Федеральное государственное бюджетное образовательное

учреждение высшего образования

«Национальный исследовательский университет «МЭИ»

|  |  |  |  |
| --- | --- | --- | --- |
| Институт: | ИВТИ | Кафедра: | УИТ |
| Направление подготовки: | | 27.04.04 Управление в технических системах | |

**ОТЧЕТ по практике**

|  |  |
| --- | --- |
| **Наименование практики:** | Учебная практика: ознакомительная практика |

**Тема:**

**СТУДЕНТ**

|  |  |
| --- | --- |
|  | / Козлов И.А. / |
| *(подпись )* | (*Фамилия и инициалы*) |

|  |  |
| --- | --- |
| Группа | А-01м-23 |
|  | *(номер учебной группы)* |

**ПРОМЕЖУТОЧНАЯ АТТЕСТАЦИЯ ПО ПРАКТИКЕ**

|  |
| --- |
|  |
| *(отлично, хорошо, удовлетворительно, неудовлетворительно, зачтено, не зачтено)* |

|  |  |
| --- | --- |
|  | / / |
| *(подпись )* | (*Фамилия и инициалы члена комиссии*) |

|  |  |
| --- | --- |
|  | / / |
| *(подпись )* | (*Фамилия и инициалы члена комиссии*) |

**Москва**

**2023**

**Оглавление**

[**АКТУАЛЬНОСТЬ РАЗРАБОТКИ СИСТЕМЫ РЕКОМЕНДАЦИЙ ДЛЯ КНИГ** 3](#_Toc164932934)

[ПОСТАНОВКА ЗАДАЧИ И ОСНОВНЫЕ ПРОБЛЕМЫ РЕКОМЕНДАТЕЛЬНЫХ СИСТЕМ 5](#_Toc164932935)

[ОСНОВНЫЕ АЛГОРИТМЫ ПРИМЕНЯЕМЫЕ В РЕКОМЕНДАТЕЛЬНЫХ СИСТЕМАХ 6](#_Toc164932936)

[**ПОСТРОЕНИЕ РЕКОМЕНДАТЕЛЬНОЙ СИСТЕМЫ НА ОСНОВЕ МЕТОДА КОНТЕНТНОЙ ФИЛЬТРАЦИИ** 10](#_Toc164932937)

# **АКТУАЛЬНОСТЬ РАЗРАБОТКИ СИСТЕМЫ РЕКОМЕНДАЦИЙ ДЛЯ КНИГ**

Рекомендательные системы – это технологии, которые предлагают пользователям персонализированные рекомендации, основанные на их предпочтениях и истории взаимодействия с платформой или сервисом. Эти системы широко применяются в различных областях, таких как электронная коммерция, стриминг мультимедиа, социальные сети, новостные порталы и другие.

Почему это актуально и какие потребности покрывает.

* Избыток информации

Пользователи часто сталкиваются с проблемой избытка выбора, когда сложно определить, что выбрать из множества вариантов. Рекомендательные системы помогают справиться с этой проблемой, предлагая персонализированные рекомендации на основе предпочтений и интересов пользователя.

* Увеличение продаж

Для бизнеса, занимающегося продажей книг, рекомендательные системы являются мощным инструментом для увеличения объема продаж. Персонализированные рекомендации способствуют увеличению конверсии и среднего чека покупателей.

* Привлечение внимания к новым продуктам

Рекомендации могут помочь в продвижении новых книг или авторов. Пользователи, заинтересованные в определенных жанрах или тематиках, могут быть направлены на новые релизы, что способствует расширению аудитории и увеличению продаж.

* Анализ данных и понимание потребительского поведения

Рекомендательные системы собирают данные о предпочтениях и действиях пользователей, что позволяет компаниям лучше понимать их потребности и предпочтения. Эта информация может быть использована для улучшения ассортимента книг, адаптации маркетинговых стратегий и оптимизации бизнес-процессов.

* Создание лояльности

Пользователи, получающие рекомендации, соответствующие их интересам, склонны возвращаться к платформе снова и снова, что способствует удержанию клиентов и повышению их уровня удовлетворенности.

