УДК 004.89
DOI 10.25205/1818-7900-2019-17-3-43-60

Разработка музыкальной рекомендательной системы на основе обработки метаданных контента

А. В. Менькин

Новосибирский государственный университет
Новосибирск, Россия

Аннотация

Музыкальная рекомендательная система (МРС) помогает пользователям музыкальных стриминговых сервисов находить интересующий их музыкальный контент. Разреженность пользовательских оценок – одна из главных проблем исследования МРС. Она вызвана тем, что пользователь оценивает лишь малую долю объектов музыкального каталога. В результате МРС часто не обладает достаточным набором данных для составления рекомендаций.

В статье предложен подход для решения проблемы разреженности пользовательских оценок на основе использования оценок связанных объектов. Описана гибридная МРС, использующая как нормализованные пользовательские оценки треков, альбомов, артистов, жанров, так и информацию о связях между объектами разных типов. Произведена оценка эффективности разработанной МРС, а также произведен сравнительный анализ предложенного подхода с коллаборативным методом предсказания пользовательских предпочтений.

Ключевые слова

музыкальные рекомендательные системы, разреженность пользовательских оценок, метаданные, музыкальный контекст

Для цитирования

Менькин А. В. Разработка музыкальной рекомендательной системы на основе обработки метаданных контента // Вестник НГУ. Серия: Информационные технологии. 2019. Т. 17, № 3. С. 43–60. DOI 10.25205/1818-7900-2019-17-3-43-60

Development of a Music Recommender System Based on Content Metadata Processing

A. V. Menkin

Novosibirsk State University
Novosibirsk, Russian Federation

Abstract

Music recommender systems (MRS) help users of music streaming services to find interesting music in the music catalogs. The sparsity problem is an essential problem of MRS research. It refers to the fact that user usually rates only a tiny part of items. As a result, MRS often has not enough data to make a recommendation.

To solve the sparsity problem, in this paper, a new approach that uses related items' ratings is proposed. Hybrid MRS based on this approach is described. It uses tracks, albums, artists, genres normalized ratings along with information about relations between items of different types in the music catalog. The proposed MRS is evaluated and compared to collaborative method for users' preferences prediction.

Keywords

music recommender systems, sparsity, metadata, music context

For citation

Menkin A. V. Development of a Music Recommender System Based on Content Metadata Processing. *Vestnik NSU. Series: Information Technologies*, 2019, vol. 17, no. 3, p. 43–60. (in Russ.) DOI 10.25205/1818-7900-2019-17-3-43-60

Введение

Рекомендательные системы – это программные инструменты и методики, которые предлагают пользователям объекты, наиболее интересные для них [1]. Музыкальная рекомендательная система (МРС) предлагает пользователям музыкального сервиса объекты музыкального каталога. Основными задачами и направлениями исследований МРС являются:

1) предсказание пользовательских оценок (треков, альбомов, артистов, жанров и др.);
2) автоматическое тегирование;
3) автоматическое составление (продолжение) плейлиста;
4) контекстно-ориентированные системы.

Задача предсказания пользовательских оценок решается для составления рекомендаций [2–5] и для восполнения пропусков в разреженной матрице пользовательских оценок [6]. Семантические дескрипторы (теги) позволяют описать объекты музыкального каталога в терминах, понятных для пользователя [7; 8], поэтому они удобны для поиска музыки. Этим вызван интерес к задачам автоматического тегирования, например, классификации жанров и распознавания эмоций [9]. Так как пользователь обычно слушает треки в составе последовательности [10], задача автоматического составления (продолжения) плейлиста [8; 11] также представляет интерес. Она связана с выбором треков, соответствующих предопределенным характеристикам (например, семантическим дескрипторам [8]), порядком выбранных треков и др. [10]. Контекст пользователя (локация, время, деятельность и др.) значительно влияет на его ситуативные предпочтения [10], поэтому разработка контекстно-ориентированных систем [8] является перспективным направлением.

Для решения задач исследований МРС используются три вида данных:

1) контентные данные;
2) музыкальный контекст;
3) пользовательские данные.

Контентные данные [12] – это значения акустических признаков, полученные в результате обработки аудиосигнала [2; 3; 8; 13]. Музыкальный контекст [14] – это данные, которые описывают музыкальные объекты, но не являются результатом обработки аудиосигнала. Например, текстовые описания объектов, информация о связях между объектами разных типов, теги, поставленные экспертной группой, и др. Пользовательские данные [15] включают в себя персональные данные [3; 5], историю воспроизведений [4; 11; 16], историю сессий (повторные воспроизведения треков, пропуски треков и др.) [4; 6], контекст пользователя [6; 8], поставленные оценки [2; 3; 5; 6], присвоенные теги [11; 13] и др. Также используются данные о взаимодействии пользователей с другими видами контента [5].

При решении задач исследований МРС возникают следующие проблемы:

1) холодный старт (новый пользователь, новый объект);
2) разреженность пользовательских оценок.

Проблема холодного старта [10] является обобщением проблемы нового пользователя и проблемы нового объекта. Проблема нового пользователя возникает, когда пользователь регистрируется в музыкальном сервисе, и МРС не обладает данными об этом пользователе, чтобы составить для него рекомендацию. Для решения этой проблемы применяются опрос пользователя при регистрации и кросс-доменный подход [5]. Проблема нового объекта возникает, когда объект добавляется в каталог, и МРС не обладает данными об этом объекте, чтобы включить его в рекомендацию. Для решения этой проблемы применяются контентно-ориентированный подход [2] и контентные гибридные системы [13; 16].

Проблема разреженности пользовательских оценок [10] возникает, когда число поставленных оценок значительно меньше числа возможных. Разреженность определяется согласно формуле

$$sparsity = \left(1 - \frac{N_R}{N_U \times N_I}\right) \times 100\%, \quad (1)$$

где N_U – число пользователей, N_I – число объектов, N_R – число пользовательских оценок (пользователь может поставить только одну оценку объекту).

