

МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»

Кафедра математической
кибернетики и компьютерных наук

**РЕКОМЕНДАТЕЛЬНАЯ СИСТЕМА ФИЛЬМОВ НА ОСНОВЕ
ОТКРЫТЫХ ДАННЫХ**

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 5 курса 551 группы
направления 09.03.04 — Программная инженерия
факультета компьютерных наук и информационных технологий
Дукаева Аслана Вахидовича

Научный руководитель

к. т. н., доцент

В. М. Соловьёв

Заведующий кафедрой

к. ф.-м. н., доцент

С. В. Миронов

Саратов 2025

СОДЕРЖАНИЕ

ВВЕДЕНИЕ	3
1 Теоретические основы рекомендательных систем.....	4
1.1 Классификация рекомендательных систем	4
1.2 Коллаборативная фильтрация: принципы и ограничения	5
1.3 Контент-ориентированные методы	5
1.4 Роль открытых данных	6
2 Реализация гибридной рекомендательной системы	7
2.1 Анализ требований.....	7
2.2 Архитектура гибридной рекомендательной системы	8
2.3 Реализация прототипа	8
2.4 Алгоритм гибридизации	8
2.5 Тестирование системы	9
ЗАКЛЮЧЕНИЕ	10

ВВЕДЕНИЕ

Современная индустрия кинематографа переживает бурное развитие, обусловленное внедрением новых технологий и ростом популярности стриминговых платформ, таких как Netflix, Amazon Prime и Кинопоиск. В условиях увеличения объемов пользовательских данных и повышения ожиданий аудитории к персонализированному контенту возникает потребность в создании интеллектуальных рекомендательных систем (РС), способных эффективно обрабатывать информацию о фильмах и предпочтениях пользователей. Такие системы должны не только предлагать релевантные рекомендации, но и решать более сложные задачи, такие как классификация контента, анализ трендов и персонализация пользовательского опыта. Актуальность данной работы определяется необходимостью преодоления ограничений традиционных подходов к созданию РС, таких как проблема “холодного старта” и разреженность данных, а также потребностью в разработке универсальных решений, применимых к различным сценариям использования в киноиндустрии и за ее пределами.

Рекомендательные системы играют ключевую роль в повышении удовлетворенности пользователей и увеличении доходов компаний, что подтверждается исследованиями, показывающими рост доходов на 20–30 при внедрении РС. Однако существующие решения имеют ряд ограничений. Контентно-ориентированные методы (CBF), основанные на анализе метаданных фильмов (жанр, актеры, режиссеры), часто игнорируют контекстные данные пользователей, что снижает персонализацию рекомендаций. Коллаборативная фильтрация (CF), опирающаяся на поведенческие данные, сталкивается с проблемой “холодного старта” для новых пользователей или объектов, а также с трудностями обработки разреженных матриц взаимодействий.

Структура и объем работы. Бакалаврская работа состоит из раздела с используемыми определениями и сокращениями, введения, 4 разделов, заключения, списка источников и 5 приложений. Общий объем работы — 51 страницы, из них 46 страниц — основное содержание и список использованных источников 25 наименования.

1 Теоретические основы рекомендательных систем

Рекомендательные системы (РС) представляют собой программные комплексы, обеспечивающие персонализированные рекомендации на основе анализа данных о предпочтениях пользователей, их поведении или характеристиках объектов. В условиях роста объемов цифровой информации РС становятся неотъемлемой частью стриминговых платформ, электронной коммерции и социальных сетей. Согласно исследованиям, внедрение РС увеличивает удовлетворенность пользователей и доходы компаний на 20–30%. В данном разделе рассматриваются основные подходы к созданию РС, их преимущества и ограничения, а также роль открытых данных в их построении.

1.1 Классификация рекомендательных систем

Классификация РС основывается на подходах к генерации рекомендаций. Выделяют следующие категории:

1. *Коллаборативная фильтрация (CF)*: основана на анализе взаимодействий пользователей с объектами, предполагая, что пользователи с аналогичными предпочтениями имеют схожие интересы. Примером является алгоритм рекомендаций Netflix, использующий историю просмотров.
2. *Контент-ориентированные методы (CBF)*: используют атрибуты объектов, такие как жанр или ключевые слова, для формирования рекомендаций на основе истории взаимодействий пользователя. Пример — система IMDb, опирающаяся на метаданные фильмов.
3. *Гибридные подходы*: комбинируют CF и CBF для повышения точности и устранения ограничений каждого метода. Платформа Кинопоиск применяет гибридный подход, объединяя анализ просмотров и жанров.
4. *Контекстно-зависимые системы*: учитывают дополнительные параметры, такие как время или местоположение.
5. *Системы, основанные на знаниях*: используют правила или онтологии для рекомендаций в специализированных доменах.

