0098300419306533
7. Padarian, J. and Fuentes, I.: Word embeddings for application in geosciences: development, evaluation, and examples of soil-related concepts, SOIL, 5, 2019 — P. 177-187, https://doi.org/10.5194/soil-5-177-2019, 2019. URL: https://soil.copernicus.org/articles/5/177/2019/
8. Pennington, J., Socher, R., and Manning, C.: Glove: Global vectors for word representation// Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014 — P. 1532-1543
9. Denisov, D. V. Analiz metodov mashinnogo obucheniia dlia tematicheskoi klassifikatsii tekstov// Mezhdunarodnyi zhurnal informatsionnykh tekhnologii i energoeffektivnosti. - 2024. - V. 9, № 4(42). - P. 5-11.
10. Kozitsyn A. S. Algoritmy tematicheskogo poiska dannykh v naukometricheskikh sistemakh // Programmnaia inzheneriia. — 2022. — V. 13, № 6. — P. 291–300.
11. Avtomatizirovannoe postroenie realistichnykh litofatsialnykh kart metodami kombinatornoi optimizatsii / A. P. Antonov, S. A. Afonin, A. S. Kozitsyn i dr. // Intellektualnye sistemy. Teoriia i prilozheniia. - 2024. - V. 28, № 4. - P. 5-20.