Эта проблема особенно актуальна для исследований МРС. Из-за того, что музыкальные каталоги содержат огромное число объектов (десятки миллионов треков [10]), пользователь обычно взаимодействует лишь с малой их долей. Также пользователь обычно слушает треки в составе последовательности и не прерывает воспроизведение, чтобы оценить их. Более того, пользователь нередко не концентрируется на воспроизводимых треках и не оценивает их. В результате доля объектов, оцененных пользователем, обычно близка к нулю. Значение разреженности одного из наибольших наборов данных "C15 – Yahoo! Music user ratings of musical tracks, albums, artists and genres, v 1.0" составляет 99,96 % (набор данных содержит пользовательские оценки треков, альбомов, артистов, жанров). Для сравнения: разреженность набора данных от Netflix составляет 98,82 % (набор данных содержит пользовательские оценки фильмов и сериалов) [10]. Разница более чем в процент существенна.

Для решения этой проблемы применяются контентные гибридные системы и методы, в которых вычисляются неявные пользовательские оценки. В контентных гибридных системах [2] обычно для предсказания пользовательской оценки трека используются оценки треков похожих по значениям акустических признаков. Также может быть обучена модель [3], которая использует значения акустических признаков совместно с персональными данными. В методах, основанных на неявных оценках [4; 6] для треков, с которыми пользователь взаимодействовал, но которые не оценивал, вычисляются неявные оценки, которые восполняют пропуски в разреженной матрице пользовательских оценок.

Для решения проблемы разреженности пользовательских оценок предлагается использовать оценки связанных (согласно метаданным контента) объектов музыкального каталога как похожих. В статье описана гибридная МРС, в которой применен предложенный подход. Этапы работы МРС:

1) определение похожих альбомов, артистов, жанров;
2) предварительный выбор треков;
3) определение похожих пользователей;
4) вычисление значений признаков;
5) обучение модели для предсказания пользовательских оценок треков;
6) предсказание пользовательских оценок треков;
7) составление рекомендации.

На этапах 1–2 для пользователя предварительно выбираются треки. Для этого определяются альбомы, артисты, жанры, которые (предположительно) нравятся ему, и выбираются треки, связанные с ними. На этапах 3–6 для каждого выбранного трека предсказывается оценка. Для этого используются оценки, поставленные похожими пользователями как этому треку, так и связанным с ним объектам. На этапе 7 выбираются треки с наибольшими предсказанными оценками.

Музыкальная рекомендательная система

Объекты представлены в виде целочисленных идентификаторов. Множество объектов I определено согласно формуле

$$I = \{0 \dots N_I - 1\},$$

где N_I – число объектов.

Метаданные контента содержат информацию о типах объектов и о связях между объектами разных типов. В результате обработки метаданных контента определены:

1) I_{tracks}, I_{albums}, $I_{artists}$, I_{genres} – разбиение множества I на множества треков, альбомов, артистов, жанров;

2) $c_{track,album}: I_{tracks} \to I_{albums} \cup \{None\}$ – отображение, которое сопоставляет с треком связанный альбом. Если трек не связан ни с одним альбомом, то сопоставляется $None$;

3) $c_{track,artist}: I_{tracks} \to I_{artists} \cup \{None\}$ – отображение, которое сопоставляет с треком связанного артиста. Если трек не связан ни с одним артистом, то сопоставляется $None$;

4) $c_{track,genres}: I_{tracks} \to \mathcal{P}(I_{genres})$ – отображение, которое сопоставляет с треком множество связанных жанров;

5) $c_{album,artist}: I_{albums} \to I_{artists} \cup \{None\}$ – отображение, которое сопоставляет с альбомом связанного артиста. Если альбом не связан ни с одним артистом, то сопоставляется $None$;

6) $c_{album,genres}: I_{albums} \to \mathcal{P}(I_{genres})$ – отображение, которое сопоставляет с альбомом множество связанных жанров.

Также определены вспомогательные отображения:

1) $c_{album,tracks}: I_{albums} \to \mathcal{P}(I_{tracks})$ – отображение, которое сопоставляет с альбомом множество связанных треков;

2) $c_{artist,tracks}: I_{artists} \to \mathcal{P}(I_{tracks})$ – отображение, которое сопоставляет с артистом множество связанных треков;

3) $c_{artist,albums}: I_{artists} \to \mathcal{P}(I_{albums})$ – отображение, которое сопоставляет с артистом множество связанных альбомов;

4) $c_{genre,tracks}: I_{genres} \to \mathcal{P}(I_{tracks})$ – отображение, которое сопоставляет с жанром множество связанных треков;

5) $c_{genre,albums}: I_{genres} \to \mathcal{P}(I_{albums})$ – отображение, которое сопоставляет с жанром множество связанных альбомов.

Пользователи представлены в виде целочисленных идентификаторов. Множество пользователей U определено согласно формуле

$$U = \{0 \dots N_U - 1\},$$

где N_U – число пользователей.

МРС использует пользовательские оценки объектов:

1) $r: U \times I \to [0,1] \cup \{None\}$ – отображение, которое сопоставляет с пользователем и объектом нормализованную оценку. Если пользователь не оценивал объект, то сопоставляется $None$;

2) $r_2: U \times I \to \{0, 1, None\}$ – отображение, которое сопоставляет с пользователем и объектом бинарную оценку. Определено согласно формуле

$$r_2(u,i) = \begin{cases} 0, & r(u,i) \in [0, t_r) \\ 1, & r(u,i) \in [t_r, 1] \\ None, & r(u,i) = None \end{cases}, \quad t_r \in (0,1],$$

где t_r – пороговое значение. Если $r(u,i)$ достигает t_r, то считается, что пользователю u нравится объект i. Если пользователь не оценивал объект, то сопоставляется $None$.

Определение похожих альбомов, артистов, жанров

Рассмотрим на примере альбомов. В первую очередь по формуле

$$I'_{albums} = \{i \in I_{albums} | c_{album,tracks}(i) \neq \emptyset\}$$

определяется множество альбомов, связанных хотя бы с одним треком, I'_{albums}.

Для каждого альбома $i \in I'_{albums}$ согласно формуле

$$v_i(u) = \begin{cases} -1, & r_2(u,i) = 0 \\ 1, & r_2(u,i) = 1 \\ 0, & r_2(u,i) = None \end{cases}, \quad u \in U,$$

строится вектор v_i.

