Модель рекомендательной системы для интерактивного радиосервиса FMhost*

[1]*Захарчук В. В.*, [1]*Игнатов Д. И.*, [1]*Константинов А. В.*, [2]*Николенко С. И.*
dignatov@hse.ru
[1]НИУ ВШЭ, Москва; [2]Математический институт им. Стеклова и Академический университет, Санкт-Петербург

Представлена модель новой рекомендательной системы для интерактивного радиосервиса FMhost. Новая рекомендательная модель сочетает коллаборативный и основанный на поведении пользователя подходы. Приводятся результаты предварительного анализа данных и описывается методика оценивания качества.

A New Model of Radio Station Recommender System for Online Radiohosting FMHost*

[1]*Zaharchuk V. V.*, [1]*Ignatov D. I.*, [1]*Konstantinov A. V.*, [2]*Nikolenko C. I.*
[1]Moscow, NRU HSE, Russia;[2]Steklov Mathematical Institute and Academic University, St. Petersburg, Russia

We proposed a new radio station recommender model for Russian radio hosting FMHost. It consists of two parts: individual-based and collaborative-based. Both subsystems use tag profiles of users and radio stations. The hybrid recommender model and its prospective quality of service evaluation are described.

Рекомендация музыки — одна из важных тем в области рекомендательных систем; см., например, труды International Society for Music Information Retrieval Conference (ISMIR), Workshop on Music Recommendation and Discovery (WOMRAD), и Recommender Systems Conference (RecSys). И хотя такие сервисы как LastFm, Yahoo!LaunchCast и Pandora хорошо известны, они работают на коммерческой основе и, более того, два последних не вещают для России. Несмотря на большое количество качественных работ по различным аспектам рекомендации музыки, проведено совсем немного исследований по рекомендациям радиостанций для онлайн радиосервисов. Эта работа посвящена российскому онлайн радиохостингу FMhost и его новой гибридной рекомендательной подсистеме.

В настоящее время фокус исследований в области информатики для музыкальной индустрии сместился от задач поиска музыки и навигации [1, 2] в сторону сервисов рекомендаций музыки [3, 4]. Тема не нова (см., например, [5]), однако сейчас она переживает второе рождение балагодаря новым возможностям больших онлайн радиосервисов предоставлять не только миллионы треков для прослушивания, но и радиохостинг. Тегирование пользователями (social tagging) — еще один из важных факторов, который позволяет применять сходство по тегам в рекомедательных системах данной предметной области [6].

Работа первых 3-х авторов проведена в рамках программы фундаментальных исследований НИУ ВШЭ в 2012 году. Работа С.И. Николенко поддержана РФФИ, гранты 12-01-00450-а,РФФИ 11-01-12135-офи-м-2011 и 11-01-00760-а, Российской президентской программой для молодых кандидатов наук, грант МК-6628.2012.1, Российской президентской программой для ведущих научных школ, грант НШ-3229.2012.1. Благодарим Рустама Тагиева, Йонаса Пульманса и Миколу Печенижского.

Многие онлайн сервисы (например, Last.fm или LaunchCast) называют свои потоки радиовещания «радиостанциями», но на самом деле они составляют плейлисты из баз данных треков с помощью рекомендательной системы, нежели чем рекомендуют радиоканал. FMhost, с другой стороны, предоставляет пользователям онлайн радиостанции в традиционном смысле этого слова: реальные диджеи проводят лайвы, радиостанции имеют свой стиль и передают настроение диджея, проводятся соревнования и т. д. Таким образом, задача рекомендаций радиостанций оригинальна, а некоторые из возникающих подзадач сервиса уникальны.

