

# Автоматическое дополнение плейлистов в рекомендательной системе\*

*Кислинский В. Г., Фролов Е., Воронцов К. В.*

kislinskiy.vg@phystech.edu; evgeny.frolov@skolkovotech.ru; vokov@forecsys.ru

Московский физико-технический институт

Работа посвящена исследованию метода совместной матричной факторизации в задаче top-N рекомендаций для автоматического продолжения плейлистов. Предлагается использовать модель матричной факторизации, учитывающую дополнительную информацию о плейлистах и треках. В данном методе латентное пространство приобретает дополнительную структуру, соответствующую признаковым описаниям плейлистов и треков. Также будет введена дополнительная регуляризация, основанная на предположении, что если объекты близки в пространстве признаков, то они также близки и в латентном факторном пространстве. Для анализа качества представленного алгоритма проводятся эксперименты на выборке из миллиона плейлистов MPD.

**Ключевые слова:** *задача top-N рекомендаций, совместная матричная факторизация, алгоритм LSE, латентное факторное пространство, коллаборативная фильтрация.*

## 1 Введение

Существует два основных подхода к задаче автоматического продолжения плейлистов: музыкальный информационный поиск и рекомендательные системы[1]. Рекомендательные системы имеют несколько основных техник, которые можно разделить на две основные группы: коллаборативная фильтрация и content-based подход. Content-based системы используют свойства объектов, создают профили пользователей, опираясь на свойства объектов, которые пользователь предпочел ранее, и рекомендуют новым пользователям наиболее подходящие к их профилю объекты. Коллаборативная фильтрация использует историю действий пользователей в системе для получения новых рекомендаций. Методы коллаборативной фильтрации основаны на предположении о том, что те, кто одинаково оценивали какие-либо объекты в прошлом, склонны давать похожие оценки в будущем. Такие методы обладают хорошей устойчивостью и способны выявлять скрытые свойства объектов, что улучшает релевантность рекомендаций. Матричная факторизация является одним из методов коллаборативной фильтрации, такой подход имеет сходство с семантическим анализом[2]. Каждому пользователю и объекту, сопоставляются некоторые факторы, которые определяют их скрытые интересы и свойства. Основным недостатком этого подхода является проблема холодного старта - неспособность строить рекомендации для новых пользователей и рекомендовать новые объекты.

В данной работе предполагается, что использование свойств объектов и пользователей улучшит качество рекомендаций, получаемых методами коллаборативной фильтрации и решит проблему холодного старта. Будет исследована модель совместной матричной факторизации, подобный подход предлагается в работе[3], где матрица объект-признак факторизуются вместе с матрицей объект-пользователь. Существуют различные техники учета дополнительной информации в модели матричной факторизации, введение допол-

нительной регуляризации[4], совместная факторизация с матрицами похожести объектов и пользователей[5].

Основной целью работы является решение проблемы холодного старта для пользователей. В задаче автоматического дополнения плейлистов пользователи это плейлисты, объекты - треки. Для плейлистов, которые содержат достаточно большой набор треков, хорошо работают методы коллаборативной фильтрации, но для новых плейлистов, которые пока содержат мало треков или не содержат совсем, коллаборативная фильтрация не может построить релевантные рекомендации. Если же строить рекомендации на основе свойств плейлистов и треков, то для новых плейлистов можно получить более подходящие рекомендации, например, был создан плейлист с именем "рок", тогда будет логичнее предложить наиболее популярные треки из рок-музыки, а не что-то другое, но такой content-based подход будет строить слишком тривиальные рекомендации для плейлистов, которые уже имеют много треков. Совместная матричная факторизация объединяет в себе оба этих подхода, будут находиться такие профили плейлистов, которые зависят не только от матрицы плейлист-трек, но и от матрицы плейлист-признак. Также будет рассмотрено предположение о том, что из близости плейлистов в пространстве признаков, следует близость в латентном факторном пространстве, пространстве профилей, для проверки этого предположения будет введена дополнительная регуляризация[3].

## 2 Постановка задачи

Задано множество треков  $\mathcal{T} = \{t_i\}_{i=1}^m$  и множество плейлистов  $\mathcal{P} = \{p_i\}_{i=1}^n$ , где  $p_i \in \mathcal{T}$ ,  $|p_i| \ll |\mathcal{T}|$ . Каждый трек описывается исполнителем и альбомом, обозначим  $\mathcal{S} = \{s_i\}_{i=1}^k$  - множество исполнителей,  $\mathcal{L} = \{l_i\}_{i=1}^d$  - множество альбомов, а каждый плейлист имеет название,  $\mathcal{A} = \{a_i\}_{i=1}^r$  - множество названий. Определим матрицу  $\mathbf{R} \in \mathbb{R}^{n \times m}$ , следующим образом  $\mathbf{R}_{ij} = 1$ , если  $t_j \in p_i$ , иначе ноль, также зададим бинарную матрицу  $\mathbf{X} \in \mathbb{R}^{n \times (k+d+r)}$ , которая описывает треки, каких авторов содержатся в плейлисте -  $k$  первых столбцов, каких альбомов  $d$  следующих столбцов, имя плейлиста -  $r$  последних столбцов. Надо для нового плейлиста  $p$  найти  $N$  наиболее подходящих треков, для этого будет строиться вектор  $\mathbf{r} \in \mathbb{R}^{m \times 1}$ ,  $i$ -ый элемент которого означает насколько трек  $t_i$  подходит плейлисту  $p$ .