Компонента $v_i(u)$ соответствует пользователю u и содержит значение из множества $\{-1, 1, 0\}$ в зависимости от того, оценивал ли пользователь альбом, и если оценивал, то понравился ли альбом ему. Далее, для каждой пары альбомов $i, i' \in I'_{albums}$ $(i \neq i')$ на основе векторов v_i и $v_{i'}$ вычисляется значение коэффициента схожести sim_I:

$$sim_I(i, i') = \begin{cases} \cos(v_i, v_{i'}), & \text{если } v_i, v_{i'} \neq \bar{0} \\ 0, & \text{иначе} \end{cases}.$$

Чем ближе значение $sim_I(i, i') > 0$ к 1, тем больше альбомы i, i' похожи. Чем ближе значение $sim_I(i, i') < 0$ к -1, тем более они различны. Если значение $sim_I(i, i')$ близко к 0, то схожесть не определена. Если хотя бы один альбом не оценен ни одним пользователем, то считается, что $sim_I(i, i') = 0$. Далее, для каждого альбома $i \in I'_{albums}$ на основе значений sim_I определяется множество похожих альбомов:

$$Sim_I(i) = \{i' \in I'_{albums} \backslash \{i\} \mid sim_I(i, i') \geq t_{sim_I}\}, \qquad t_{sim_I} \in (0, 1],$$

где t_{sim_I} – пороговое значение.

Для артистов, жанров, связанных хотя бы с одним треком, – аналогично.

Также при построении вектора v_i для альбома i, если пользователь u не оценивал альбом, но оценивал хотя бы один связанный трек, предлагается вычислять среднее значение нормализованных оценок пользователя, поставленных связанным трекам, и в зависимости от того, достигает ли это значение t_r, восполнять компоненту $v_i(u) = 0$ значением из множества $\{-1, 1\}$. В результате восполнения нулевых компонент векторов v строятся векторы v', которые могут быть использованы вместо векторов v при вычислении значений sim_I. Аналогично для артистов, жанров (используются оценки связанных треков и альбомов).

Предварительный выбор треков

Для пользователя $u \in U$ выбираются треки, которые в дальнейшем могут быть включены в рекомендацию. В первую очередь согласно формулам

$$I^1_{albums}(u) = \{i \in I'_{albums} | r_2(u, i) = 1\},$$
$$I^0_{albums}(u) = \{i \in I'_{albums} | r_2(u, i) = 0\},$$
$$I^{None}_{albums}(u) = \{i \in I'_{albums} | r_2(u, i) = None\}$$

определяются множества альбомов, связанных хотя бы с одним треком, которые нравятся пользователю, не нравятся и которые он не оценивал ($I^1_{albums}(u)$, $I^0_{albums}(u)$, $I^{None}_{albums}(u)$ соответственно).

Для каждого альбома $i \in I^{None}_{albums}(u)$ определяется число похожих альбомов, которые нравятся пользователю и не нравятся, $c^1(u, i)$, $c^0(u, i)$ соответственно:

$$c^1(u, i) = \left\| I^1_{albums}(u) \cap Sim_I(i) \right\|,$$
$$c^0(u, i) = \left\| I^0_{albums}(u) \cap Sim_I(i) \right\|.$$

На основе этих чисел предсказывается бинарная оценка:

$$\widehat{r_2}(u, i) = \begin{cases} 1, & c^1(u, i) \geq c^0(u, i), c^1(u, i) > 0 \\ 0, & c^1(u, i) < c^0(u, i) \\ None, & c^1(u, i) = c^0(u, i) = 0 \end{cases}.$$

Если оба эти числа равны нулю, то оценка не предсказывается.

Для артистов, жанров, связанных хотя бы с одним треком, – аналогично. Для пользователя случайно выбираются $N_{preselection} > 0$ треков, связанных с альбомами, артистами, жанрами, которые (предположительно) нравятся ему (согласно r_2 или $\widehat{r_2}$).

Определение похожих пользователей

В первую очередь для каждого пользователя $u \in U$ строится вектор v_u:

$$v_u(i) = \begin{cases} -1, & r_2(u,i) = 0 \\ 1, & r_2(u,i) = 1 \\ 0, & r_2(u,i) = None \end{cases}, \; i \in I.$$

Компонента $v_u(i)$ соответствует объекту i и содержит значение из множества $\{-1, 1, 0\}$ в зависимости от того, оценивал ли пользователь объект, и если оценивал, то понравился ли объект ему. Далее, для каждой пары пользователей $u, u' \in U$ $(u \neq u')$ на основе векторов v_u и $v_{u'}$ вычисляется значение коэффициента схожести sim_U:

$$sim_U(u,u') = \begin{cases} \cos(v_u, v_{u'}), & \text{если } v_u, v_{u'} \neq \bar{0} \\ 0, & \text{иначе} \end{cases}.$$

Чем ближе значение $sim_U(u,u') > 0$ к 1, тем больше пользователи u, u' похожи. Чем ближе значение $sim_U(u,u') < 0$ к -1, тем больше они различны. Если значение $sim_U(u,u')$ близко к 0, то схожесть не определена. Если хотя бы один пользователь не оценивал ни один объект, то считается, что $sim_U(u,u') = 0$. Далее, для каждого пользователя $u \in U$ на основе значений sim_U определяется множество похожих пользователей $Sim_U(u)$. Для этого выбираются $N_{Sim_U} > 0$ пользователей с наибольшими положительными значениями sim_U.

Вычисление значений признаков

Для пары пользователь $u \in U$, трек $i \in I_{tracks}$ вычисляются значения признаков $f_{k=0\ldots6}(u,i) \in [0,1] \cup \{None\}$. В качестве признаков $f_{k=0\ldots6}(u,i)$ выбраны средние взвешенные значения нормализованных оценок, поставленных (похожими) пользователями из множества $Sim_U(u)$ объектам из множеств $I_{k=0\ldots6}(i)$ (табл. 1).