Таким образом, рекомендательные системы не просто помогают пользователю выбрать товар, но и решают ряд других важных для бизнеса задач.

Для продвижения своих продуктов рекомендательные системы используют такие сервисы как

* Amazon - онлайн-ритейлер книг и других товаров, который
* Goodreads - это крупнейшая в мире социальная сеть для любителей книг.
* Bookmate - это цифровая платформа для чтения книг.
* Audible - это платформа для аудиокниг, которая предлагает аудиокниги и подкасты.
* Google Play Books - предлагает каталог книг и аудиокниг для чтения и прослушивания

ПОСТАНОВКА ЗАДАЧИ И ОСНОВНЫЕ ПРОБЛЕМЫ РЕКОМЕНДАТЕЛЬНЫХ СИСТЕМ

Формальная постановка задачи

Дано:

- множество пользователей,

- множество продуктов,

- матрица рейтингов размера n × m, где на месте будет стоять некоторое число, если пользователь оценил продукт и пусто в противном случае.

N - требуемое число рекомендаций, которые хотим получить от системы.

Требуется найти:

Для данного пользователя найти N-мерный вектор , где продукты , k ∈ N еще не оценены этим пользователем, то есть в матрице рейтингов стоит пусто на месте , а также, чтобы эти продукты наиболее точно удовлетворяли предпочтениям пользователя, то есть прогнозные рейтинги , были наибольшими.

Оценка качества будет описана отдельно применительно к коллаборативной фильтрации.

Проблематика рекомендательных систем включает в себя ряд вызовов и проблем:

* Фильтрование информации: Рекомендательные системы помогают пользователям справляться с информационной перегрузкой, но они также могут уменьшить разнообразие информации, предоставляя рекомендации, которые соответствуют их существующим интересам. Это может привести к "пузырьковой изоляции", когда пользователи ограничиваются контентом, который соответствует их мнениям и интересам, и не видят разнообразных точек зрения.
* Приватность и безопасность данных: Рекомендательные системы собирают и анализируют большое количество личных данных пользователей. Это может вызвать опасения в отношении приватности и безопасности данных, особенно если эти данные попадают в руки злоумышленников или используются в недобросовестных целях.
* Байесовский выбор: Рекомендательные системы могут уделять слишком много внимания более популярным или монотонным элементам, так как они часто получают больше взаимодействий от пользователей. Это может привести к неразнообразию рекомендаций и игнорированию менее известных, но более интересных вещей.
* Холодный старт: Рекомендательные системы могут столкнуться с проблемой "холодного старта", когда новые пользователи или элементы не имеют истории взаимодействия. В этом случае системе трудно предоставить релевантные рекомендации.
* Фильтрация нежелательного контента: Рекомендательные системы должны также учитывать этические и законодательные нормы, чтобы избегать рекомендаций нежелательного или вредного контента, такого как фейковые новости, порнография, насилие и др.

ОСНОВНЫЕ АЛГОРИТМЫ ПРИМЕНЯЕМЫЕ В РЕКОМЕНДАТЕЛЬНЫХ СИСТЕМАХ

**Коллаборативная фильтрация:**

На сегодняшний день, этот метод можно считать самым распространенным.

Метод коллаборативной фильтрации основан на схожести пользователей. Алгоритмы ищут пользователей со схожими интересами, затем рекомендуют пользователям товары или услуги, которыми воспользовались другие пользователи платформы. Часто в качестве меры схожести используют косинусное расстояние.

**Фильтрация на основе контента:**

Этот метод рекомендует элементы на основе их сходства с предпочтениями пользователя. Например, для рекомендации книг, этот метод учитывает жанры, авторов, аннотации и другие характеристики книг, которые пользователь предпочел ранее. Как правило строится так называемый «профиль пользователя», он отражает его интересы и старается найти книгу, которую пользователь захочет приобрести.

