АЛГОРИТМЫ ПОСТРОЕНИЯ ПЕРСОНАЛИЗИРОВАННЫХ МУЗЫКАЛЬНЫХ РЕКОМЕНДАТЕЛЬНЫХ СИСТЕМ

Орлова А.С.

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Петюкевич Н.С. – м.т.н.,ассистент

На сегодняшний день существует большое количество сервисов для прослушивания музыки. Отличительной чертой подобных сервисов является наличие персонализированного подхода к каждому пользователю. Задача персонализированного подхода заключается в том, чтобы, основываясь на предпочтениях пользователя, предложить музыкальную композицию, которая будет ему интересна. Решением данной задачи занимаются специальные программы, называемые рекомендательными системами.

Рекомендательные системы – это программные средства, которые способны предсказать, какие объекты будут интересны пользователю при наличии определённой информации о его предпочтениях. Под объектом понимается то, что именно система рекомендует пользователю [1]. В данном случае в качестве объекта выступает музыкальная композиция.

Существует две основные стратегии построения рекомендательных систем - фильтрация на основе содержания и коллаборативная фильтрация. При фильтрации на основе содержания создаются профили пользователей и объектов. Профили пользователей могут включать различную информацию о самом пользователе, а также его ответы на заранее составленные вопросы, профили объектов могут включать названия жанров, имена исполнителей, оценки пользователей. Метод коллаборативной фильтрации не требует определения множества характеристик, а основывается только на информации о действиях и поведении пользователей в прошлом. Далее будут рассматриваться алгоритмы, применяемые для построения рекомендательных систем методом коллаборативной фильтрации.

В некоторых коллаборативных рекомендательных системах используется алгоритм Байеса. В его основе лежит теорема Байеса, которая заключается в вычислении вероятности наступления события (выставления пользователем той или иной оценки) в условиях, когда на основе наблюдений известна лишь некоторая информация о событиях. Основным недостатком алгоритма Байеса является требование большой выборки для обучения, а также тот факт, что значения спрогнозированных вероятностей не всегда являются достаточно точными. Поэтому в последнее время используются другие алгоритмы.

В центре большинства рекомендательных систем находится так называемая матрица предпочтений. Это матрица, строки которой соответствуют пользователям, а столбцы композициям (рисунок 1). На пересечении некоторых строк с некоторыми столбцами матрица заполнена оценками. Оценка – это показатель того, насколько пользователю понравилась прослушанная композиция по заданной шкале (например, от 1 до 5). Незаполненные поля соответствуют непрослушанным композициям. Основное допущение данного метода состоит в том, что те, кто одинаково оценивали какие-либо объекты в прошлом, склонны давать похожие оценки другим объектам и в будущем.

	Track 1	Track 2	Track 3	Track 4	Track 5
User 1	5	1	3	5	2
User 2	?	?	?	2	5
User 3	4	?	3	?	1
User 4	1	5	5	3	3

Рисунок 1 – Матрица оценок пользователей

Задача заключается в том, чтобы предположить, как оценил бы пользователь ту или иную композицию. Например, какую оценку поставил бы третий пользователь песне номер 2 или 4.

Самым популярным алгоритмом для нахождения данной оценки является алгоритм *k*-ближайших соседей [2]. Алгоритм включает следующие шаги:

– вычисление степени схожести пользователей;
– нахождение *k* самых похожих пользователей;
– рекомендация пользователю тех композиций, которые он еще не прослушивал.

Оценки каждого пользователя можно представить в виде вектора. Чем ближе два вектора находятся друг к другу, тем более похожими являются соответствующие оценки пользователей, а

значит, и более похожими являются их интересы. Расстояние между векторами в свою очередь можно найти из векторного произведения. Запишем векторное произведение двух векторов:

$$A \cdot B = \|A\| \|B\| \cos\theta,$$

где θ - угол между векторами.