Онлайн сервис FMhost

FMhost — это интерактивная радиосеть, портал, который позволяет пользователям прослушивать и организовывать вещание собственной радиостанции. Существует четыре основных категории пользователей портала: 1) неавторизованный пользователь; 2) слушатель; 3) диджей (DJ); 4) владелец радиостанции. Возможности пользователя в системе варьируются в зависимости от их статуса. Неавторизованные пользователи могу прослушивать любые станции, но не имеют права участвовать в голосованиях и стать диджеем, также воспользоваться рекомендательной системой и рейтингами. Слушатели, в отличие от неавторизованных пользователей, могут голосовать за треки, лайвы и радиостанции. Они могут использовать рекомендательную систему или участвовать в рейтинговании,подписаться на лайвы, радиостанции или диджея. Они также могут провести лайв и стать диджеями.

Существует три типа широковещания: 1) перенаправление потока с другого сервера; 2) трансляция автодиджея; 3) «живое вещание».

FMhost был первым проектом такого рода в России начиная с 2009. Сейчас, воодушевленные успехом FMhost `http://host.fm`, существуют несколько радиовещательных порталов, такие как `http://frodio.com/`, `http://myradio24.com/`, `http://fmhosting.ru/` и т.п. В конце 2011 FMhost был закрыт для серьезной реорганизации, переработки кода и разработки новой архитектуры рекомендательной системы. Предыдущая версия рекомендательной системы испытывала несколько проблем, среди которых несоответствие тегов и наличие собственных треков пользователей без описания в виде тегов. Однако опрос FMhost для примерно 100 респондентов показал, что более чем половине из них нравилась предыдущая версия нашей рекомендательной системы и более чем 80% ответов были положительными или нейтральными.

Модели, алгоритмы и архитектура рекомендательной системы

Наша модель использует три исходные матрицы данных. Первая матрица отслеживает $A = (a_{ut})$ количество посещений пользователем u радиостанций с некоторым тегом t. Каждая радиостанция r вещает аудиотреки с некоторым множеством тегов T_r. Множества всех пользователей, радиостанций и тегов обозначены U, R и T соответственно. Вторая матрица $B = (b_{rt})$ содержит информацию о том, как много треков с тегом t было проиграно радиостанцией r. И третья матрица $C = (c_{ur})$ хранит число посещений пользователем u радиостанции r. Для каждой из этих матриц мы обозначим v^A, v^B, и v^C соответсвующие вектора, содержащие суммы элементов: $v^A = \sum\limits_{t\in T} a_{ut}$, $v^B = \sum\limits_{t\in T} b_{rt}$ и $v^C = \sum\limits_{r\in R} a_{ur}$. Для каждой матрицы A, B, C соответствующая матрица частот посещений обозначена как A_f, B_f и C_f; матрица частот получена нормализацией исходной матрицы посещений с соответсвующим вектором посещений, например, $A_f = (a_{ut} \cdot (v_u^A)^{-1})$. Наша модель не является статической; матрицы A, B и C изменяются после посещения пользователем u радиостанции r с тегом t, т. е. каждое значение a_{ut}, b_{rt} и c_{ur} увеличивается на 1 после посещения.

Модель состоит из трех основных блоков: рекоменадтельная система, основанная на модели индивидуального поведения (Individual-Based Recommender System, (IBRS)), модель коллаборативной рекомендательной системы (Collaborative-Based Recommender System (CBRS)) и конечная рекомендательная система (End Recommender Systems (ERS)), которая агрегирует результаты двух предыдущих.

Каждая модель имеет собственную алгоритмическую реализацию. Т. к. обе предыдущие работы [7, 8] и эта работа неявно используют идеи бикластеризации мы продолжаем именование наших основных алгоритмов с помощью акронима RecBi; на этот раз это семейство RecBi3. Мы назвали алгоритмы для трех предложенных выше моделей RecBi3.1, RecBi3.2 и RecBi3.3 соответственно.

