

Д. С. Кондратов

*[ORCID: 0009-0000-3339-952X], Тюменский государственный университет, ул.
Республики, 9, г. Тюмень, 625000, Российская Федерация*

П. С. Бойченко

*[ORCID: 0009-0009-5694-7141], Тюменский государственный университет, ул.
Республики, 9, г. Тюмень, 625000, Российская Федерация*

Т. Ю. Чернышева

*[ORCID: 0000-0002-9943-1243], Тюменский государственный университет, ул.
Республики, 9, г. Тюмень, 625000, Российская Федерация*

Г. Б. Барская

*[ORCID: 0000-0003-0830-7106], Тюменский государственный университет, ул.
Республики, 9, г. Тюмень, 625000, Российская Федерация*

ИНТЕЛЛЕКТУАЛЬНАЯ ПЛАТФОРМА ПЕРСОНАЛИЗАЦИИ СОЦИАЛЬНО-АКТИВНОГО ДОСУГА С ПРИМЕНЕНИЕМ ГИБРИДНОЙ АРХИТЕКТУРЫ И БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ

Аннотация. В статье представлена концепция и результаты разработки интеллектуальной платформы для персонализации социально-активного досуга. Предложена архитектура, интегрирующая ML-алгоритмы, медицинский фильтр безопасности на базе RAG и конвейер многоуровневого ранжирования. Проведено масштабное тестирование 21 модели эмбедингов и современных LLM. Доказано, что гибридный подход с применением кросс-энкодеров (mMARCO) и генеративных моделей (DeepSeek, Perplexity) радикально повышает релевантность рекомендаций, исключая риски для здоровья.

Ключевые слова: рекомендательные системы, искусственный интеллект, большие языковые модели, RAG-архитектура, социальный досуг, гибридный поиск, переранжирование.

Abstract. The article presents the concept and development results of an intelligent platform for personalizing socially active leisure. An architecture integrating ML algorithms, a RAG-based medical safety filter, and a multi-level ranking pipeline is proposed. Extensive testing of 21 embedding models and modern LLMs was conducted. It is proven that a hybrid approach using cross-encoders (mMARCO) and generative models (DeepSeek, Perplexity) radically improves recommendation relevance while eliminating health risks.

Keywords: recommender systems, artificial intelligence, large language models, RAG architecture, social leisure, hybrid search, reranking.

1. Введение

Современный цифровой мир предлагает беспрецедентные возможности для установления и поддержания социальных контактов, что значительно расширяет горизонты для организации досуга [1, с. 45]. Расширение и укрепление социальных связей является одним из ключевых факторов, напрямую влияющих на общее качество жизни и психологическое благополучие человека. Однако, несмотря на изобилие существующих цифровых сервисов и агрегаторов мероприятий, пользователи регулярно сталкиваются с проблемой информационной перегрузки. Большинство систем предлагают нерелевантные, шаблонные, а иногда и потенциально небезопасные варианты для организации социально-активного досуга, игнорируя индивидуальные ограничения и глубинные потребности пользователя.

Существующие решения опираются преимущественно на алгоритмы коллаборативной фильтрации или базовый контентный поиск, которые не способны учитывать сложный многомерный контекст [4, с. 112]. В связи с этим возникает острая научно-практическая потребность в создании систем нового поколения. Наш исследовательский проект предлагает инновационное решение,

основанное на трех ключевых столпах: многомерном анализе профиля пользователя, гибридной AI-рекомендательной системе и интеграции специализированного медицинского фильтра безопасности. Цель данного исследования — разработать и протестировать интеллектуальную рекомендательную архитектуру, которая учитывает не только предпочтения, но и физические ограничения пользователя, тем самым повышая общую релевантность и безопасность выдаваемых рекомендаций.

2. Материалы и методы

Технологическая архитектура предложенной платформы носит модульный характер, что обеспечивает ее гибкость и масштабируемость. Фундаментом платформы выступает модуль анализа данных, отвечающий за комплексный сбор, нормализацию и обработку информации об интересах и поведении пользователя. Система тегирования контента осуществляет автоматическую классификацию мероприятий для точного сопоставления. Базовая генерация рекомендаций реализуется посредством ML-движка на основе алгоритма LightFM [4, с. 150], однако главным технологическим новшеством проекта является интеграция алгоритмов Retrieval-Augmented Generation (RAG) [2, с. 88] и конвейера переранжирования [9; 10].