**Гибридные методы:**

Это комбинация различных методов, таких как фильтрация на основе контента и коллаборативная фильтрация, чтобы улучшить качество рекомендаций. Как правило, метод довольно сложен в реализации, однако способен выдавать неплохие результаты, так как учитывает как информацию о конкретном пользователе, так и действия схожих пользователей.

Таблица 1. Общее сравнение методов рекомендательных систем

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Метод | Описание | Вычислительная сложность | Хранение данных | Масштабируемость | Интерпретируемость |
| Контеная фильтрация | Анализирует характеристики контента, чтобы рекомендовать похожие элементы. | Низкая | Низкая | Высокая | Высокая |
| Коллаборативная фильтрация: | Находит пользователей с похожими предпочтениями и рекомендует им то, что понравилось другим. | Средняя | Средняя | Средняя | Средняя |
| Гибридные методы: | Сочетают несколько методов для достижения более точных рекомендаций. | Варьируется | Варьируется | Варьируется | Варьируется |

**Онлайн и офлайн рекомендации.**

Глобально рекомендательные системы можно разделить на онлайн и офлайн, из названий уже примерно ясно как они работают. Онлайн работают мгновенно, офлайн же используют более сложные алгоритмы, но могут выдавать более качественный результат. Дело в том, что алгоритмы бывают вычислительно сложными и не всегда могут выдавать результат мгновенно, поэтому часто на практике используют несколько систем.

# **МЕТОД КОНТЕНТНОЙ ФИЛЬТРАЦИИ В РЕКОМЕНДАТЕЛЬНЫХ СИСТЕМАХ**

Метод контентной фильтрации в рекомендательных системах основан на анализе содержания объектов (фильмов, книг, музыки, товаров и т.д.), которые пользователь уже оценил или с которыми он взаимодействовал. Система пытается найти новые объекты, похожие на те, которые уже понравились пользователю, и рекомендует их ему.

Существует два основных типа контентной фильтрации:

* Фильтрация на основе характеристик: Этот метод использует метаданные объектов, такие как жанр, автор, тема, описание и т.д., для определения их сходства.
* Фильтрация на основе текста: Этот метод использует анализ текста (например, отзывы пользователей, описания объектов) для определения их сходства.

Преимущества метода контентной фильтрации

* Простота реализации: Этот метод относительно прост в реализации по сравнению с другими методами, такими как коллаборативная фильтрация.
* Прозрачность: Этот метод легко понять и объяснить пользователям.
* Эффективность: Этот метод может быть очень эффективным для пользователей, которые уже имеют достаточно обширную историю взаимодействия с системой.

Недостатки метода контентной фильтрации

* Проблема "холодного старта": Этот метод не работает для пользователей, которые еще не имеют истории взаимодействия с системой.
* Ограниченная точность: Этот метод может быть неточным для пользователей, которые имеют широкий спектр интересов.
* Проблема "фильтрационного пузыря": Этот метод может привести к тому, что пользователи будут видеть только те объекты, которые похожи на те, которые они уже видели, что может ограничить их кругозор.
* В целом, метод контентной фильтрации является эффективным инструментом для создания рекомендательных систем. Однако его следует использовать в сочетании с другими методами, такими как коллаборативная фильтрация, чтобы повысить точность и эффективность рекомендаций.

Пример использования метода контентной фильтрации:

Сервис Netflix использует метод контентной фильтрации, чтобы рекомендовать фильмы своим пользователям. Netflix анализирует историю просмотров пользователя, чтобы определить, какие жанры и темы ему нравятся. Затем Netflix использует эту информацию, чтобы рекомендовать пользователю другие фильмы, которые могут ему понравиться.

# **ПОСТРОЕНИЕ РЕКОМЕНДАТЕЛЬНОЙ СИСТЕМЫ НА ОСНОВЕ МЕТОДА КОНТЕНТНОЙ ФИЛЬТРАЦИИ**

**Описание выборки**

Выборка на основе информации с сайта Goodreads содержит в себе информацию о пользователях и книгах.

Имеется 2 таблицы данных.

