

Иванов Максим Сергеевич, Гурко Андрей Владимирович

Санкт-Петербургский Горный Университет императрицы Екатерины II

СИСТЕМА ГЕНЕРАЦИИ АНАЛИТИЧЕСКОГО КОНТЕНТА МИНЕРАЛЬНО-СЫРЬЕВОГО КОМПЛЕКСА

Аннотация: В статье рассматривается разработка системы генерации аналитического контента для научных исследований в области минерально-сырьевого комплекса. Основное назначение системы — автоматизация создания графического и текстового материала для исследований и отчетов с возможностью динамического прогнозирования данных и визуализации ключевых показателей. Система поддерживает генерацию многоформатных графиков и предсказание будущих значений на основе оптимального выбора метода математической регрессии. Контент может редактироваться пользователем и экспортироваться в виде PDF-документов, а доступ к статистическим данным обеспечивается через API.

Ключевые слова: Генерация контента, минерально-сырьевой комплекс, прогнозирование данных, математическая регрессия, визуализация, API, аналитика.

Annotation: The article discusses the development of a system for generating analytical content for scientific research in the field of the mineral resource complex. The main purpose of the system is to automate the creation of graphical and textual material for research and reports with the ability to dynamically predict data and visualize key indicators. The system supports the generation of multiformat graphs and the prediction of future values based on the optimal choice of mathematical regression method. The content can be edited by the user and exported as PDF documents, and access to statistical data is provided through the API.

Keywords: Content generation, mineral resource complex, data forecasting, mathematical regression, visualization, API, analytics.

Введение

Современные научные исследования требуют значительных временных затрат на подготовку аналитических отчетов, визуализацию данных и прогнозирование ключевых показателей. Особенно актуальной эта задача становится в сфере минерально-сырьевого комплекса, где работа с большими массивами данных требует эффективных инструментов обработки и представления информации [1]. Для решения данной задачи разрабатывается система генерации аналитического контента, которая позволяет автоматизировать процесс подготовки материалов для научных публикаций. Система поддерживает создание текстового контента и визуализацию данных в виде различных типов графиков: линейных, столбчатых, круговых, а также более сложных диаграмм - полярных, пузырьковых и смешанных [2]. Ключевой особенностью системы является возможность прогнозирования данных на основе различных методов регрессии. Предлагается применить метод автоматического выбора наилучшего метода прогнозирования (линейный, логарифмический, экспоненциальный, полиномиальный или степенной) на основе коэффициента детерминации [3].

Для реализации системы используются веб-технологии и библиотеки, обеспечивающие удобное взаимодействие с данными и их визуализацию

Основные этапы и методы разработки

Для построения базы данных на реляционной модели предполагается использование PostgreSQL [4]. Основные сущности системы: секции контента, текстовые блоки, темы, графики и прогнозируемые данные представлены на рис. 1 [5].