Таблица 1

Признаки и соответствующие множества объектов

Table 1

Features and Corresponding Sets of Items

Признак	Множество	Объекты множества
$f_0(u,i)$	$I_0(i)$	Трек i
$f_1(u,i)$	$I_1(i)$	Альбом, связанный с треком i
$f_2(u,i)$	$I_2(i)$	Треки, связанные с альбомом
$f_3(u,i)$	$I_3(i)$	Артист, связанный с треком i
$f_4(u,i)$	$I_4(i)$	Треки и альбомы, связанные с артистом
$f_5(u,i)$	$I_5(i)$	Жанры, связанные с треком i
$f_6(u,i)$	$I_6(i)$	Треки и альбомы, связанные с жанрами

Идея заключается в том, что если похожие пользователи не оценивали трек i (это вероятно из-за проблемы разреженности пользовательских оценок), то, возможно, они оценивали связанные альбом, артиста, жанры. В этом случае значение $f_0(u,i)$ не определено, но хотя бы

одно из значений $f_1(u,i)$, $f_3(u,i)$, $f_5(u,i)$ можно вычислить. Если похожие пользователи не оценивали и связанный альбом, то, возможно, они оценивали другие треки, связанные с ним. В этом случае значение $f_1(u,i)$ не определено, но значение $f_2(u,i)$ можно вычислить. Аналогично для связанных артиста, жанров. Более того, когда значение f_0 можно вычислить, использование значений $f_{k=1\ldots6}$ может способствовать увеличению эффективности предсказания оценок треков.

Множества объектов $I_{k=0\ldots6}(i)$ (см. табл. 1) определены согласно формулам

$$I_0(i) = \{i\},$$

$$I_1(i) = \begin{cases} \{c_{track,album}(i)\}, & c_{track,album}(i) \neq None \\ \emptyset, & иначе \end{cases},$$

$$I_2(i) = \bigcup_{i' \in I_1(i)} c_{album,tracks}(i'),$$

$$I_3(i) = \begin{cases} \{c_{track,artist}(i)\}, & c_{track,artist}(i) \neq None \\ \emptyset, & иначе \end{cases},$$

$$I_4(i) = \bigcup_{i' \in I_3(i)} \left(c_{artist,tracks}(i') \cup c_{artist,albums}(i')\right),$$

$$I_5(i) = c_{track,genres}(i),$$

$$I_6(i) = \bigcup_{i' \in I_5(i)} \left(c_{genre,tracks}(i') \cup c_{genre,albums}(i')\right),$$

Значения признаков $f_{k=0\ldots6}(u,i)$ вычисляются по формулам

$$R_{k=0\ldots6}(u,i) = \{(u',i',r') | u' \in Sim_U(u), i' \in I_k(i), r' = r(u',i') \neq None\},$$

$$f_{k=0\ldots6}(u,i) = \frac{\sum_{(u',i',r') \in R_k(u,i)} (sim(u,u') \times r')}{\sum_{(u',i',r') \in R_k(u,i)} sim(u,u')}, \; R_k(u,i) \neq \emptyset.$$

Если $R_k(u,i) = \emptyset$, то значение $f_k(u,i)$ не определено. В этом случае считается, что $f_k(u,i) = None$.

Обучение модели для предсказания пользовательских оценок треков

На этапах определения похожих пользователей и вычисления значений признаков подготавливаются тренировочное, валидационное (для выбора модели) и тестовое множества. Эти множества содержат кортежи: пользователь $u \in U$, трек $i \in I_{tracks}$, значения признаков $f_{k=0\ldots6}(u,i)$, оценка $r(u,i)$ (или $r_2(u,i)$). Важно, чтобы при подготовке этих множеств используемые оценки не пересекались. С использованием тренировочного множества обучается модель (модели) для предсказания оценок треков $\hat{r}$ (или $\widehat{r_2}$). В качестве входных данных модель использует предварительно обработанные значения $f_{k=0\ldots6}$. Обученная модель предсказывает оценки для предварительно выбранных треков.

Составление рекомендации

Для пользователя $u \in U$ составляется список треков, которые, предположительно, нравятся ему. Для этого из предварительно выбранных треков выбираются $N_{recommendation} > 0$ треков с наибольшими предсказанными нормализованными оценкам, достигающими t_r (в порядке убывания предсказанной нормализованной оценки). Если предсказываются бинарные оценки, то трек i включается в рекомендацию, если предсказанная бинарная оценка $\widehat{r_2}(u,i) = 1$.

Реализация МРС

Для реализации МРС выбран набор данных “C15 – Yahoo! Music user ratings of musical tracks, albums, artists and genres, v 1.0” (тренировочное подмножество “большого” поднабора данных, табл. 2). Пользователи и объекты (треки, альбомы, артисты, жанры) представлены в виде целочисленных идентификаторов. Для треков опционально определены связанные альбом, артист, жанры; для альбомов – связанные артист, жанры. Пользовательские оценки представлены в виде целочисленных значений из множества $\{0 \dots 100\}$, хотя 97,69 % оценок кратны 10.

Таблица 2

Основные характеристики используемой части набора данных

Table 2

Characteristics of the Dataset’s Applied Part

Показатель	Значение
Число пользователей	1000990
Число объектов	624961
Число треков	507172
Число альбомов	88909
Число артистов	27888
Число жанров	992
Число пользовательских оценок	252800275
Разреженность (согласно формуле (1))	99,96 %

Оценки (при помощи деления на 100) приведены к значениям из отрезка $[0, 1]$. Выбрано пороговое значение нормализованной оценки $t_r = 0{,}5$. Это наибольшее значение, которое достигается как минимум половиной нормализованных оценок (58,67 %, при $t_r = 0{,}51$ – 47,44 %).

Определение похожих альбомов, артистов, жанров и предварительный выбор треков

Недостатком набора данных является то, что не все альбомы, артисты, жанры связаны с треками. Поэтому на этапах определения похожих альбомов, артистов, жанров и предварительного выбора треков рассматривались 52 187 альбомов (58,7 % от числа всех альбомов), 19 691 артист (70,61 %), 205 жанров (20,67 %). На основе оценок пользователей из множества $\{0 \dots 19999\}$ определены множества похожих альбомов, артистов, жанров. Для этого использовались значения sim_I, вычисленные на основе векторов v, v', и пороговые значения t_{sim_I}: 0.01, 0.05, 0.1, 0.15, 0.2, 0.25 (12 подходов, для каждого подхода множества определены по отдельности). Для альбомов выбирается подход, при котором достигается наибольшая точность предсказания бинарных оценок. Для артистов, жанров – аналогично. Для пользователя с применением выбранных подходов предварительно выбираются $N_{preselection} = 40$ треков.

