

Усовершенствование коллаборативной фильтрации

Ю. А. Морозов¹, С. Э. Сараджишвили²

Санкт-Петербургский политехнический университет Петра Великого (СПбПУ)

¹stonefiz@gmail.com, ²SSaradg@yandex.ru

Аннотация. В данной работе проводится эксперимент комбинации метрик для расчета схожести, при создании коллаборативной фильтрации. Используется коэффициент Отиаи и евклидово расстояние, в итоге получается рекомендательная система, которая выдает удовлетворительный результат.

Ключевые слова: рекомендательные системы, коллаборативная фильтрация, коэффициент Отиаи, евклидово расстояние

I. РЕКОМЕНДАТЕЛЬНЫЕ СИСТЕМЫ

В последние годы рекомендательные системы получают всё большее распространение в интернете, в различных сервисах будь то рекомендации видео, музыки или товаров.

Рекомендательные системы – это системы, которые предназначены для того, чтобы рекомендовать пользователю информацию на основании множества различных критериев. [1] Рекомендации, как правило, основываются на данных о конкретном пользователе, на информации о том, каким образом пользователь взаимодействовал с системой ранее. Таким образом, формируются рекомендации специально для данного пользователя.

Большинство рекомендательных систем при реализации персональных рекомендаций также могут учитывать принадлежность пользователя к группе и популярность контента. В рекомендательных системах выделяют: множество субъектов (пользователи, клиенты), множество объектов (предметы, артисты, ресурсы) и пространство откликов множества субъектов над множеством объектов. В случае рекомендательной системы похожих музыкальных исполнителей в качестве объектов выделяются музыкальные исполнители, в качестве субъектов выделяются пользователи или клиенты этой системы, а пространство откликов можно характеризовать как количество прослушивания пользователями различных исполнителей.

Существуют различные виды и подходы для реализации рекомендательных систем. Основными можно назвать следующие: контент-ориентированные – это те системы, которые рекомендуют что-либо на базе знаний об услугах или объектах; на основе знаний – пользователям рекомендуют товары на основе тех, что они уже положили в корзину или создают подборки, связанные с теми, которые они уже посмотрели и коллаборативная

фильтрация – например, первый и второй пользователь играют на гитаре, в приставку и любят пиццу. Третий пользователь тоже любит пиццу и учится играть на гитаре. Приставку он покупать не планировал, но рекомендуемая система предположит, что она ему нужна на основе пересечений с другими пользователями.

Мы будем рассматривать коллаборативную фильтрацию в рекомендательных системах, на примере реализации системы рекомендации музыкального контента.

II. КОЛЛАБОРАТИВНАЯ ФИЛЬТРАЦИЯ

К коллаборативной фильтрации относятся те методы и алгоритмы, которые основываются на данных о предыдущих сеансах работы пользователей с этой же системой. [1] Коллаборативную фильтрацию можно схематично визуализировать, таким образом, как это показано на рисунке ниже.



Рис. 1. Визуализация коллаборативной фильтрации

Данный вид рекомендательных систем может включать в себя множество методов. Общей чертой для них всех является выбор общих данных. В этих методах фильтрации в качестве источника для создания рекомендаций используются пересечения оценок/количество прослушиваний/количество просмотров контента пользователями. Считается, что с течением времени люди мало меняют свои вкусы, и, если вы в чем-то с кем-то согласились в прошлом, вы, вероятно, согласитесь с ним и в будущем.

На рисунке ниже представлен пример совместной оценки пользователями информации. Под знаком вопроса должен быть результат, который рассчитан на основе общих данных, используя методы коллаборативной фильтрации.

A					
B					
C					
D					
E				?	

Рис. 2. Пример системы, основанной на совместной фильтрации

Мы будем рассматривать настройку коллаборативной фильтрации для рекомендательной системы музыки. С использованием данных музыкальных плейлистов с сайта Last.fm [2]

A. Подготовка

На начальном этапе мы подготовили и обработали данные, с которыми будем работать в дальнейшем. Изначально у нас на входе имеются данные в нескольких файлах о пользователях, артистах и прослушиваниях пользователями этих артистов. В файлах о пользователях, содержатся все существующие пользователи в системе last.fm, в данных об артистах и прослушиваниях – сведения об артистах и количестве их прослушиваний.

Мы осуществляем проверки, что пользователи действительно существуют путем проверки соответствия данных о пользователях в файле о них и их существования и удаления лишних в файлах с данными об артистах и прослушиваниях, а также очищаем данные от нулевых значений, неизвестных исполнителей и пропущенных значений.

Далее, мы выделяем множество субъектов, множество объектов и пространство откликов множества субъектов над множеством объектов, формируя сводные матрицы вида: пользователь, исполнитель – показатель прослушивания для дальнейшего преобразования и расчета, похожих друг на друга исполнителей.

B. Евклидово расстояние

Когда матрицы подготовлены для расчетов. Первым делом мы будем рассчитывать Евклидово расстояние, позволяющее измерить расстояние между двумя точками. Это кратчайшее расстояние от одной точки до другой по прямой. Нетрудно понять, что эта прямая является гипотенузой прямоугольного треугольника [2].