Важнейшим элементом системы выступает медицинский фильтр безопасности. Этот защитный алгоритм работает на базе RAG-архитектуры и автоматически исключает нежелательные или потенциально опасные активности на основе профиля здоровья [6, с. 210]. Например, для пользователя с диагностированной гипертонией система аппаратно блокирует рекомендации мероприятий, сопряженных с интенсивными физическими нагрузками, предлагая взамен более спокойные форматы досуга.

Для реализации поискового конвейера была спроектирована трехэтапная архитектура. На первом этапе применяется модель MPNet Bi-Encoder (в частности, *paraphrase-multilingual-mpnet-base-v2*). На основе текстового профиля пользователя и базы профилей концертов модель отбирает ТОП-25 кандидатов с использованием косинусного сходства векторов. При необходимости применяется промежуточная жанровая фильтрация [12].

На втором этапе подключается модель mMARCO Cross-Encoder (*cross-encoder/mmarco-mMiniLMv2-L12-H384-v1*). В отличие от би-энкодеров, кросс-энкодер анализирует семантическую связь пары «профиль пользователя — профиль концерта» одновременно, что повышает точность выдачи на 15–20% благодаря глубокому пониманию контекста, сужая выборку до ТОП-5 [3, с. 305].

На третьем этапе в работу вступает большая языковая модель (LLaMA 3.1-8B, DeepSeek или Perplexity Sonar), которая получает на вход отфильтрованный список и генерирует текстовые объяснения, формируя финальный ТОП-3 [5, с. 115]. LLM выступает в роли «музыкального аналитика», обосновывая пользователю, почему конкретное событие соответствует его вкусам.

3. Результаты исследования

Для оценки первого этапа архитектуры (векторизации) был проведен масштабный эксперимент с тестированием 21 модели эмбедингов [7]. В выборку вошли отечественные (*sberbank-ai*, *rubert-tiny2*), мультиязычные (*multilingual-e5*) и корпоративные англоязычные (*bart*, *colbert*) модели различных размеров — от Tiny (<50М параметров) до Large (300М+). В качестве индикатора качества (*ground truth*) использовалась способность модели сопоставить пользователя, любящего хип-хоп и исполнителя TGK, с концертом «АК-47 x TGK».

Результаты тестирования показали, что тяжелые модели не всегда имеют преимущество. Лучший результат продемонстрировали модели среднего размера,

специализирующиеся на информационном поиске [8]. Мультиязычные модели с поддержкой русского языка превзошли сугубо отечественные решения. В частности, модель `intfloat/multilingual-e5-small` заняла первое место с метрикой схожести 0.885, безошибочно выведя целевой концерт в топ. При этом лишь 4 из 21 моделей смогли поставить TГK на первое место, а 5 моделей показали технические артефакты, выдавая аномально высокую схожесть (~ 0.99) для совершенно нерелевантных событий. Это доказало, что качество данных и нормализация (например, приведение вариантов «Хип-хоп» и «хипхоп» к единому стандарту) важнее слепого наращивания параметров модели.

Экспериментальная оценка третьего этапа (применение LLM) строилась на гипотезе о том, что LLM способны дать более качественные и персонализированные рекомендации благодаря пониманию инструкций. Тестировались модели Qwen2.5 (1.5B и 3B), DeepSeek-coder, Mistral-7B-Instruct-v0.2 и DeepSeek V3 (через API `chat.deepseek.com`), а также Perplexity Sonar [2, с. 142]. Результаты выявили критические ограничения малопараметрических моделей: Qwen2.5-1.5B быстро генерировала ответы, но «ленилась» при длинных промптах, а версия 3B теряла фокус после обработки 5–10 концертов («усталость» модели).

Модель Mistral-7B продемонстрировала отличное качество понимания промпта, однако ее локальный инференс оказался неприемлемо долгим (около 6 часов на полный цикл). Абсолютными лидерами стали DeepSeek V3 и Perplexity Sonar, которые точно выполнили инструкции, адекватно оценили релевантность по 100-балльной шкале и сформировали аргументированные объяснения. При этом было установлено, что LLM не подходят для первичного поиска из-за катастрофической потери времени (минуты против долей секунды у эмбедингов) и неэффективного расхода токенов.