Определение похожих пользователей

Для каждого пользователя из множества $\{0 \dots 19999\}$ оценки треков разделены на тренировочные, валидационные и тестовые в долях 75, 12,5, 12,5 % соответственно. Все оценки альбомов, артистов, жанров использовались как тренировочные. Для построения векторов v

и вычисления значений sim_U использовались только тренировочные оценки. Для пользователей из множества $\{0 \dots 199\}$ хотя бы с одной оценкой трека (123 пользователя) вычислены значения sim_U относительно пользователей из множества $\{0 \dots 19999\}$. Для определения множеств похожих пользователей Sim_U выбирались $N_{Sim_U} = 1000$ пользователей с наибольшими положительными значениями sim_U. В результате для 119 из 123 пользователей определена 1 000 похожих пользователей (для остальных 4 пользователей – меньшее число). В среднем по 123 множествам Sim_U среднее значение sim_U составило 0,1481, а среднее квадратичное отклонение значений sim_U внутри множества – 0,03.

Вычисление значений признаков

Для оценок треков, поставленных пользователями из множества $\{0 \dots 199\}$, вычислены значения признаков $f_{k=0\dots6}$. В результате подготовлены тренировочное, валидационное и тестовое множества с мощностями 16769, 2797, 2787 соответственно (табл. 3).

Таблица 3

Доли пропусков значений признаков, %

Table 3

Features' Missing Parts, %

Множество / Признак	f_0	f_1	f_2	f_3	f_4	f_5	f_6
Тренировочное	32,55	20,6	17,87	13,71	13,36	4,94	2,17
Валидационное	33,25	20,38	17,73	14,48	13,55	4,83	2,32
Тестовое	32,15	19,27	17,19	13,17	12,67	5,06	1,94

В тестовом множестве значение признака f_0 (среднее взвешенное значение нормализованных оценок трека, поставленных похожими пользователями) не определено в 32,15 % случаев. Доли пропусков значений признаков $f_{k=1\dots6}$ значительно меньше. Отметим, что в используемой части набора данных 8,33 % треков не связаны ни с одним альбом, 12,52 % – ни с одним артистом, 6,15 % – ни с одним жанром (в коммерческих музыкальных каталогах эта проблема отсутствует).

Модель для предсказания пользовательских оценок треков и составление рекомендации

Для реализации модели выбран метод машинного обучения «случайный лес» [17]. Метод применен со значениями параметров «число деревьев»: 10, 20, 30, 40, 50, 60, 70; «минимальное число элементов в листе»: 1, 5, 10, 15, 20 (35 вариантов). Пропуски в значениях признаков $f_{k=0\dots6}$ заменяются средним значением признака в тренировочном множестве (для признаков $f_{k=0\dots6}$ эти значения равны 0,5747, 0,5593, 0,547, 0,6021, 0,5579, 0,6243, 0,573 соответственно). С использованием тренировочного множества обучены модели для предсказания нормализованных и бинарных пользовательских оценок треков. С использованием валидационного множества выбираются модели с наибольшей эффективностью предсказания оценок. С использованием выбранной (для нормализованных или бинарных оценок) модели выбираются $N_{recommendation} = 10$ треков для рекомендации.

Оценка эффективности реализованной МРС

Предварительный выбор треков

Для оценки эффективности предсказания бинарных оценок альбомов, артистов, жанров использовались оценки пользователей из множества {20000 ... 29999}.

На основе оценок пользователей из множества {0 ... 29999} выявлено, что в **90,55 %** случаев, когда пользователь оценивал как альбом, так и хотя бы один связанный трек, бинарная оценка альбома совпадает с приведенным к бинарному значению (с использованием порогового значения t_r) средним значением нормализованных оценок связанных треков. Для артистов, жанров совпадения выявлены в **87,09** и **65,89 %** случаев соответственно. Это означает, что если пользователю нравится альбом, артист или жанр, то с большой вероятностью (с меньшей для жанра) ему нравятся связанные треки. Таким образом, при высокой точности предсказания бинарных оценок альбомов, артистов, жанров предварительный выбор треков является эффективным.

Для каждого пользователя из множества {20000 ... 29999} с двумя и более оценками альбомов (4 652 пользователя) оценки альбомов случайно разделены пополам на оценки, которые предсказываются и которые используются для предсказания. Всего предсказывалось 223 639 оценок, значение 0 принимают 46,3 %, значение 1 – 53,7 %. Для каждого подхода получены точность по формуле

$$ACC = \frac{1}{\|T\|}\sum_{T}[\widehat{r_2} = r_2] \times 100\% \qquad (2)$$

и доля не предсказанных оценок (табл. 4).

Таблица 4

Результаты предсказания бинарных оценок альбомов
(все подходы)

Table 4

Albums Binary Ratings Prediction
(All Approaches)

t_{sim_I}	ACC, %		Доля $\widehat{r_2} = None$, %	
	v	v'	v	v'
0,01	78,51	78,62	0,85	0,47
0,05	79,57	**79,99**	1,61	1,33
0,1	76,96	76,61	6,47	7,25
0,15	66,1	63,39	20,55	24,16
0,2	49,49	44,15	41,32	48,13
0,25	34,91	28,17	59,27	67,47

Выбран подход с использованием векторов v' и значения $t_{sim_I} = 0{,}05$. По сравнению с аналогичным подходом на основе векторов v доля альбомов с хотя бы одним значением sim_I, достигающим t_{sim_I}, увеличилась с 93 до 95,86 % (для этих альбомов потенциально может быть предсказана бинарная оценка). Для выбранного подхода получены доли, в которых для бинарных оценок со значениями 0 и 1 предсказаны значения 0, 1 и $None$ (табл. 5).

Таблица 5

Результаты предсказания
бинарных оценок альбомов со значениями 0 и 1
(выбранный подход)

Table 5

Albums Binary Ratings Equal to 0 and 1 Prediction
(Selected Approach)

$r_2 \setminus \widehat{r_2}$	0, %	1, %	$None$, %
0	76,55	22,33	1,13
1	15,54	82,96	1,5

Для артистов – аналогично (табл. 6). Всего предсказывалось 359 378 оценок (9 831 пользователь), значение 0 принимают 39,39 %, значение 1 – 60,61 %.