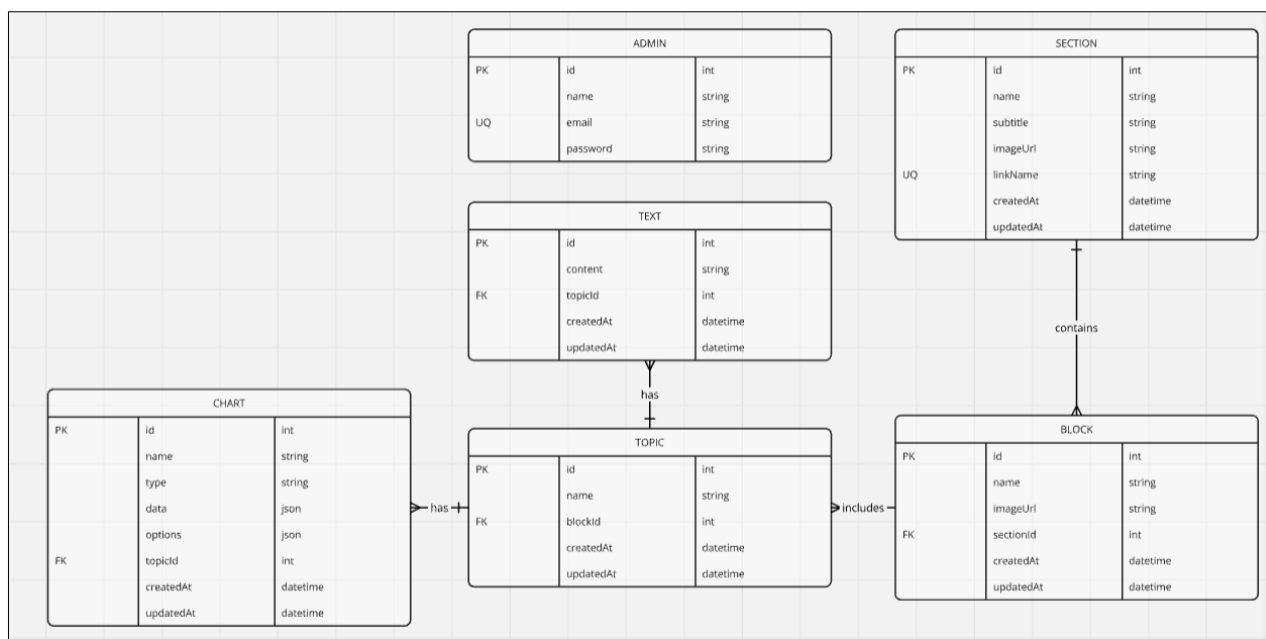


Рис. 1. ER-диаграмма базы данных [материал автора]

Для прогнозирования временных рядов применены методы регрессии с использованием библиотеки Regression.js. Библиотека Regression.js предоставляет несколько методов для выполнения регрессионного анализа. Основные методы, которые она поддерживает, включают линейную регрессию, полиномиальную регрессию, регрессию с использованием экспоненциальной функции и другие. Рассмотрим поддерживаемые данной библиотекой методы и покажем пример решения задачи выбора метода для построения прогноза:

1. Модель линейной регрессии имеет вид:

$$y = m \cdot x + b; (1)$$

y — результат, m —наклон, x — независимая переменная (входные данные), b — пересечение с осью y .

2. Модель полиномиальной регрессии:

$$y = a_n + a_{n-1}x^{n-1} + \dots + a_1 + a_0; (2)$$

где: n — степень полинома, a_i — коэффициенты полинома.

3. Модель экспоненциальной регрессии имеет вид:

$$y = a \cdot c^{bx}; (3)$$

a — начальное значение, b — скорость роста, e — основание логарифма.

4. Модель логарифмической регрессии:

$$y = a + b \cdot \ln(x); (4)$$

a — свободный член, b — коэффициент наклона, $\ln(x)$ —логарифм переменной.

5. Модель логистической регрессии использует:

$$y = \frac{1}{1+e^{-(a+b \cdot x)}}; (5)$$

где: a — свободный член, b — коэффициент наклона, e — основание натурального логарифма.

Для оценки адекватности моделей выбран **коэффициент детерминации** (R^2 , R-квадрат). Коэффициент показывает, какую долю вариации зависимой переменной можно объяснить с помощью независимых переменных в модели.

Коэффициент детерминации рассчитывается по следующей формуле:

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}; (6)$$

где: S_{res} (сумма квадратов остатков) — это сумма квадратов разностей между фактическими значениями и предсказанными значениями модели:

$$S_{res} = \sum (y_i - \hat{y}_i)^2; (7)$$

S_{tot} (общая сумма квадратов) — это сумма квадратов разностей между фактическими значениями и средним значением зависимой переменной:

$$S_{tot} = \sum (y_i - \bar{y})^2; (8)$$

Значение R^2 варьируется от 0 до 1: $R^2 = 0$: Модель не объясняет никакой вариации в данных. Это означает, что предсказания модели совпадают с средним значением зависимой переменной. $R^2 = 1$: Модель идеально объясняет все вариации в данных. Все предсказанные значения совпадают с фактическими значениями. Значения между 0 и 1: Чем ближе R^2 к 1, тем лучше модель объясняет данные. Коэффициент детерминации часто используется для оценки качества регрессионных моделей и помогает исследователям и аналитикам понять, насколько хорошо модель справляется с предсказанием зависимой переменной на основе независимых переменных. Эти формулы и примеры дают общее представление о методах регрессии, доступных в библиотеке Regression.js. Выбор метода зависит от характера данных и цели анализа. На каждом этапе вычисляется коэффициент детерминации (R^2), который помогает выбрать наиболее точную модель прогнозирования (см. рис. 2) [6].

