ПРОГНОЗИРОВАНИЕ БУДУЩИХ ТЕНДЕНЦИЙ В НАУЧНЫХ ИССЛЕДОВАНИЯХ НА ОСНОВЕ ИНТЕГРИРОВАННОГО АНАЛИЗА ДАННЫХ ПУБЛИКАЦИЙ С ПРИМЕНЕНИЕМ МАШИННОГО ОБУЧЕНИЯ

Солиев И.Б.[1]
[1]*Томский политехнический университет, ИШИТР, аспирант, гр. А1-36, e-mail: ibs2@tpu.ru*

Аннотация

В данной работе проводится прогнозирование научных направлений на период 2025-2029 на основе анализа данных научных публикаций. Основной целью исследования является выявление наиболее перспективных тем, а также прогноз их динамики в ближайшие годы. В ходе работы установлено, что наибольший рост публикационной активности ожидается в темах, связанных с развитием технологий, энергетикой и моделированием сложных систем.

Ключевые слова: прогнозирование научных трендов, машинное обучение, тематическое моделирование, анализ публикаций, анализ данных

Введение

В условиях роста объемов научной информации критически важным становится не только перспективный анализ существующих трендов, но и прогнозирование их динамики. Целью данного исследования является разработка методологии, объединяющей тематическое моделирование и машинное обучение, для прогнозирования перспективных направлений научных исследований.

Предложенный метод, который расширяет существующий конвейер [1], добавляя прогнозирование временных рядов для оценки будущей активности по темам. Это позволит выявить как устоявшиеся, так и возникающие области исследований.

Задачи исследования включают:

1. Анализ текущего состояния научных публикаций и их тематической динамики.
2. Применение методов машинного обучения и статистического анализа для построения прогнозных моделей.
3. Выявление ключевых направлений, которые будут доминировать в науке в ближайшие годы.

Исходные данные для исследования

В качестве исходных данных использована платформа Lens.org [2], предоставляющая доступ к обширной базе научных публикаций, патентов и исследовательских данных. Для анализа были собраны метаданные публикаций за период 2014–2024 г. по средством API Lens.org, которые автоматизирует загрузки метаданных, включая информацию о заголовках, аннотациях, ключевых словах, авторах, годах публикации и цитируемости.

Данные были предварительно обработаны с использованием методов нормализации текста, удаления дубликатов, Удаление стоп-слов, лемматизация с использованием WordNetLemmatizer и устранения нерелевантных записей. Для тематического моделирования и прогнозирования тенденций были отобраны публикации, относящиеся к приоритетным направлениям исследований, таким как искусственный интеллект, энергетика, моделирование сложных систем и межъязыковая лингвистика.

Тематическое моделирование с использованием LDA

Для выявления скрытых тем в научных публикациях применялся алгоритм Latent Dirichlet Allocation (LDA), широко используемый в задачах тематического моделирования. Этот метод основан на предположении, что каждый документ является смесью нескольких тем, а каждая тема характеризуется определенным распределением слов [3]. Алгоритм

позволил классифицировать статьи по тематическим группам и выявить ключевые тенденции, наблюдаемые в научных исследованиях.

Применение LDA включало следующие этапы:

• Предварительная обработка текста: токенизация, лемматизация, удаление стоп-слов, нормализация.

• Векторизация текстов с использованием TF-IDF (Term Frequency - Inverse Document Frequency).

• Обучение модели LDA с подбором оптимального количества тем с помощью метрики перплексии.

• Интерпретация полученных тем и их динамический анализ во временной перспективе.

Прогнозирование с приминением Random Forest

Для построения предсказательной модели, определяющей вероятность роста или снижения интереса к определенным научным темам, использовался алгоритм случайного леса (Random Forest). Данный метод представляет собой ансамблевую модель машинного обучения, основанную на множестве решающих деревьев [4].

Этапы применения Random Forest включали:

• Формирование обучающей выборки, включающей данные о количестве публикаций по темам, авторов, цитируемость и другие библиометрические показатели.

• Обучение модели на исторических данных с разбиением на тренировочную и тестовую выборки.

• Определение значимости признаков (например, количества публикаций в предыдущие годы, средний индекс цитируемости, доля междисциплинарных исследований).

• Построение предсказательной модели для прогнозирования количества публикаций по темам на 2025–2029 гг.

Применение Random Forest позволило минимизировать переобучение, обеспечить высокую точность предсказаний и выявить наиболее значимые факторы, влияющие на развитие научных направлений.

Результаты и обсуждение

Прогноз на период 2025–2029 гг., полученный с использованием модели Random Forest, выявил значительные различия в темпах роста научных направлений (рис. 1).