Чем меньше значение угла θ, тем ближе друг к другу находятся векторы. Степень схожести пользователей может быть выражена как значение косинуса угла между векторами (чем ближе значение косинуса к 1, тем более похожие интересы у пользователей):

$$\cos\theta = \frac{A \cdot B}{\|A\| \|B\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \sqrt{\sum_{i=1}^{n} B_i^2}}.$$

После того, как найдены пользователи со схожими интересами, целевому пользователю могут быть порекомендованы композиции.

Зачастую такие матрицы разрастаются в объемах, становятся слишком большими и их неудобно, сложно и долго обрабатывать. Существует ещё один алгоритм, который используется для создания рекомендательных систем - факторизация матриц. С помощью этого метода изначальную матрицу предпочтений можно представить в виде произведения двух прямоугольных матриц меньшей размерности (рисунок 2). Одна матрица может быть представлена как матрица пользователей, где строки - это пользователи, а столбцы - это скрытые признаки. Вторая матрица - матрица песен, где строки - это скрытые признаки, а столбцы представляют собой сами песни.

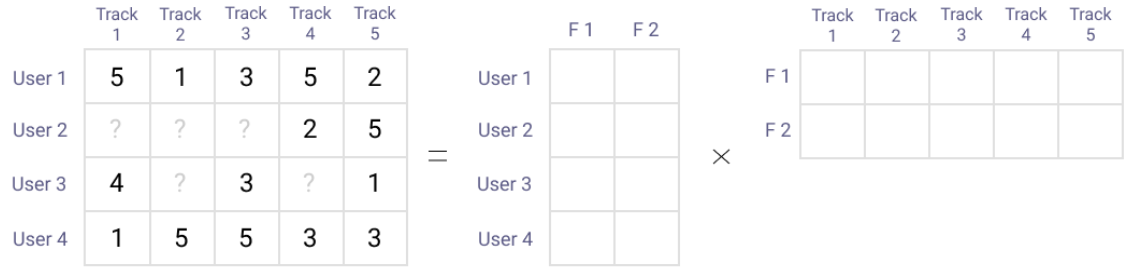

Рисунок 2 – Разложение матрицы оценок пользователей

Чем больше количество признаков, тем точнее будет приближение оценки, но в то же время алгоритм будет сложнее, работать будет дольше и размер матриц будет больше. Количество признаков следует выбирать разумно, проанализировав конкретную задачу.

Одним из методов разложения матриц является мультипликативное правило обновления [3].

где A – исходная матрица оценок размерностью $n \cdot m$;

$$H_{ij} = H_{ij} \frac{(W^T A)_{ij}}{(W^T W H)_{ij} + \varepsilon}, \qquad W_{ij} = W_{ij} \frac{(A H^T)_{ij}}{(W H H^T)_{ij} + \varepsilon},$$

H – матрица песен, заполненная начальными случайными значениями, размерностью $k \cdot m$;

W – матрица пользователей, заполненная начальными случайными значениями, размерностью $n \cdot k$.

Данный метод является итеративным, в конце каждой итерации вычисляется ошибка. Алгоритм останавливается, когда значение ошибки приближается к 0:

$$\frac{\|A - WH\|}{\|A\|} \approx 0.$$

Данный метод решает проблему обработки большого количества данных, а также даёт хорошее приближение оценок пользователей.

Список использованных источников:

1. Francesco Ricci, Lior Rokach, Bracha Shapira, Paul B. Kantor. Recommender Systems Handbook. 2011.
2. Lars Elden. 2007. Matrix Methods in Data Mining and Pattern Recognition (Fundamentals of Algorithms). Soc. for Industrial and Applied Math., Philadelphia, PA, USA.
3. Michael W. Berry, Murray Browne, Amy N. Langville, V. Paul Pauca, Robert J. Plemmons, Algorithms and applications for approximate nonnegative matrix factorization, Computational Statistics & Data Analysis, Volume 52, Issue 1, 2007, p. 155-173.