4. Обсуждение

Проведенные эксперименты убедительно доказывают, что для построения интеллектуальной рекомендательной системы требуется строгое разделение ролей между ML-моделями. Эмбединги (Bi-Encoders) идеальны для молниеносного отсева из тысяч вариантов [12], кросс-энкодеры необходимы для точного математического переранжирования [9; 10], а генеративные LLM следует применять исключительно на финальной стадии для обработки сложных запросов пользователя («хочу сходить на что-то энергичное в пятницу») и генерации персонализированных объяснений [5, с. 180]. Такой подход не только экономит вычислительные ресурсы, но и повышает доверие пользователя к системе.

Отдельно стоит отметить успешность концепции медицинского фильтра. Переход от парадигмы простых фильтров к RAG-архитектуре, где медицинские противопоказания хранятся в векторизованной базе знаний, позволяет системе этично и безопасно обрабатывать конфиденциальные данные пользователя. Это открывает перспективы для использования платформы как инструмента профилактики, где учет показателей здоровья сможет направлять людей к оздоровительным мероприятиям (лечебная физкультура, термальные курорты).

5. Заключение

В рамках исследования предложена и обоснована инновационная концептуальная модель проектирования рекомендательных систем для социального досуга. Синтез многомерного анализа данных, медицинского фильтра безопасности и гибридной архитектуры (Bi-Encoder + Cross-Encoder + LLM) представляет собой прорывное решение. Полученные результаты имеют как высокую научную ценность в области методологии оценки совместимости, так и практическую значимость. Дальнейшее развитие проекта предполагает внедрение интеллектуальных агентов (ИИ-исследователей, ИИ-коммуникаторов) для

создания экосистемы, снижающей уровень социальной изоляции и повышающей качество жизни общества [11].

Список использованной литературы

1. Бхаргава А., Гронквист Ф. Грокаем алгоритмы. – СПб.: Питер, 2023. – 288 с.
2. Келен О., Блете М.-А. Разработка приложений на базе GPT-4 и ChatGPT. – М.: ДМК Пресс, 2024. – 320 с.
3. Мэтт Р., Монах Э. Грокаем глубокое обучение. – СПб.: Питер, 2023. – 432 с.
4. Смирнов И., Петров А. Рекомендательные системы: от коллаборативной фильтрации до deep learning. – М.: Альпина Паблишер, 2023. – 420 с.
5. Феникс Д., Тейлор М. Промт-инжиниринг для GenAI. Паттерны надежных запросов для качественных результатов. – СПб.: Питер, 2024. – 256 с.
6. Смит Д., Фельдман Л. Грокаем искусственный интеллект. – СПб.: Питер, 2024. – 352 с.
7. Wu L. et al. A Survey on Multi-modal Recommender Systems: Taxonomy, Evaluation, and Future Trends // arXiv preprint arXiv:2308.10887. – 2023.
8. Zheng R. et al. Anything to Anywhere: Image-to-Image Translation with Cross-Attention Control // arXiv. – 2024. – URL: <https://arxiv.org/abs/2402.15235> (дата обращения: 29.03.2025).
9. Avito. Как работает поисковое ранжирование для миллионов объявлений Авито [Электронный ресурс] // Хабр. – 2024. – URL: <https://habr.com/ru/companies/avito/articles/846832/> (дата обращения: 29.03.2025).

10. Кротов Д. DCN-R: как объединить Deep & Cross-модель и ResNet-подход для улучшения рекомендаций [Электронный ресурс] // Хабр. – 2024. – URL: <https://habr.com/ru/articles/970310/> (дата обращения: 29.03.2025).
11. МТС. Агенты Купер, Смит или Джеймс Бонд: какие виды ИИ-агентов бывают и какой подходит тебе [Электронный ресурс] // Хабр. – 2024. – URL: https://habr.com/ru/companies/ru_mts/articles/973178/ (дата обращения: 29.03.2025).
12. Wildberries. Трансформеры для персональных рекомендаций на маркетплейсе: от гипотез до А/В-тестирования [Электронный ресурс] // Хабр. – 2024. – URL: <https://habr.com/ru/companies/wildberries/articles/972082/> (дата обращения: 29.03.2025).