## 3 Описание метода

Задача матричной факторизации заключается в нахождение двух матриц меньшей размерности, произведение которых приближает исходную. В задаче рекомендаций ищется приближение матрицы рейтингов, в наших терминах  $\mathbf{R} \approx \mathbf{UV}$ , где  $\mathbf{U}$  - матрица профилей плейлистов,  $\mathbf{V}$  - матрица профилей треков. Предполагая, что профили плейлистов зависят от того какие исполнители, альбомы входят в плейлист, какие названия у плейлистов, можно записать  $\mathbf{X} \approx \mathbf{UH}$ , где  $\mathbf{H}$  - матрица, задающая малоразмерное латентное представление авторов, альбомов, названий. Таким образом приходим к следующей задаче оптимизации:

$$\begin{aligned} \mathbf{U}, \mathbf{V}, \mathbf{H} = \arg \min_{\mathbf{U}, \mathbf{V}, \mathbf{H}} & \alpha \|\mathbf{R} - \mathbf{UV}\|_F^2 + (1 - \alpha) \|\mathbf{X} - \mathbf{UH}\|_F^2 + \\ & \lambda (\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2 + \|\mathbf{H}\|_F^2), \\ & s.t. \mathbf{U} \geq 0, \mathbf{V} \geq 0, \mathbf{H} \geq 0 \end{aligned} \quad (1)$$

где  $\lambda$  положительный коэффициент регуляризации, а  $\alpha \in [0, 1]$ , если  $\alpha > 0.5$ , то матрица  $\mathbf{U}$  больше зависит от распределения треков по плейлистам, если  $\alpha < 0.5$ , то наоборот, пред-

полагается, что  $U$  зависит от распределения авторов, альбомов по плейлистам и названий плейлистов.

### 3.1 Введение дополнительной регуляризации на основе близости плейлистов в пространстве признаков

Близость плейлистов в пространстве признаков можно оценивать евклидовым расстоянием. Составим граф близости плейлистов, в котором вершинами будут плейлисты, каждый плейлист соединим ребром с  $l$  ближайшими плейлистами, где  $l$  ранг разложения матриц  $\mathbf{R}$  и  $\mathbf{X}$ . По полученному графу составляется матрица смежности  $\mathbf{A}$ , где  $A_{ij} = 1$  если вершина  $i$ , соединена с вершиной  $j$ , ноль иначе. Предполагая, что если плейлисты близки в реальном пространстве признаков, то они также близки в пространстве профилей, с помощью матрицы  $\mathbf{A}$  можно определять близость, соответственных профилей:

$$\begin{aligned} S &= \frac{1}{2} \sum_{i,j=1}^n \|u_i - u_j\|^2 A_{ij} = \sum_{i=1}^n (u_i^T u_i) D_{ii} - \sum_{i,j=1}^n (u_i^T u_j) A_{ij} = \\ &= Tr(\mathbf{U}^T \mathbf{D} \mathbf{U}) - Tr(\mathbf{U}^T \mathbf{A} \mathbf{U}) = Tr(\mathbf{U}^T \mathbf{L} \mathbf{U}) \end{aligned} \quad (2)$$

где  $\mathbf{D}$  диагональная матрица, на  $ii$  месте которой стоит сумма  $i$  строки матрицы  $\mathbf{A}$ , матрица  $\mathbf{L} = \mathbf{D} - \mathbf{A}$ . Перепишем (1), учитывая  $S$ :

$$\begin{aligned} \mathbf{U}, \mathbf{V}, \mathbf{H} &= \arg \min_{\mathbf{U}, \mathbf{V}, \mathbf{H}} \alpha \|\mathbf{R} - \mathbf{U}\mathbf{V}\|_F^2 + (1 - \alpha) \|\mathbf{X} - \mathbf{U}\mathbf{H}\|_F^2 + \\ &\quad \beta Tr(\mathbf{U}^T \mathbf{L} \mathbf{U}) + \lambda (\|\mathbf{U}\|_F^2 + \|\mathbf{V}\|_F^2 + \|\mathbf{H}\|_F^2) \\ &\quad s.t. \mathbf{U} \geq 0, \mathbf{V} \geq 0, \mathbf{H} \geq 0 \end{aligned} \quad (3)$$