Каждый подход имеет свои преимущества и ограничения, что подчеркивает актуальность разработки универсальных гибридных систем.

1.2 Коллаборативная фильтрация: принципы и ограничения

Коллаборативная фильтрация (CF) — один из наиболее распространенных методов РС. Она предполагает, что пользователи с похожими предпочтениями в прошлом будут иметь схожие интересы в будущем. Алгоритмы CF используют матрицу взаимодействий «пользователь–объект», которая часто является разреженной. Различают два подхода:

- *User-Based CF*: рекомендации формируются на основе сходства пользователей. Сходство между пользователями u и v вычисляется, например, с помощью косинусной меры:

$$\text{sim}(u, v) = \frac{\sum_{i \in I} r_{u,i} \cdot r_{v,i}}{\sqrt{\sum_{i \in I} r_{u,i}^2} \cdot \sqrt{\sum_{i \in I} r_{v,i}^2}},$$

где $r_{u,i}$ — рейтинг, выставленный пользователем u объекту i , а I — множество объектов, оцененных обоими пользователями.

- *Item-Based CF*: акцент на схожести объектов, основанной на оценках пользователей, что более масштабируемо из-за меньшего числа объектов.

Ограничения CF включают:

- *Проблема холодного старта*: отсутствие данных о новых пользователях или объектах.
- *Разреженность данных*: малая заполненность матрицы взаимодействий.
- *Скалируемость*: высокая вычислительная сложность при больших объемах данных.

Для преодоления этих проблем применяются методы матричной факторизации, такие как SVD или ALS.

1.3 Контент-ориентированные методы

Контент-ориентированные методы (CBF) формируют рекомендации на основе характеристик объектов (жанр, режиссер, актеры) и истории взаимодействий пользователя. Профиль пользователя P_u сопоставляется с профилем объекта P_i с использованием, например, косинусного сходства:

$$\text{score}(u, i) = \cos(P_u, P_i) = \frac{P_u \cdot P_i}{\|P_u\| \cdot \|P_i\|}.$$

Преимущества CBF:

- Возможность рекомендаций для новых объектов при наличии их описания.
- Независимость от данных других пользователей, что решает проблему холодного старта для объектов.
- Прозрачность рекомендаций.

Ограничения:

- Ограниченная персонализация из-за фокуса на ранее выбранных категориях.
- Зависимость от качества метаданных.
- Недостаток разнообразия рекомендаций.

Для улучшения CBF применяются методы обработки естественного языка (NLP), такие как TF-IDF или word embeddings.

1.4 Роль открытых данных

Открытые данные, такие как MovieLens, IMDb и TMDb, играют ключевую роль в построении РС, предоставляя метаданные о фильмах и пользовательские оценки. Преимущества их использования:

- Обогащение профилей объектов (жанры, актеры, онтологии из Wikidata).
- Решение проблемы холодного старта для новых объектов.
- Прозрачность и интерпретируемость рекомендаций.
- Доступность и масштабируемость.

Сложности включают:

- Неполнота и ошибки в данных (например, отсутствие информации о малоизвестных фильмах в IMDb).
- Гетерогенность форматов (JSON, CSV, RDF).
- Необходимость значительных ресурсов для обработки больших наборов данных, таких как MovieLens.

Для преодоления этих вызовов применяются методы ETL (извлечение, преобразование, загрузка), очистка данных и алгоритмы NLP. Открытые данные обеспечивают гибкость и демократизацию разработки РС, что делает их основой для универсальных систем.

2 Реализация гибридной рекомендательной системы

Разработка универсальной гибридной рекомендательной системы фильмов на основе открытых данных требует четкого определения требований, проектирования модульной архитектуры, реализации прототипа и его тестирования. В данном разделе описаны этапы создания системы, включая анализ требований, архитектуру, реализацию, алгоритм гибридизации и результаты тестирования, демонстрирующие высокую эффективность системы.

2.1 Анализ требований

Разработка системы началась с анализа требований, обеспечивающих научную обоснованность и воспроизводимость результатов. Функциональные требования включают:

- *Сбор и преобразование данных*: автоматизированный импорт из открытых источников (MovieLens, TMDb) с очисткой и нормализацией данных.
- *Формирование датасета*: создание репрезентативного набора данных (не менее 10 тысяч фильмов и 100 тысяч оценок) с разнообразием жанров и периодов.
- *Гибридный подход*: реализация алгоритмов CF (SVD, ALS) и CBF (TF-IDF, word embeddings) с настраиваемыми весами.
- *Оценка качества*: использование метрик Precision@K, Recall@K, RMSE, NDCG с автоматизированным тестированием.
- *Прототип*: модульная реализация на Python с интерфейсом для экспериментов.