IBRS. Модель **IBRS** использует матрицы A_f и B_f и предназначена для формирования некоторому пользователю $u_0 \in U$ топ N рекомендаций, представленных математически специальной структурой $Top_N(u)$. Формально $Top_N(u_0)$ — это тройка $(R_{u_0}, \preceq_{u_0}, \text{rank})$, где R_{u_0} — множество не более N радиостанций, рекомендуемых конкретному пользователю u_0, $\preceq_{u_0}$ является хорошо определенным квазипорядком (рефлексивным, транзитивным и полным) на множестве R_{u_0}, а rank — функция, которая отображает каждую радиостанцию r из R_{u_o} в $[0, 1]$. Без ограничения общности мы сичтаем здесь, что чем выше значение rank, тем выше релевантность для пользователя радиостанции.

Алгоритм **RecBi3.1** вычисляет по 1-норме расстояние между пользователем u_0 и радиостанцией r, т. е. $d(u_0, r) = \sum\limits_{t\in T} |a_{u_0 t} - b_{rt}|$. Затем вычисляются все расстояния между пользователем u_0 и радиостанциями $r \in R$. Далее алгоритм строит отношение $\prec_{u_0}$ согласно следующему правилу: $r_i \preceq r_j$ тогда и только тогда, когда $d(u_0, r_i) \leqslant d(u_0, r_i)$. Функция rank определена на R_{u_0} согласно следующему правилу:

$$\text{rank}(r_i) = 1 - d(u_0, r_i) / \max_{r_j \in R} d(u_0, r_j).$$

Окончательно, после отбора N радиостанций для N наибольших значений rank на множестве R_{u_0}, мы получаем структуру $Top_N(u_0)$, которая представляет ранжированный список радиостанций, рекомендуемых пользователю u_0.

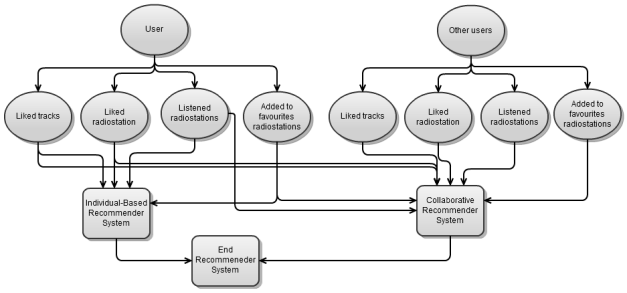

Рис. 1. Архитектура рекомендательной системы

Как показано на рисунке 1 наша модель принимает во внимание не только прослушанные треки, но и треки понравившиеся пользователю, понравившиеся радиостанции и радиостанции добавленные в фавориты. Для уточнения модели IBRS мы настраиваем ее параметры с помощью алгоритма SMARTS, известного в теории принятия решений [9]. Согласно методу и решению эксперта, мы должны принять во внимание каждый тег трека прослушанной радиостанции, понра-

вившуюся радиостанцию, понравившийся трек и радиостанцию-фаворит с различными весами. Процедура SMARTS предоставила нам четыре веса согласно оценкам наших экспертов взаимной важности критериев, а именно 0.07, 0.16, 0.3 и 0.47. В методе SMARTS, мы рассматриваем каждый тип тега как критерий с двумя терминальными значениями 0 и 100% действительной шкалы. Некоторый тег t может иметь несколько типов одновременно; в этом случае алгоритм добавляет к a_{ut} общий вес тега (т. е. сумму весов) после посещения пользователем u некоторой радиостанции с этим тегом.

В случае, когда существует несколько элементов с равными рангами, таким образом, что $Top_N(u)$ не однозначно определен, мы просто отбираем первые элементы согласно некоторому произвольному упорядочению (например, по лексикографическому порядку имен станций).