### 3.2 Оптимизационная проблема

Полученная задача оптимизации не является выпуклой по параметрам  $(\mathbf{U}, \mathbf{V}, \mathbf{H})$ , следовательно поиск глобального минимума затруднительная задача. Предлагается итеративный алгоритм поиска стационарной точки оптимизируемого функционала, который был получен и доказан в статье[3]. Матрицы  $(\mathbf{U}, \mathbf{V}, \mathbf{H})$  будут обновляться по следующим правилам:

$$\begin{aligned} \mathbf{U} &= \mathbf{U} \odot \frac{\alpha \mathbf{R} \mathbf{V}^T + (1 - \alpha) \mathbf{X} \mathbf{H}^T + \beta \mathbf{A} \mathbf{U}}{\alpha \mathbf{U} \mathbf{V} \mathbf{V}^T + (1 - \alpha) \mathbf{U} \mathbf{H} \mathbf{H}^T + \beta \mathbf{D} \mathbf{U} + \lambda \mathbf{U}} \\ \mathbf{V} &= \mathbf{V} \odot \frac{\alpha \mathbf{U}^T \mathbf{R}}{\alpha \mathbf{U}^T \mathbf{U} \mathbf{V} + \lambda \mathbf{V}} \\ \mathbf{H} &= \mathbf{H} \odot \frac{(1 - \alpha) \mathbf{U}^T \mathbf{X}}{(1 - \alpha) \mathbf{U}^T \mathbf{U} \mathbf{H} + \lambda \mathbf{H}}, \end{aligned}$$

где  $\odot$  - поэлементное умножение,  $\oslash$  - поэлементное деление.

### 3.3 Получение рекомендаций

Для того чтобы получить рекомендации для нового плейлиста  $p$  составим строку признаковов описания  $\mathbf{x} \in \mathbb{R}^{1 \times (k+d+r)}$ , из системы  $\mathbf{x} = \mathbf{u}\mathbf{H}$ , найдем профиль плейлиста  $\mathbf{u}$ , методом наименьших квадратов. Теперь умножая вектор  $\mathbf{u}$  на матрицу  $\mathbf{V}$ , получим вектор  $\mathbf{r}$ ,  $i$ -ый элемент, которого означает насколько трек  $t_i$  подходит плейлисту  $p$ , после этого из вектора  $\mathbf{r}$  выбирается top- $N$  значений с индексами  $\{i_1, \dots, i_N\}$  и плейлисту  $p$  рекомендуются треки  $\{t_{i_1}, \dots, t_{i_N}\}$ .

## 4 Оценка качества алгоритма

Качество полученных рекомендаций будет оцениваться с помощью двух метрик: R-precision и  $nDCG$  - normalized discounted cumulative gain. Первая из них будет показывать насколько, рекомендованные треки соответствуют интересам пользователя, вторая помимо релевантности треков, показывает качество ранжирования треков. Пусть  $R$  список рекомендованных треков,  $G$  - список истинных треков.

R-precision определяется, как отношение правильно рекомендованных треков на количество истинных треков.

$$\text{R-precision} = \frac{|G \cap R_{1:|G|}|}{|G|}$$

Чтобы определить  $nDCG$ , определим сначала  $DCG$  и  $IDCG$  - идеальный  $DCG$ ,  $DCG$  увеличивается, если релевантно рекомендованные треки находятся выше в списке  $top - N$  рекомендаций.

$$DCG = 1 + \sum_{i=2}^{|R|} \frac{rel_i}{\log_2 i}$$

$$IDCG = 1 + \sum_{i=2}^{|R|} \frac{1}{\log_2 i}$$

Теперь определим  $nDCG$ :

$$nDCG = \frac{DCG}{IDCG}$$

## 5 Базовый вычислительный эксперимент

Оценка качества алгоритма и сравнение с другими методами будет проводиться на случайной подвыборке из миллиона плейлистов, содержащей 7657 плейлистов и 8560 треков, при этом каждый плейлист содержит не менее пяти треков. По данной подвыборке проводится следующая кроссвалидация: множество плейлистов разбивается на пять примерно равных частей, качество оценивается на каждой из пяти частей поочередно, на остальных обучается, при этом на тестовых плейлистах скрывается часть треков.

## Литература

- [1] Geoffray Bonnin, Dietmar Jannach. Automated Generation of Music Playlists: Survey and Experiments. ACM Computing Surveys (CSUR). 2014
- [2] Scott C. Deerwester, Susan T. Dumais, Thomas K. Landauer, George W. Furnas, Richard A. Harshman. Indexing by Latent Semantic Analysis. JASIS. 1990
- [3] Martin Saveski, Amin Mantrach. Item Cold-Start Recommendations: Learning Local Collective Embeddings. Proceeding RecSys '14 Proceedings of the 8th ACM Conference on Recommender systems Pages 89-96. 2014
- [4] Yifan Chen, Xiang Zhao. Leveraging High-Dimensional Side Information for Top-N Recommendation. CoRR. 2017
- [5] Cold-Start Item and User Recommendation with Decoupled Completion and Transduction Iman Barjasteh, Rana Forsati, Farzan Masrour, Abdol-Hossein Esfahanian, Hayder Radha. Cold-Start Item and User Recommendation with Decoupled Completion and Transduction. 2015