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

Кафедра информатики и программирования

Реализация и сравнительный анализ методов построения рекомендательных систем

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 4 курса 441 группы
направления 02.03.03 Математическое обеспечение и администрирование информационных систем
факультета компьютерных наук и информационных технологий
Овчаровой Анастасии Александровны

Научный руководитель:
зав. кафедрой, к.ф.-м.н., доцент _____ Огнева М. В.

подпись, дата

Зав. кафедрой:
к.ф.-м.н., доцент _____ Огнева М. В.

подпись, дата

Саратов 2020

ВВЕДЕНИЕ

Актуальность темы.

На данный момент, объем информации и количество услуг, доступных пользователям, значительно растет. Количество услуг настолько велико, что пользователь не может физически посмотреть все возможные предложения. Именно поэтому системы рекомендаций становятся неотъемлемой частью веб-приложений, предоставляющих пользователю свои услуги. В качестве услуг могут быть представлены как товары в интернет-магазинах, так и всевозможный мультимедиа-контент: книги, музыка, фильмы, видео, приложения, игры, а также другие пользователи.

Системы рекомендаций используются, по большей части, для предоставления пользователю объектов, которые могут его заинтересовать в данный момент времени. Довольно часто они используются в электронной коммерции. В последнее время эти системы используют также в розничной торговли, справочных центрах, при поиске ПО, для научных статей и т.п. Это можно объяснить тем, что рекомендации предоставляются пользователю автоматически, основываясь на тех действиях, которые он (или другие пользователи) уже совершил (покупки, выставленные оценки и т.д.) и на основе обратной связи (заказы в магазинах, переход по ссылкам и т.п.).

Существуют различные способы построения рекомендательных систем, основанные на контентной фильтрации, коллаборативной фильтрации, гибридной фильтрации и т.д., каждый из которых имеет свои достоинства и недостатки.

Цель бакалаврской работы – построение рекомендательной системы по поиску нужной информации с использованием различных алгоритмов фильтрации, а также их сравнительная характеристика.

Поставленная цель определила **следующие задачи**:

1. Дать определение рекомендательной системы и рассмотреть подходы к ее построению;
2. Рассмотреть алгоритм ассоциативных правил;

3. Рассмотреть алгоритм коллаборативной фильтрации;
4. Рассмотреть алгоритм контентной фильтрации;
5. Реализовать алгоритм ассоциативных правил;
6. Применить алгоритмы ассоциативных правил для построения рекомендательной системы фильмов;
7. Применить алгоритм коллаборативной фильтрации для построения рекомендательной системы фильмов;
8. Применить алгоритм контентной фильтрации для построения рекомендательной системы фильмов;
9. Применить алгоритмы коллаборативной и контентной фильтраций для построения простой гибридной рекомендательной системы фильмов;
10. Сравнить результаты работы полученных систем.

Методологические основы рекомендательных систем представлены в работах Николенко С.А., Ломаш Д. А., Хлопина К. В., Xiaoyuan Su and Taghi M. Khoshgoftaar, Князевой А.А., Колобова О.С..

В теоретической части необходимо разобрать понятие рекомендательной системы, рассмотреть подходы к ее построению, в частности, алгоритм коллаборативной фильтрации, алгоритм контентной фильтрации и алгоритм на основе ассоциативных правил.

В практической части необходимо:

1. Реализовать рекомендательную систему на основе коллаборативной фильтрации;
2. Рекомендательную систему на основе контентной фильтрации;
3. Рекомендательную систему на основе ассоциативных правил;
4. Простую гибридную рекомендательную систему на основе контентной и коллаборативной фильтраций;
5. Сравнить полученные рекомендательные системы.

Структура и объём работы. Бакалаврская работа состоит из введения, 6 разделов, заключения, списка использованных источников и 4 приложений. Общий объем работы – 83 страниц, из них 23 страницы – основное

содержание, включая 0 рисунков и 22 таблицы, список использованных источников информации – 23 наименования.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Основные понятия» посвящен определению рекомендательной системы, алгоритмов ее реализации, а также метрик сравнения.

Рекомендательные системы – методы и инструменты для предложения пользователям объектов для их использования. Формально, задача ставится следующим образом: имеется множество пользователей и множество объектов. Необходимо для пользователя предложить набор объектов так, чтобы максимизировать функционал качества Q.