Таблица user:

'ID' – id текущего пользователя,

'Name' – название прочитанной им книги,

'Rating' – рейтинг, который пользователь поставил книге, оценки варьируются от 1 до 5.

Таблица book:

'Id' – id книги,

'Name' – название книги,

'Authors' – автор книги,

'ISBN' – международный стандартный книжный номер,

'Rating' – средний рейтинг книги проставленный пользователями,

'PublishYear' – год публикации,

'PublishMonth' – месяц публикации,

'PublishDay' – день публикации,

'Publisher' – издатель,

'RatingDist5' – количество оценок 5,

'RatingDist4' – количество оценок 4,

'RatingDist3' – количество оценок 3,

'RatingDist2' – количество оценок 2,

'RatingDist1' – количество оценок 1,

'RatingDistTotal' – общее число оценок пользователей,

'CountsOfReview' – общее число написанных ревью,

'Language' – язык книги,

'pagesNumber' – число страниц в книге,

'Description' – текстовое описание книги (аннотация).

Основным признаком для контентной фильтрации будет служить описание книг, прочитанных пользователем.

Число пользователей 3822

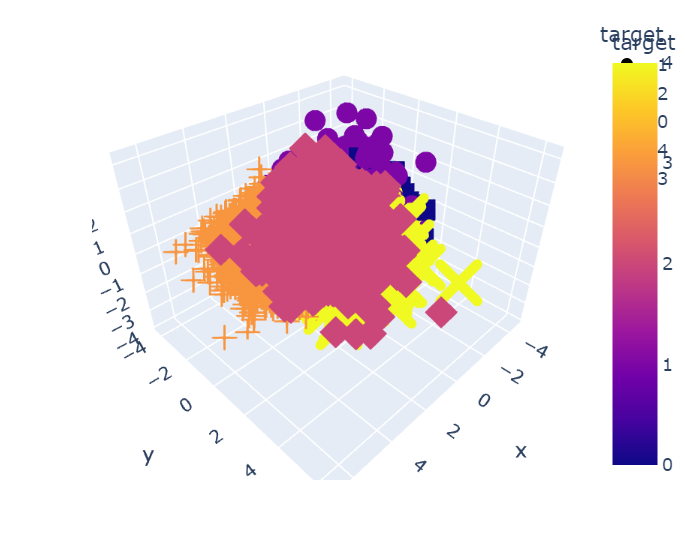
Число прочитанных книг 30470

Пример описания

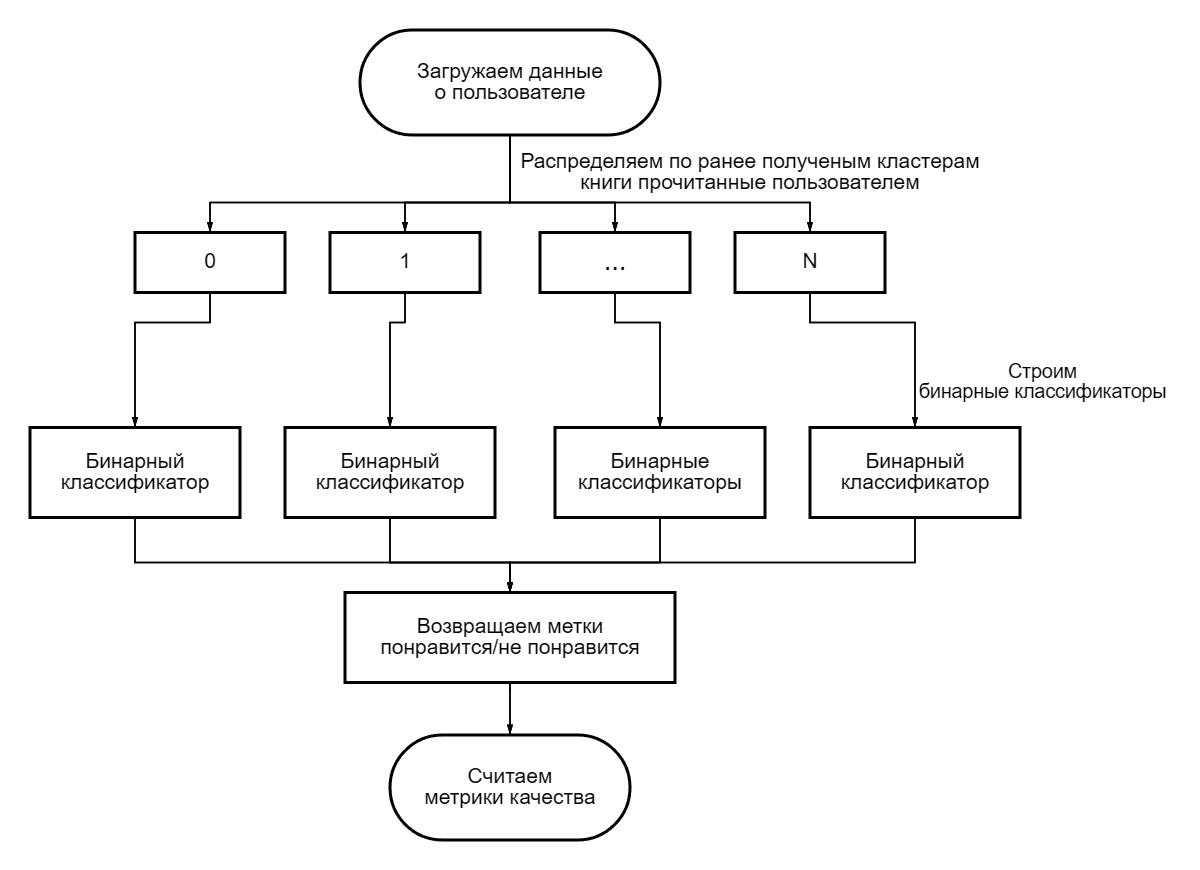
«Just when you thought it was safe to go back to the bookstore, Douglas Adams presents The Restaurant at the End of the Universe.<br />The Time: Infinity<br />The Place: The Universe<br />The Survivors: Arthur Dent (Earthling) - a mild-mannered, out-to-lunch space traveler, searching the cosmos for a decent cup of tea. Ford Prefect (alien) - unflappable researcher for the most successful and useful book, The Hitchhiker's Guide to the Galaxy. Trillian (earthling) - sexy space cadet who gave Arthur the cold shoulder at a cocktail party; she is now the girl friend of Zaphod Beeblebrox (alien) - two headed, three-armed ex-head honcho of the Universe.