Таблица 6

Результаты предсказания бинарных оценок артистов
(все подходы)

Table 6

Artists Binary Ratings Prediction
(All Approaches)

t_{sim_I}	ACC, %		Доля $\widehat{r_2} = None$, %	
	v	v'	v	v'
0,01	81,81	81,79	0,12	0,07
0,05	83,41	83,3	0,21	0,15
0,1	84,06	**84,15**	1,05	0,93
0,15	81,7	81,82	4,49	4,38
0,2	73,72	73,36	14,26	14,69
0,25	59,53	57,81	31,5	33,61

Выбран подход с использованием векторов v' и значения $t_{sim_I} = 0{,}1$. По сравнению с аналогичным подходом на основе векторов v доля артистов с хотя бы одним значением sim_I, достигающим t_{sim_I}, увеличилась с 91,13 до 94,57 %. Для выбранного подхода получены доли, в которых для бинарных оценок со значениями 0 и 1 предсказаны значения 0, 1 и $None$ (табл. 7).

Таблица 7

Результаты предсказания
бинарных оценок артистов со значениями 0 и 1
(выбранный подход)

Table 7

Artists Binary Ratings Equal to 0 and 1 Prediction
(Selected Approach)

$r_2 \setminus \widehat{r_2}$	0, %	1, %	$None$, %
0	75,46	23,42	1,12
1	9,39	89,8	0,81

Для жанров – аналогично (табл. 8). Всего предсказывалось 32 596 оценок (6 498 пользователей), среди которых значение 0 принимают 38,92 %, значение 1 – 61,08 %.

Таблица 8

Результаты предсказания бинарных оценок жанров
(все подходы)

Table 8

Genres Binary Ratings Prediction
(All Approaches)

t_{sim_I}	ACC, %		Доля $\widehat{r_2} = None$, %	
	v	v'	v	v'
0,01	84,35	83,91	0,01	0
0,05	84,95	84,17	0,05	0
0,1	85,89	84,85	0,23	0,05
0,15	**86,39**	86,09	0,85	0,52
0,2	86,09	86,03	1,93	1,94
0,25	84,62	84,9	4,08	3,82

Выбран подход на основе векторов v и значения $t_{sim_I} = 0{,}15$. Доля жанров с хотя бы одним значением sim_I, достигающим t_{sim_I}, составляет 100 %. Для выбранного подхода получены доли, в которых для бинарных оценок со значениями 0 и 1 предсказаны значения 0, 1 и $None$ (табл. 9).

Таблица 9

Результаты предсказания
бинарных оценок жанров со значениями 0 и 1
(выбранный подход)

Table 9

Genres Binary Ratings Equal to 0 and 1 Prediction
(Selected Approach)

$r_2 \setminus \widehat{r_2}$	0, %	1, %	$None$, %
0	86,02	13,53	0,44
1	12,26	86,63	1,11

Предсказание пользовательских оценок треков

На основе оценок пользователей из множества $\{0 \ldots 199\}$ с хотя бы одной оценкой трека (123 пользователя) и похожих пользователей из множества $\{0 \ldots 19999\}$ подготовлены валидационное и тестовое множества с мощностями 2 797, 2 787 соответственно (табл. 10).

Таблица 10

Доли оценок кратных 0,1 в валидационном и тестовом множествах, %

Table 10

Parts of Ratings that Are Multiples of 0.1 in the Validation and the Test Sets, %

Мн. $\setminus r$	0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Валид.	35	0,14	0,21	6,97	0,21	10,87	0,32	11,23	1,22	17,27	9,72
Тест.	33,55	0,43	0,07	8,07	0,14	10,8	0,47	10,44	1,69	16,22	11,12

Метод «случайный лес» применен со значениями параметров «число деревьев»: 10, 20, 30, 40, 50, 60, 70; «минимальное число элементов в листе»: 1, 5, 10, 15, 20. Для обученных моделей с использованием валидационного множества получены значения RMSE (табл. 11):

$$RMSE = \sqrt{\frac{1}{\|T\|}\sum_T (\hat{r} - r)^2}.$$

Таблица 11

RMSE предсказания нормализованных оценок треков из валидационного множества (все модели)

Table 11

Tracks Normalized Ratings Prediction RMSE for the Validation Set (All Models)

Ч. д.	Минимальное число элементов в листе				
	1	5	10	15	20
10	0,3133	0,3071	0,309	0,3107	0,3119
20	0,3065	0,3035	0,3048	0,3095	0,3106
30	0,3043	0,3028	0,3059	0,3083	0,311
40	0,3026	0,3014	0,3049	0,3084	0,3097
50	0,303	0,3018	0,3049	0,3075	0,3103
60	0,3036	0,30101	0,3048	0,3075	0,3099
70	0,3018	**0,30095**	0,3042	0,3079	0,3104

Выбрана модель, для которой метод «случайный лес» применен со значениями параметров «число деревьев» – 70, «минимальное число элементов в листе» – 5.

Также рассмотрен коллаборативный метод, в котором значение признака f_0 (среднее взвешенное значение нормализованных оценок трека, поставленных похожими пользователями) используется в качестве предсказания. В тех случаях, когда значение f_0 не определено (для тестового множества в 32,15 % случаев), используется среднее значение f_0 в тренировочном множестве. Для выбранной модели и коллаборативного метода с использованием тестового множества получены значения RMSE и MAE (табл. 12):

$$MAE = \frac{1}{\|T\|}\sum_T |\hat{r} - r|.$$

Таблица 12

Результаты предсказания нормализованных оценок треков из тестового множества

Table 12

Tracks Normalized Ratings Prediction for the Test Set

	RMSE	MAE
Выбранная модель	0,3037	0,2354
Зн. пр. f_0 или ср. зн.	0,3946	0,3177

Выбранная модель превосходит коллаборативный метод. Также значения RMSE и MAE получены для подмножества тестового множества, для которого значения f_0 определены (табл. 13).