Можно выделить три основных подхода к построению рекомендательных систем:

1. Контентная фильтрация (content-based);

2. Коллаборативная фильтрация (collaborative filtering);

3. Гибридная фильтрация.

Помимо этих подходов существует множество других, один из них— метод ассоциативных правил.

Контентная фильтрация основана на принципе подбора объектов, похожих по тем или иным характеристикам на те, которые ранее понравились пользователю системы. Контентная рекомендательная система основана на методах поиска, сравнения и фильтрации информации. Описанный подход, в основном, используется для текстов: документы, сайты, блоги; или объекты, которые могут быть описаны с ключевыми словами. В рекомендательных системах, основанных на контентной фильтрации, релевантность рекомендуемых объектов находится из оценок пользователя, данных похожим объектам. Например, чтобы предложить человеку документ, рекомендательная система на основе контентной фильтрации пытается найти схожесть между различными объектами, которые ранее были высоко оценены пользователем (сравниваются тематики, авторы и, возможно, даже главы документов).

Коллаборативная фильтрация является методом построения рекомендательных систем, который основан на предположении, что пользователи с аналогичными оценками ранее просмотренных элементов оценят то же самое в будущем. Группа пользователей, которая наиболее похожа на данного пользователя, называется «соседями».

В основном, коллаборативная фильтрация используется для фильтрации информации, которая была получена при взаимодействии большого количества людей, данных и т.д. Приложения, которые используют коллаборативную фильтрацию, чаще всего работают с большими объемами данных.

Среди методов коллаборативной фильтрации можно выделить два: «Основанный на пользователях» и «Основанный на сущностях».

Основной принцип подхода, основанного на пользователях: порекомендовать товары, которые покупают схожие пользователи, усреднить рейтинги товара, проставленные другими пользователями, с весами по степени схожести пользователей.

Основной принцип подхода, основанного на сущностях: порекомендовать товары, похожие на уже приобретенные, усреднить рейтинги уже оцененных товаров с весами по степени похожести на неоцененный товар.

Ассоциативные правила – основой этого алгоритма является использование априорных данных о частоте выбора клиентами определенных товарных групп. Основное правило данного алгоритма гласит: «Все подмножества множества продуктов с высоким спросом, также встречаются часто».

Гибридная рекомендательная система — объединяет несколько подходов в одной системе. Гибридная система, объединяющая методы A и B, пытается использовать преимущества A для исправления B.

Существует различные способы гибридизации. Рассмотрим некоторые из них:

1. Взвешенная комбинация;

2. Переключение;
3. Смесь рекомендаций;
4. Конвейер;
5. Комбинирование признаков;
6. Усиление признаков.

Для оценки качества работы существует множество метрик качества. В основном, это метрики оценки точности предполагаемого значения и реального, если таковой имеется. Рассмотрим некоторые из них.

RMSE (Root Mean Squared Error, пер. средняя квадратичная ошибка) – ошибка вычисляется как корень из суммы квадратов разниц между предсказываемым значением и реальным значением (1).

$$RMSE = \sqrt{\frac{\sum_{i \in n}(P_i - R_i)^2}{n}} \quad (1)$$

Далее можно оценить еще некоторые характеристики рекомендательных систем, основываясь на всем списке рекомендаций. Также, нужно иметь данные о том, какие позиции в списке соответствуют запросу, а какие нет. Определим длину всего списка рекомендаций как L, множество соответствующих рекомендаций – T, множество несоответствующих (нерелевантных) позиций в списке – F. Тогда характеристику точности можно представить в виде формулы (2).

$$Accuracy = \frac{\|T\|}{L} \quad (2)$$

Введем еще одно множество G – множество элементов, которые должны были быть порекомендованы, но в списке не оказались. Тогда характеристику полноты можно представить в виде формулы (3).

$$Re\ call = \frac{\|T\|}{\|T\| + \|G\|} \quad (3)$$

Второй раздел «Практическая часть» посвящен реализации рекомендательных систем и их сравнительному анализу.