Формула для вычисления евклидова расстояния выглядит следующим образом:

$$d(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}$$

Мы будем рассчитывать евклидово расстояние для пар пользователь – исполнитель. После проведенных расчетов мы получим матрицу следующего вида, изображенную на рис. 3.

a day to remember	a perfect circle	a tribe called quest	abba	ac/dc	aerosmith	aerop rock	af1	against me!	air	y1
0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	
0.0	0.0	0.000000	0.0	0.0	0.0	0.021277	0.0	0.0	0.021277	
0.0	0.0	0.032258	0.0	0.0	0.0	0.000000	0.0	0.0	0.032258	
0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.037037	
0.0	0.0	0.000000	0.0	0.0	0.0	0.000000	0.0	0.0	0.000000	

Рис. 3. Рассчитанная матрица с Евклидовым расстоянием

C. Коэффициент Отиаи

Коэффициент Отиаи – это бинарная мера сходства, которая часто используется в рекомендательных системах [4]. Данный коэффициент рассчитывает косинус угла между векторами. Мы будем использовать его для того, чтобы сравнить вектора артистов. Благодаря, предыдущим расчетам мы уже рассчитали коэффициенты расстояния между векторами. Теперь необходимо применить угловой коэффициент Отиаи, который будет характеризовать следующим образом соотношения: чем меньше угол, тем ближе прослушивания/похожесть артистов друг на друга. Формула имеет следующий вид:

$$\text{similarity}(A, B) = (\sum_{i=1}^n A_i \times B_i) / (\sqrt{\sum_{i=1}^n A_i^2} \times \sqrt{\sum_{i=1}^n B_i^2})$$

D. Результаты расчетов

После применения рассчитанных коэффициентов к матрицам мы получим матрицу такого вида, как показано на рисунке ниже:

	2pac	3 doors down	30 seconds to mars	50 cent	65daysofstatic	a day to remember
2pac	1.000000	0.007332	0.021072	0.198916	0.015376	0.008932
3 doors down	0.007332	1.000000	0.067555	0.008503	0.002758	0.000000
30 seconds to mars	0.021072	0.067555	1.000000	0.007345	0.027696	0.014395
50 cent	0.198916	0.008503	0.007345	1.000000	0.012034	0.000000
65daysofstatic	0.015376	0.002758	0.027696	0.012034	1.000000	0.005081
...	...	...	...	...	...	...
wilee	0.009770	0.001095	0.000000	0.007940	0.013463	0.000000
within temptation	0.012041	0.029584	0.015477	0.008238	0.002805	0.000000
yann tiersen	0.006791	0.013091	0.008044	0.008836	0.043058	0.000000

Рис. 4. Рассчитанная матрица соответствий

Е. Тестирование рекомендаций

После того, как мы получили матрицу соответствий. Можно визуализировать похожих друг на друга исполнителей на основе наших данных. На рисунке ниже визуализация похожих исполнителей на Eminem.

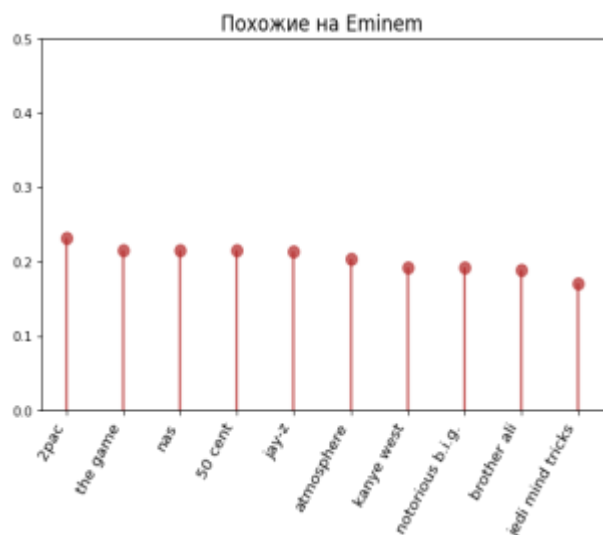


Рис. 5. Похожие исполнители на Eminem

По данному рисунку отображаются такие исполнители как 2pac, the game, 50 cent и другие рэп-исполнители. Стоит отметить, что данные исполнители принадлежат жанру рэп, как и исходный, на которого ищутся похожие. Также они похожи по звучанию и смысловому содержанию. Можно сказать, что в данном случае система рекомендаций отработала корректно.

На рис. 6 можно увидеть исполнителей, которые похожи на Rihanna.