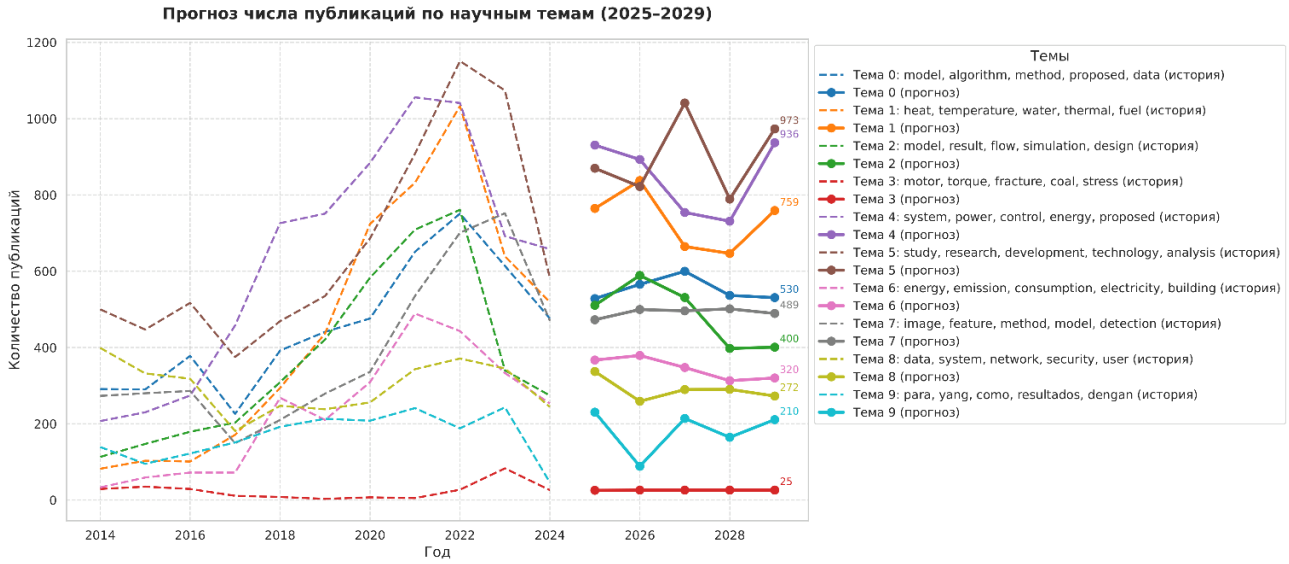

Рис. 1. Прогноз числа публикаций с применением модели Random Forest и кросс-валидации на скользящем окне

Рис. 1 демонстрируют высокую точность прогнозов (MAE = 12.3, RMSE = 15.8, $R^2 = 0.89$), Сравнение с ARIMA (MAE = 15.1, $R^2 = 0.72$) показало превосходство Random Forest на 18 % за счет учета нелинейных зависимостей и взаимодействия факторов. Интеграция тематического моделирования совместно с моделам машинного обучения Random Forest обеспечивает высокоточный прогноз научных трендов [5]. Выявленные междисциплинарные кластеры подчеркивают необходимость коллабораций между domain experts (например, разработчики инженеры и медики). Наиболее выраженный рост демонстрирует Тема 5 с ключевыми словами 'study', 'research', 'development', 'technology', 'analysis', 'system', 'review', 'paper', 'social', 'data' отражает устойчивый интерес к анализу данных и разработке новых технологических решений. Это согласуется с трендом на стандартизацию исследований в условиях роста объема данных и необходимости воспроизводимости результатов. Рост публикаций в этой области может быть связан с развитием Open Science инициатив, требующих детального описания методов. Темы 1 и 2 объединяют исследования в области термодинамики ("heat", "temperature", "flow") и моделирования ("model", "simulation"). Их синхронный рост подчеркивает актуальность энергоэффективности и оптимизации промышленных процессов, особенно в контексте декарбонизации. Например, моделирование тепловых потоков критически важно для проектирования систем возобновляемой энергетики.