В данном разделе были реализованы:

1. Рекомендательная система на основе ассоциативных правил;
2. Рекомендательная система на основе коллаборативной фильтрации;
3. Рекомендательная система на основе контентной фильтрации;
4. Простая гибридная рекомендательная система на основе коллаборативной и контентной фильтраций.

После чего был проведен сравнительный анализ полученных рекомендательных систем. Для этого были реализованы такие метрики:

1. RMSE;
2. Recall;
3. Accuracy.

Сравнительный анализ показал, что:

1. Рекомендательная система на основе контентной фильтрации работает очень медленно и неэффективно, так как рекомендует все фильмы, похожего жанра;

2. На больших данных рекомендательная система на основе контентной фильтрации работает приблизительно до нескольких часов;

3. Рекомендательная система на основе коллаборативной фильтрации оказалась самой быстрой в выполнении и второй по точности;

4. Рекомендательная система на основе ассоциативных правил работает довольно быстро, т.к. сразу выводит общие рекомендации для всех пользователей, но у данной системы невысокая точность и высокий уровень ошибки за счет того, что она не учитывает предпочтения конкретного пользователя;

5. Гибридная рекомендательная система дает более точные результаты и работает эффективнее, за счет того, что контентная фильтрация дополнительно проверяет предсказания коллаборативной;

6. Несмотря на эффективность, время выполнения гибридной рекомендательной системы больше, чем время выполнения рекомендательной системы на основе коллаборативной фильтрации.

Гибридизация, действительно, способна помочь улучшить результаты рекомендательных систем, причем даже относительно плохо работающая система может помочь улучшить результат хорошо работающей системы при использовании гибридизации.

ЗАКЛЮЧЕНИЕ

В данной работе были разобраны и построены рекомендательные системы выполняющие поиск и выдачу рекомендаций фильмов

1. На основе контентной фильтрации;
2. На основе коллаборативной фильтрации;
3. На основе ассоциативных правил.

После чего, была построена простая гибридная рекомендательная система, на основе коллаборативной и контентной фильтраций.

Для каждой из этих систем были получены оценки точности и ошибки, а также время выполнения. На основе этих данных был проведен сравнительный анализ, который показал, что:

1. Рекомендательная система на основе контентной фильтрации работает медленно и неэффективно;
2. Рекомендательная система на основе коллаборативной фильтрации оказалась самой быстрой в выполнении и второй по точности;
3. Рекомендательная система на основе ассоциативных правил работает довольно быстро, но у данной системы невысокая точность и высокий уровень ошибки;
4. Гибридная рекомендательная система дает более точные результаты и работает эффективнее, но, несмотря на это, имеет не самое быстрое время выполнения.

Основные источники информации:

1. Николенко С.А. Рекомендательные системы. СПб // Центр Речевых Технологий — 2012 — С. 53.

2. Ломаш Д. А., Хлопин К. В. Вестник Ростовского государственного университета путей сообщения // Ростовский государственный университет путей сообщения — 2013— С. 75-84.

3. Xiaoyuan Su and Taghi M. Khoshgoftaar. A Survey of Collaborative Filtering Techniques// Hindawi Publishing Corporation, 2009.

4. Князева А.А., Колобов О.С., Турчановский И.Ю. «Способы построения гибридной рекомендательной системы на основе данных о заказах библиотеки» // Институт вычислительных технологий, Томский филиал, г. Томск – 2015.

5. Сергей Николаенко Рекомендательные системы. [Электронный ресурс].– URL: chrome extension://ilhapdfjlmhfdgdbefpinebijmhjijpn/https://logic.pdmi.ras.ru/~sergey/teaching/mlstc12/15-recommender.pdf (Дата обращения 10.11.2019).

6. Абсолютное юзабилити. Эра рекомендательных систем. [Электронный ресурс].– URL: https://rusability.ru/usability/absolyutnoe-yuzabiliti-era-rekomendatelnyih-sistem/ (Дата обращения 10.11.2019).

7. Алгоритмы выделения ассоциативных правил. [Электронный ресурс].– URL: https://ranalytics.github.io/data-mining/054-Association-Rules-Algos.html (Дата обращения 10.11.2019).

8. Поиск ассоциативных правил. [Электронный ресурс].– URL: https://ami.nstu.ru/~vms/lecture/data_mining/rules.htm (Дата обращения 10.11.2019).