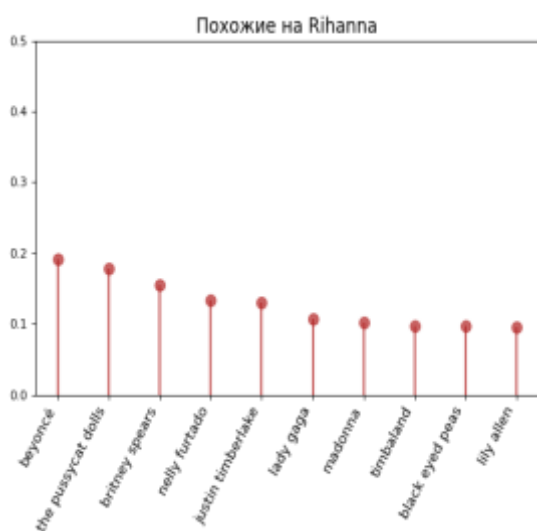


Рис. 6. Похожие исполнители на Rihanna

Исполнители похожие на Rihanna, как видно на рисунке – это такие исполнители как Beyonce, Britney spears, The pussycat dolls и другие. Все они похожи по

жанрам, мотиву музыки и даже в большинстве своем по половой принадлежности солисток.

Для следующего примера мы возьмем такого исполнителя как группа – Queen и посмотрим, какой результат будет для них. На рис. 7 изображен результат для группы Queen.

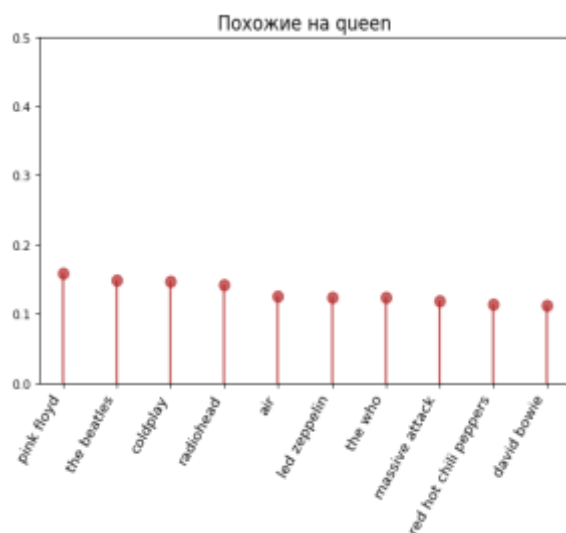


Рис. 7. Похожие исполнители на Queen

Исполнители похожие на Queen, как можно увидеть на рисунке – это такие исполнители как Pink Floyd, The Beatles, Coldplay, Radiohead, Air и другие. Данные исполнители по своей жанровой принадлежности действительно похожи, а также по звучанию, манере исполнения и т. д.

Проведем последнее тестирование и в качестве исполнителя возьмем такого исполнителя, как Madonna. Результат рассчитанных похожих исполнителей будет следующий:

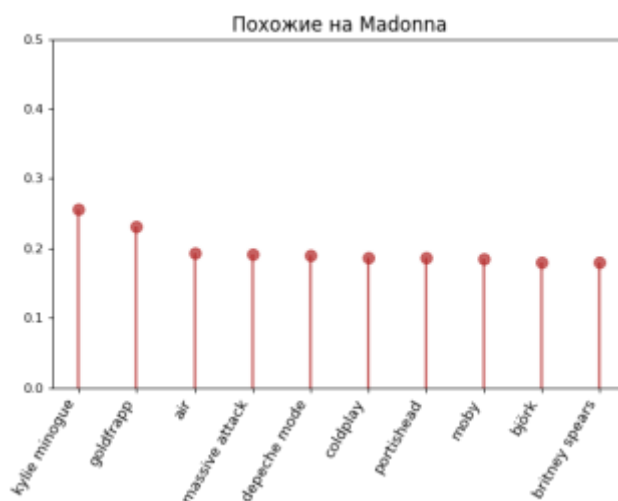


Рис. 8. Похожие исполнители на Madonna

Как можно увидеть на рисунке, исполнители похожие на Madonna – это такие исполнители как kylie Minogue,

goldfrapp, air, Britney spears и другие. И в самом деле, данные исполнители по жанровой принадлежности похожи и звучанию похожи на Madonna. Можно сделать вывод, что система рекомендаций для исполнителей отработала корректно.

III. ЗАКЛЮЧЕНИЕ

В заключении можно сделать вывод, что эксперимент с комбинацией коэффициентов Евклидова расстояния и коэффициента Отиаи для коллаборативной фильтрации удался, а предполагаемые похожие исполнители похожи друг на друга по жанровым принадлежностям и репертуару. Таким образом, наша настроенная

рекомендательная система работает корректно и соответствует требованиям к функциональности рекомендательной системы.

СПИСОК ЛИТЕРАТУРЫ

- [1] Фальк Ким. Рекомендательные системы на практике / пер. с англ. Д.М. Павлова. Москва: ДМК Пресс, 2020. 448 с.
- [2] Набор данных Last.fm dataset 360K. Режим доступа: <https://www.upf.edu/web/mtg/lastfm360k>.
- [3] Герасименко Е.М. Методы машинного обучения в научных исследованиях: Учебное пособие. Ростов-на-Дону: Южный федеральный университет, 2020.
- [4] Порошкина В.В. Меры подобия в рекомендательных системах // Аллея науки. 2019. № 31.С. 909-915.