Нефункциональные требования:

- *Воспроизводимость*: документирование процессов, открытость кода (GitHub), использование стандартного оборудования (4 ядра, 16 ГБ ОЗУ).
- *Гибкость*: модульная архитектура для интеграции новых алгоритмов.
- *Эффективность*: предобработка данных за 1 час, масштабируемость до 10 миллионов записей.

Ограничения включают качество данных и ограниченные вычислительные ресурсы, что учтено при проектировании.

2.2 Архитектура гибридной рекомендательной системы

Архитектура системы спроектирована как модульная, чтобы обеспечить гибкость и воспроизводимость. Она включает следующие компоненты (см. рисунок ??):

- *Импорт данных*: загрузка из MovieLens и TMDb через CSV и API.
- *Предобработка*: очистка данных, удаление пропусков, нормализация рейтингов.
- *Построение признаков*: создание матрицы взаимодействий и контентных признаков (one-hot кодирование жанров, TF-IDF для описаний).
- *Обучение*: использование LightFM с функцией потерь WARP для создания эмбедингов пользователей и фильмов.
- *Генерация рекомендаций*: формирование персонализированных списков фильмов.
- *Оценка качества*: вычисление метрик Precision@K, Recall@K, RMSE.

2.3 Реализация прототипа

Прототип системы реализован на Python с использованием библиотек LightFM, pandas и SQLite. Технологический стек выбран за его богатую экосистему и поддержку воспроизводимости. Основные этапы реализации:

- *Импорт данных*: загрузка данных из MovieLens (20M оценок) и TMDb через API.
- *ETL-процесс*: очистка данных (удаление дубликатов, заполнение пропусков средним рейтингом), нормализация (приведение к единому формату), сохранение в SQLite.
- *Обучение модели*: создание матрицы взаимодействий и обучение LightFM (30 эпох, WARP loss).
- *Интерфейс*: консольный интерфейс для тестирования рекомендаций.

2.4 Алгоритм гибридизации

Гибридизация реализована через позднее объединение (late fusion), где оценки коллаборативной ($R_{cf}(u, i)$) и контент-ориентированной ($R_{cb}(u, i)$) моделей комбинируются по формуле:

$$R_{hyb}(u, i) = \alpha \cdot R_{cf}(u, i) + (1 - \alpha) \cdot R_{cb}(u, i),$$

где $\alpha \in [0, 1]$ — эмпирически подбираемый коэффициент, обеспечивающий баланс между CF и CBF. Этот подход позволяет устранить проблему «холодного старта» и повысить персонализацию рекомендаций.

2.5 Тестирование системы

Система протестирована на наборах данных MovieLens и IMDb. Используются метрики:

- *Precision@10*: доля релевантных фильмов среди первых 10 рекомендаций.
- *Recall@10*: доля найденных релевантных фильмов.
- *F1-мера*: гармоническое среднее Precision@10 и Recall@10.
- *RMSE*: ошибка предсказания рейтингов.

Результаты:

- Precision@10 = 0.85 (против 0.78 для CBF и 0.80 для CF).
- F1-мера = 0.89 для классификации жанров.
- RMSE = 0.92 для предсказания рейтингов.

Сравнение с базовыми методами (User-Based и Item-Based CF) показало превосходство гибридного подхода: Accuracy = 0.84 при 500 пользователях против 0.69 и 0.68 соответственно (см. рисунок 2 диплома). Тестирование проводилось с 5-кратной кросс-валидацией, что подтверждает воспроизводимость результатов.

ЗАКЛЮЧЕНИЕ

В результате исследования разработана универсальная гибридная рекомендательная система фильмов на основе открытых данных, объединяющая контентно-ориентированные и коллаборативные подходы. Цель работы достигнута: создан прототип на Python с использованием LightFM, протестированный на наборах данных MovieLens и IMDb. Система показала высокую эффективность: $\text{Precision@10} = 0.85$, F1-мера = 0.89 для классификации жанров, $\text{RMSE} = 0.92$ для предсказания рейтингов [?]. Научная новизна заключается в универсальной архитектуре, решающей проблему «холодного старта», и оценке системы на нестандартных задачах. Практическая значимость определяется возможностью применения в стриминговых платформах и аналитических сервисах. Перспективы включают интеграцию мультимодальных данных и оптимизацию для реального времени.