<br /><br />This fantastic foursome encounters paranoid androids, existential elevators, improbability drives, and kill-o-zap blasters on an interstellar romp that takes them from a thirty megaton rock concert on Kakrafoon to Milliways, the restaurant at the end of the Universe, where the main dish of the day introduces itself and the floor show is doomsday.<br /><br />How will it all end? Will they discover the origin of the bathtub? Will they find the significance of gin and tonic? Will they find out the question to the answer 42? For the answers to theses and other cosmic questions, climb aboard and Douglas Adams's zany spaceship and get ready to blast off. Take an improbable ride; your table is ready at The Restaurant at the end of the Universe.<br /><br />--jacket flap»

Таким образом, требуется построить некое отображение интересов пользователя, на основе описаний прочитанных им книг. В итоге довольно размытая задача построение профиля сводится к задаче текстового анализа, а задача рекомендаций сводится к задаче классификации, регрессии или ранжирования.

Распределение данных, построенное с помощью word2vec, разделение на кластеры с помощью kmeans



**Алгоритм получения рекомендаций**



|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | precision | recall | f1-score | support |
| 0 | 0.47 | 0.51 | 0.49 | 13306 |
| 1 | 0.72 | 0.69 | 0.70 | 24219 |
| accuracy |  |  | 0.62 | 37525 |
| macro avg | 0.59 | 0.60 | 0.60 | 37525 |
| weighted avg | 0.63 | 0.62 | 0.63 | 37525 |

В качестве классификатора использовали метод к-ближайших соседей, по итогу получили accuracy = 0.62.