Таблица 13

Результаты предсказания нормализованных оценок треков из подмножества тестового множества, для которого значения f_0 определены

Table 13

Tracks Normalized Ratings Prediction for the Part of the Test Set where f_0 Is Calculated

	RMSE	MAE
Выбранная модель	0,3093	0,2419
Зн. пр. f_0	0,374	0,2808

Даже в тех случаях, когда коллаборативный метод предсказывает нормализованную оценку (значение f_0 определено), выбранная модель предсказывает ее с меньшими значениями ошибок RMSE и MAE.

Для выбранной модели получены значения важности признаков $f_{k=0\ldots6}$ (табл. 14). Для этого для каждого дерева вычисляется среднее взвешенное значение уменьшения неопределенности (англ. impurity) в вершинах, где используется признак. В качестве весов используются доли элементов тренировочного множества, достигающих вершины. Важность признака – это среднее этих значений, вычисленных для каждого дерева. При обучении моделей в качестве меры неопределенности используется среднеквадратичная ошибка (MSE).

Таблица 14

Важность признаков для предсказания нормализованных оценок треков (выбранная модель)

Table 14

Features Importance for the Normalized Ratings Prediction (Selected Model)

f_0	f_1	f_2	f_3	f_4	f_5	f_6
0,0986	0,0655	0,1166	0,1168	**0,3437**	0,1106	**0,1481**

Наиболее важными являются признаки f_4 и f_6 (средние взвешенные значения нормализованных оценок, поставленных похожими пользователями трекам и альбомам артиста; трекам и альбомам жанров соответственно).

Аналогично обучены модели для предсказания бинарных оценок. Среди бинарных оценок из валидационного множества значение 0 принимают 43,15 %, значение 1 – 56,85 %. Среди бинарных оценок из тестового множества значение 0 принимают 43,09 %, значение 1 – 56,91 %. Для обученных моделей с использованием валидационного множества получены значения точности (согласно формуле (2), табл. 15).

Таблица 15

Точность предсказания бинарных оценок треков
из валидационного множества (все модели), %

Table 15

Tracks Binary Ratings Prediction Accuracy for the Validation Set
(All Models), %

Ч. д.	Минимальное число элементов в листе				
	1	5	10	15	20
10	76,87	76,83	76,69	76,94	76,08
20	77,55	77,51	77,08	76,51	75,94
30	77,4	77,51	77,26	77,05	76,22
40	77,51	77,69	77,08	76,62	76,44
50	78,01	77,69	77,08	76,73	76,51
60	77,94	77,65	77,01	76,76	76,44
70	**78,05**	77,87	77,37	77,15	76,51

Выбрана модель, для которой метод «случайный лес» применен со значениями параметров «число деревьев» – 70, «минимальное число элементов в листе» – 1.

Аналогично рассмотрен коллаборативный метод, в котором приведенное к бинарному значение признака f_0 используется в качестве предсказания. В тех случаях, когда значение f_0 не определено (для тестового множества в 32,15 % случаев), используется предопределенное значение бинарной оценки – 0 или 1. Для выбранной модели и коллаборативного метода с использованием тестового множества получены значения точности и TPR, TNR, PPV, NPV (табл. 16):

$$TPR = \frac{\sum_T[(r_2=1)\wedge(\widehat{r_2}=1)]}{\sum_T[r_2=1]},$$

$$TNR = \frac{\sum_T[(r_2=0)\wedge(\widehat{r_2}=0)]}{\sum_T[r_2=0]},$$

$$PPV = \frac{\sum_T[(r_2=1)\wedge(\widehat{r_2}=1)]}{\sum_T[\widehat{r_2}=1]},$$

$$NPV = \frac{\sum_T[(r_2=0)\wedge(\widehat{r_2}=0)]}{\sum_T[\widehat{r_2}=0]}.$$

Таблица 16

Результаты предсказания бинарных оценок треков из тестового множества

Table 16

Tracks Binary Ratings Prediction for the Test Set

	ACC, %	TPR	TNR	PPV	NPV
Выбранная модель	**77,04**	0,8045	0,7252	0,7945	0,7375
Зн. пр. f_0 или 0	64,08	0,8815	0,3231	0,6323	0,6736
Зн. пр. f_0 или 1	65,81	0,6141	0,7161	0,7407	0,5842

Выбранная модель превосходит коллаборативный метод. Также значения точности и TPR, TNR, PPV, NPV получены для подмножества тестового множества, для которого значения f_0 определены (табл. 17).

Таблица 17

Результаты предсказания бинарных оценок треков из подмножества тестового множества, для которого значения f_0 определены

Table 17

Tracks Binary Ratings Prediction for the Part of the Test Set where f_0 Is Calculated

	ACC, %	TPR	TNR	PPV	NPV
Выбранная модель	76,04	0,8124	0,6776	0,8007	0,6938
Зн. пр. f_0	72,03	0,8382	0,5322	0,7407	0,6736

Даже в тех случаях, когда коллаборативный метод предсказывает бинарную оценку, выбранная модель предсказывает ее с большей точностью и значениями TNR, PPV, NPV.

Для выбранной модели аналогично получены значения важности признаков $f_{k=0\ldots6}$ (при обучении моделей используется неопределенность Джини, табл. 18).

Таблица 18

Важность признаков для предсказания бинарных оценок треков (выбранная модель)

Table 18

Features Importance for the Binary Ratings Prediction (Selected Model)

f_0	f_1	f_2	f_3	f_4	f_5	f_6
0,1483	0,0971	**0,1564**	0,1132	**0,196**	0,1258	**0,1632**

Наиболее важными являются признаки f_4, f_6, f_2, f_0 (среднее взвешенное значение нормализованных оценок, поставленных похожими пользователями трекам и альбомам артиста; трекам и альбомам жанров; трекам альбома; треку соответственно).

Заключение

В статье предложен подход для решения проблемы разреженности пользовательских оценок в исследованиях МРС. Он заключается в использовании оценок связанных (согласно метаданным контента) объектов как похожих. На основе предложенного подхода разработана и реализована МРС. Выполнена оценка реализованной МРС.