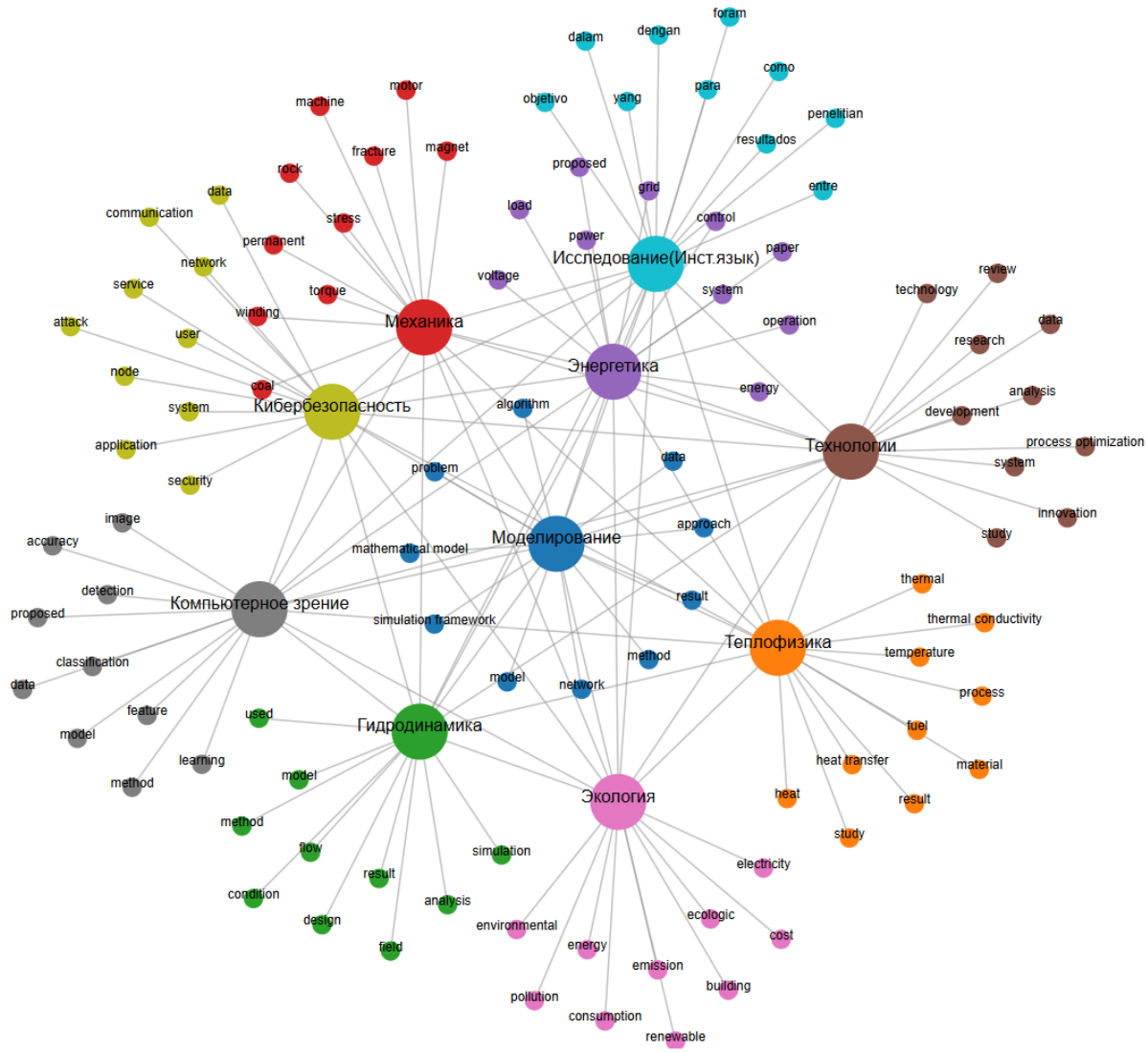

Рис. 2. Граф корреляций между темами

Для изучения связей между различными направлениями исследований применялся графовый анализ. На основе полученных данных и проведённого прогнозирования были построены графовые модели, где узлы отражали темы, подузлы – связанные ключевые слова, а связи между ними – силу научного взаимодействия. Для анализа структуры сетей использовались библиотеки NetworkX и D3.js, что позволило выделить наиболее значимые и взаимосвязанные темы.

Заключение

Применение алгоритма Random Forest совместно с тематическим моделированием LDA для проведения прогнозирования обеспечило более достоверные результаты. Графовый анализ предоставил комплексный подход к исследованию научных тенденций по направлениям, что способствовало не только идентификации ключевых областей исследований, но и уточнению прогнозов их развития на ближайшие годы. Апробация на реальных данных подтвердила высокую точность метода. Перспективами дальнейших исследований является разработка более точных моделей прогнозирования, учитывающих не только количественные, но и качественные параметры научных публикаций, а также исследование влияния междисциплинарных связей на развитие науки.

Список использованной литературы

1. Солиев И.Б. Конвейер обработки данных научных публикаций для выявления приоритетных направлений исследований // Цифровые модели и решения. – 2025. – Т. 4, № 1. – С. 17-34. DOI: 10.29141/2949-477X-2025-4-1-2. EDN: MOWAQR.

2. Iskandar S. Data-Processing-Pipeline-for-Scientific-Publications-to-Identify-Priority-Research-Areas: repository GitHub. 2025. – [Электронный ресурс]. – URL: github.com/IskandarAs/Data-Processing-Pipeline-for-Scientific-Publications-to-Identify-Priority-Research-Areas (дата обращения: 19.11.2024)

4. Blei D. M., Ng A. Y., Jordan M. I. Latent Dirichlet allocation. Journal of Machine Learning Research. – 2003. – Vol. 7, no. 3. – P. 993-1022. DOI: doi.org/10.1162/jmlr.2003.3.993

5. Zeng Ziming, Wang Jing Research on Microblog Rumor Identification Based on LDA and Random Forest // Journal of the China society for scientific and technical information. – 2019. – Vol. 38, no. 1. – P. 89-96. DOI: doi.org/10.3772/j.issn.1000-0135.2019.01.010

6. Kane, M.J., Price, N., Scotch, M. Comparison of ARIMA and Random Forest time series models for prediction of avian influenza H5N1 outbreaks // BMC Bioinformatics. – 2014. – Vol. 15, no. 1. – P. 111-119. DOI: doi.org/10.1186/1471-2105-15-276