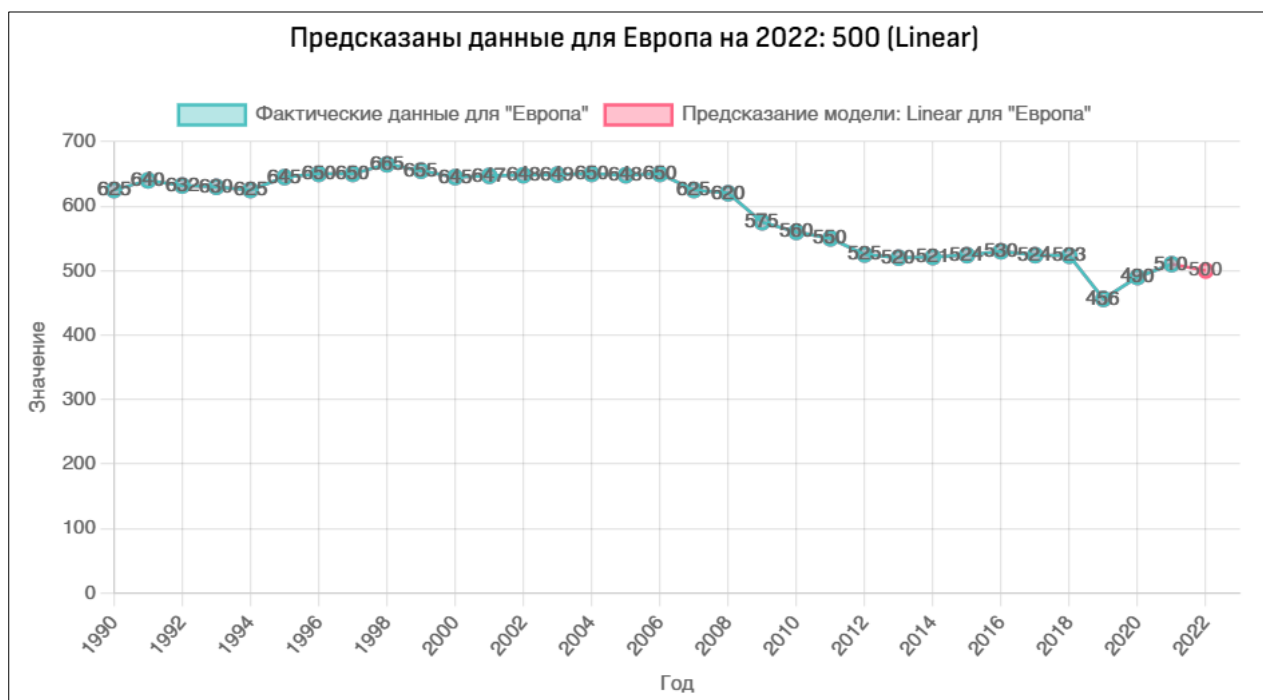


Рис. 2. Прогнозирование данных графиков выбором оптимального метода математической регрессии [материалы автора]

Разработка интерфейса пользователя: Веб-интерфейс создан с использованием Next.js (React), что обеспечивает высокую производительность и удобство работы для пользователей [7,8]. В интерфейсе предусмотрены модули управления контентом, создания и редактирования графиков, а также функции экспорта данных [9,10].

API для доступа к данным: Система предоставляет RESTful API для получения данных о контенте и статистике. Это позволяет интегрировать систему с внешними аналитическими инструментами. API поддерживает выгрузку данных в формате JSON, что облегчает их последующую обработку [11,12].

Рассмотрим пример применения метода рационального выбора прогнозирования с использованием линейной регрессии в Regression.js [13,14,15]:

Рассмотрим пример принятия решения о выборе метода прогнозирования на основе расчета коэффициента детерминации R^2 для линейной и полиномиальной регрессии. Используем небольшой набор данных и сравним результаты (см. табл. 1).

Таблица 1 [материалы автора]

Набор данных для принятия решения о выборе метода прогнозирования на основе вычисления коэффициента детерминации (R^2)

x	1	2	3	4	5
y	2	3	5	7	11

Пусть имеется следующая зависимость.