Выявлено, что если пользователю нравится альбом, артист или жанр, то с большой вероятностью (в меньшей степени для жанра) ему нравятся и треки, связанные с ним. При этом получена точность предсказания бинарных пользовательских оценок альбомов, артистов, жанров: 79,99, 84,15 и 86,39 % соответственно. Это означает, что предварительный выбор треков, реализованный в МРС, является эффективным.

Метод предсказания пользовательских оценок треков, реализованный в МРС, превосходит коллаборативный метод (использующий оценки похожих пользователей, поставленные треку) как на всем тестовом множестве (в тех 32,15 % случаев, когда коллаборативный метод не может предсказать оценку, используется предопределенное значение), так и на его подмножестве, для которого коллаборативный метод может предсказать оценку. При этом наиболее важными являются признаки, значения которых вычислены на основе оценок похожих пользователей, поставленных объектам, связанным с артистом и жанрами трека. Это означает, что использование оценок связанных объектов в реализованной МРС действительно способствует решению проблемы разреженности пользовательских оценок.

Список литературы / References

1. **Ricci F., Rokach L., Shapira B.** Recommender Systems: Introduction and Challenges. In: Recommender Systems Handbook. Springer US, 2015, p. 1–34. DOI 10.1007/978-1-4899-7637-6_1
2. **Su J.-H., Chiu T.-W.** An Item-Based Music Recommender System Using Music Content Similarity. In: Intelligent Information and Database Systems. Springer, Berlin, Heidelberg, 2016, p. 179–190. DOI 10.1007/978-3-662-49390-8_17
3. **Cheng R., Tang B.** A Music Recommendation System Based on Acoustic Features and User Personalities. In: Trends and Applications in Knowledge Discovery and Data Mining. Springer International Publishing, 2016, p. 203–213. DOI 10.1007/978-3-319-42996-0_17
4. **Hanna P.** Considering Durations and Replays to Improve Music Recommender Systems. *EAI Endorsed Transactions on Self-Adaptive Systems*, 2018, vol. 0, no. 0, p. 156379. DOI 10.4108/eai.5-2-2018.156379
5. **Fernández-Tobías I., Braunhofer M., Elahi M., Ricci F., Cantador I.** Alleviating the new user problem in collaborative filtering by exploiting personality information. *User Modeling and User-Adapted Interaction*, 2016, vol. 26, no. 2, p. 221–255. DOI 10.1007/s11257-016-9172-z
6. **Kordumova S., Kostadinovska I., Barbieri M., Pronk V., Korst J.** Personalized Implicit Learning in a Music Recommender System. In: User Modeling, Adaptation, and Personalization. Springer, Berlin, Heidelberg, 2010, p. 351–362. DOI 10.1007/978-3-642-13470-8_32
7. **Schedl M., Knees P., McFee B., Bogdanov D., Kaminskas M.** Music Recommender Systems. In: Recommender Systems Handbook. Springer, US, 2015, p. 453–492. DOI 10.1007/978-1-4899-7637-6_13
8. **Cheng Z., Shen J.** On Effective Location-Aware Music Recommendation. *ACM Transactions on Information Systems*, 2016, vol. 34, no. 2, p. 1–32. DOI 10.1145/2846092
9. **Knees P., Schedl M.** Semantic Labeling of Music. In: Music Similarity and Retrieval: An Introduction to Audio- and Web-based Strategies. Springer, Berlin, Heidelberg, 2016, p. 85–104. DOI 10.1007/978-3-662-49722-7_4
10. **Schedl M., Zamani H., Chen C.-W., Deldjoo Y., Elahi M.** Current challenges and visions in music recommender systems research. *International Journal of Multimedia Information Retrieval*, 2018, vol. 7, no. 2, p. 95–116. DOI 10.1007/s13735-018-0154-2
11. **Vall A., Dorfer M., Schedl M., Widmer G.** A Hybrid Approach to Music Playlist Continuation Based on Playlist-song Membership. In: Proc. of the 33rd Annual ACM Symposium on Applied Computing – SAC '18. ACM, 2018, p. 1374–1382. DOI 10.1145/3167132.3167280
12. **Knees P., Schedl M.** Basic Methods of Audio Signal Processing. In: Music Similarity and Retrieval: An Introduction to Audio- and Web-based Strategies. Springer, Berlin, Heidelberg, 2016, p. 33–50. DOI 10.1007/978-3-662-49722-7_2
13. **Domingues M. A., Gouyon F., Jorge A. M., Leal J., Vinagre J., Lemos L., Sordo M.** Combining usage and content in an online recommendation system for music in the Long Tail. *International Journal of Multimedia Information Retrieval*, 2013, vol. 2, no. 1, p. 3–13. DOI 10.1007/s13735-012-0025-1
14. **Knees P., Schedl M.** Contextual Music Meta-data: Comparison and Sources. In: Music Similarity and Retrieval: An Introduction to Audio- and Web-based Strategies. Springer, Berlin, Heidelberg, 2016, p. 107–132. DOI 10.1007/978-3-662-49722-7_5
15. **Knees P., Schedl M.** Listener-Centered Data Sources and Aspects: Traces of Music Interaction. In: Music Similarity and Retrieval: An Introduction to Audio- and Web-based Strategies. Springer, Berlin, Heidelberg, 2016, p. 161–177. DOI 10.1007/978-3-662-49722-7_7
16. **McFee B., Barrington L., Lanckriet G.** Learning Content Similarity for Music Recommendation. *IEEE Transactions on Audio, Speech, and Language Processing*, 2012, vol. 20, no. 8, p. 2207–2218. DOI 10.1109/TASL.2012.2199109

17. **Breiman L.** Random Forests. *Machine Learning*, 2001, vol. 45, no. 1, p. 5–32. DOI 10.1023/A:1010933404324

Материал поступил в редколлегию
Received
20.05.2019

Сведения об авторе / Information about the Author

Менькин Александр Валерьевич, магистрант факультета информационных технологий Новосибирского государственного университета (ул. Пирогова, 1, Новосибирск, 630090, Россия)

Alexander V. Menkin, Undergraduate Student, Faculty of Information Technologies, Novosibirsk State University (1 Pirogov Str., Novosibirsk, 630090, Russian Federation)

a.menkin@g.nsu.ru
ORCID 0000-0002-6364-962X