Покажем расчет коэффициентов математической модели и коэффициента детерминации, отвечающего за адекватность модели на примере линейной регрессии - $mx + b$; m — коэффициент наклона прямой, b — свободный член.

m и b — вычисляются по формулам

$$m = \frac{n(\sum xy) - (\sum x)(\sum y)}{n(\sum x^2) - (\sum x)^2}; \quad (9)$$

$$b = \sum (y_i - \bar{y})^2; \quad (10)$$

S_{res} и S_{tot} — вычисляются по формулам

$$S_{res} = \sum (y_i - \hat{y}_i)^2; \quad (11)$$

$$S_{tot} = \sum (y_i - \bar{y})^2; \quad (12)$$

расчетов мы получили:

$$R^2 = 0.92$$

В качестве второго варианта для сравнения проведем расчеты по уравнению второй степени: $y = ax^2 + bx + c$.

1. Найдем коэффициенты a , b и c с помощью метода наименьших квадратов.

2. Подсчитаем S_{res} и S_{tot} и определим:

$$R_{polynomial}^2 = 0.98$$

Теперь у нас есть два значения R^2 . Линейная регрессия: Значение $R^2 = 0.92$ говорит о том, что 92% вариации может быть объяснено линейной моделью. Полиномиальная регрессия: Значение $R^2 = 0.98$ указывает на то, что полиномиальная модель объясняет 98% вариации. Этот результат лучше, чем линейная модель. Однако, следует учитывать возможность переобучения и проверять модель на новых данных.

Заключение

Разрабатываемая система автоматизирует процесс подготовки аналитической информации. Она позволяет пользователям создавать текстовые отчеты и визуализировать данные в виде различных графиков. Механизм прогнозирования данных на основе методов регрессии, интегрированный в информационную систему поможет дополнять аналитику прогнозами на будущие периоды. Прогнозируемые значения помогают формировать дополнительные выводы для исследований, а функция экспорта Применение данных методов к системе генерации аналитического контента предоставляет исследователям и аналитикам эффективный инструмент для автоматизации подготовки отчетов и генерации другого контента. В системе сочетаются функции текстовой генерации, визуализации данных и прогнозирования ключевых показателей. Благодаря интеграции методов регрессии

пользователи могут оценивать будущие значения на основе существующих данных, что расширяет аналитические возможности системы и делает систему удобным инструментом для подготовки статей и отчетов.

Открытое API позволяет использовать систему как часть более крупных аналитических платформ, а возможность экспорта данных в PDF упрощает подготовку итоговых материалов. Дальнейшее развитие системы может включать расширение функций прогнозирования и интеграцию с другими статистическими библиотеками для повышения точности расчетов и удобства работы пользователей.

Метод имеет ограничения. Высокое значение R^2 не всегда говорит, что модель хороша. Оно может быть высоким даже в случае переобучения, когда модель сложна и начинает подстраиваться под шум в данных. А при добавлении новых независимых переменных R^2 может увеличиваться, даже если эти переменные не являются значимыми. Следует рекомендовать использовать скорректированный коэффициент детерминации R^2 , который учитывает количество независимых переменных.

Литература

1. Гурко, А. В. Финансовое положение предприятия и методы машинного обучения / А. В. Гурко, О. В. Шишкина // Экономика и управление в сфере услуг: современное состояние и перспективы развития : XX Всероссийская научно-практическая конференция, Санкт-Петербург, 07 февраля 2023 года. – Санкт-Петербург: Санкт-Петербургский гуманитарный университет профсоюзов, 2023. – С. 104-106. – EDN SIBPNN.

2. Панова, А. С. Предиктивная диагностика промышленного оборудования на основе распознавания технических звуков / А. С. Панова, Р. С. Федьков, А. В. Гурко // Анализ и прогнозирование систем управления в промышленности, на транспорте и в логистике : XXIII Международная научно-практическая конференция молодых учёных, студентов и аспирантов, Санкт-Петербург, 18-20 апреля 2023 года. – Санкт-Петербург: ООО "Медиапапир", 2023. – С. 133-138. – EDN GFZSNQ.

3. Свидетельство о государственной регистрации программы для ЭВМ № 2024682752 Российская Федерация. Программа для решения минимаксной обобщенной задачи о назначениях при одновременном поступлении работ: № 2024680927: заявл. 11.09.2024: опубл. 26.09.2024 / К. В. Матрохина, А. В. Гурко, Е. С. Черевко; заявитель Федеральное государственное бюджетное образовательное учреждение высшего образования «Санкт-Петербургский горный университет императрицы Екатерины II». – EDN OE00AE.

4. Багшыева, О. Базы данных и системы для их управления / О. Багшыева, А. Хуммедова, А. Давутгулыев, М. Багшыев // Символ науки. 2024. №3-1-1. URL: <https://cyberleninka.ru/article/n/bazy-dannyh-i-sistemy-dlya-ih-upravleniya> (дата обращения: 25.02.2025).

5. Свидетельство о государственной регистрации программы для ЭВМ № 2024682814. Программа для распределения файлов по уровням памяти вычислительной системы методом ветвей и границ / К. В. Матрохина, А. В. Гурко, В. А. Илюхин. – URL: <https://digital.spmi.ru/patent/frontend/web/filesize/2024682814.pdf> (дата обращения: 12.02.2025 г.)

6. Куницына, А. Г. Прогнозирование с помощью регрессионного анализа / А. Г. Куницына, Л. А. Винковская // Достижения науки и образования. 2020. №9 (63). URL: <https://cyberleninka.ru/article/n/prognozirovanie-s-pomoschyu-regressionnogo-analiza> (дата обращения: 25.02.2025 г.).

7. Князев, И. В. Анализ работы приложения с использованием server-side rendering: миграция, настройка и развертывание приложения Next.js // Sciences of Europe. 2021. №76-1. URL: <https://cyberleninka.ru/article/n/analiz-raboty-prilozheniya-s-ispolzovaniem-server-side-rendering-migratsiya-nastroyka-i-razvertyvanie-prilozheniya-next-js> (дата обращения: 25.02.2025).

8. Тельо-Родригес, М. Путеводитель по проектированию удобных web-API / М. Тельо-Родригес, Х. Очаран-эрнандес, Х. К. Перес-арриага, К. Лимон К., А. Х. Санчес-гарсия // Труды ИСП РАН. 2021. №1. URL: <https://cyberleninka.ru/article/n/putevoditel-po-proektirovaniyu-udobnyh-web-api> (дата обращения: 25.02.2025).
9. regression // npm URL: <https://www.npmjs.com/package/regression> (дата обращения: 03.03.2025).
10. Газизуллин, Н.И. Разработка клиентской части веб-приложения с использованием технологий SPA / Н. И. Газизуллин, И. Е. Плещинская // StudNet. 2020. №8. URL: <https://cyberleninka.ru/article/n/razrabotka-klientskoy-chasti-veb-prilozheniya-s-ispolzovaniem-tehnologiy-spa> (дата обращения: 25.02.2025).
11. Разяпов, Р. Б. Обзор технологий внутреннего позиционирования / Р. Б. Разяпов, В. Е. Жигалов, А. В. Гурко // Анализ и прогнозирование систем управления в промышленности, на транспорте и в логистике : XXIII Международная научно-практическая конференция молодых учёных, студентов и аспирантов, Санкт-Петербург, 18–20 апреля 2023 года. – Санкт-Петербург: ООО "Медиапапир", 2023. – С. 144-150. – EDN TUXJMG
12. Deep Learning for Generic Object Detection: A Survey / Li. Liu, W. Ouyang, X. Wang [et al.] // International Journal of Computer Vision. – 2020. – Vol. 128, No. 2. – P. 261-318. – DOI 10.1007/s11263-019-01247-4. – EDN KDEMGX. - URL: <https://elibrary.ru/item.asp?id=52926901> (дата обращения: 18.02.2025)
13. Srivastava, G. User-interactive salient object detection using YOLOv2, lazy snapping, and gabor filters / G. Srivastava, R. Srivastava // Machine Vision and Applications. – 2020. – Vol. 31, No. 3. – P. 17. – DOI 10.1007/s00138-020-01065-6. – EDN TJKOFA. - URL: <https://scihub.ru/10.1007/s00138-020-01065-6> (дата обращения: 18.02.2025)
14. The Study of Mathematical Models and Algorithms for Face Recognition in Images Using Python in Proctoring System / Sh. U. Niyazbekova, A. A. Nurpeisova, A. Shaushenova [et al.] // Computation. – 2022. – Vol. 10, No. 8. – DOI 10.3390/computation10080136. - URL: <https://elibrary.ru/item.asp?id=50380060> (дата обращения: 18.02.2025)
15. The Study of Mathematical Models and Algorithms for Face Recognition in Images Using Python in Proctoring System / Sh. U. Niyazbekova, A. A. Nurpeisova, A. Shaushenova [et al.] // Computation. – 2022. – Vol. 10, No. 8. – DOI 10.3390/computation10080136. - URL: <https://elibrary.ru/item.asp?id=50380060> (дата обращения: 18.02.2025)