CBRS. Модель **CBRS** использует матрицу C_f. Матрица также порождает вектор n^C, который хранит общее число прослушанных радиостанций для каждого пользователея $u \in U$. Этот вектор также изменяется по времени, а его значения используются как пороги для преобразования матрицы C_f в матрицу расстояний D следующим образом:

$$d_{ijr} = \begin{cases} |c_{fir} - c_{fjr}|, & if\ c_{fir} \geqslant n_i^{-1} \text{и } c_{fjr} \geqslant n_j^{-1} \\ |c_{fir} + c_{fjr}|, & if\ c_{fir} > n_i^{-1} \text{и} \\ & c_{fjr} < n_j^{-1} \text{или наоборот} \end{cases}$$

Это расстояние принимает во внимание количество n_u^C всех посещений радиостанций для пользователя u и рассматривает ее обратное значение как порог для принятия решения о том, стоит ли рассматривать радиостанцию r как популярную для данного пользователя. Таким образом пользователи с различными знаками $c_{fir} - n_i^{-1}$ и $c_{fjr} - n_j^{-1}$ становятся более удаленными, чем в случае обычного абсолютного расстояния. Это расстояние d_{ij} действительно служит своеобразным поляризующим фильтром и в разделе 1 мы сравниваем его с обычным подходом.

После вычисления D алгоритм **RecBi3.2** строит список $Top_k(u_0) = (U_{u_0}, \preceq_{u_0}, \text{sim})$ из k пользователей сходных с нашим целевым пользователем u_0, который ожидает рекомендаций, где $\text{sim}(u) = 1 - d_{uu_0} / \max_{u' \in U} d_{u'u_0}$. Мы определяем множество всех радиостанций, прослушанных пользователем u_0, как $L(u_0) = \{r | c_{fur} \neq 0\}$. Сходным образом мы определяем

$$\begin{aligned} Top_N(u_0) &= (R_{u_0}, \preceq_{u_0}, \text{rank}), \text{ где} \\ \text{rank}(r) &= \text{sim}(u^*) \cdot c_{fu^*r} \text{ и} \\ u^* &= \arg \max_{u \in U_{u_0}, r \in R/L(u_0)} \text{sim}(u) \cdot c_{fur}. \end{aligned}$$

Стоит упомянуть, что $\text{rank} : r \mapsto [0, 1]$.

ERS. После того как IBRS и CBRS модели предоставили рекомендации мы имеем два ранжированных списка рекомендованных радиостанций $Top_N^I(u_0)$ и $Top_N^C(u_0)$ для нашего целевого пользователя u_0 из IBRS и CBRS соответственно. Подмодель **ERS** предлагает простое решение для агрегирования этих списков в итоговую структуру рекомендаций $Top_N^E(u_0) = (R_{u_0}^E, \preceq_{u_0}^E, \text{rank}^E)$. Для каждого $r \in R_{u_0}^C \cup R_{u_0}^I$, функция $\text{rank}^E(r)$ вычисляет r с помощью взвешенной суммы

$$\beta \cdot \text{rank}^C(r) + (1 - \beta) \cdot \text{rank}^I(r),$$

где мы задаем $\beta \in$ [0,1], $\text{rank}^C(r) = 0$ для всех $r \notin R^C$ и $\text{rank}^I(r) = 0$ для всех $r \notin R^I$. Алгоритм **RecBi3.3** добавляет N лучших радиостанций согласно этому критерию в множество $R_{u_0}^C$.

Оценка качества обслуживания

Для оценки качества разработанной системы мы предлагаем разновидность скользящего контроля[10]. Прежде чем продолжить детальное описание этой процедуры, мы обсудим некоторые важные результаты анализа данных FMhost за период с 2009 по 2011.

Базовая статистика. Хорошо известный факт в социальных сетях, что данные часто следуют так называемому степенному распределению [11]. Для того, чтобы решить какое количество пользователей мы должны принимать во внимание для формирования рекомендаций, мы провели простой статистический анализ активности 20% пользователей (только зарегистрированные) и радиостанций.

Таблица 1. Посещаемость пользователей и радиостанций.

Dataset	n	$\langle x \rangle$	σ	α	*p-value*
User dataset	4187	5.86	12.9	2.46(0.096)	**0.099**
Radio dataset	2209	11.22	60.05	2.37(0.22)	**0.629**

Таблица 1 показывает значения p-value для статистических тестов [11], которые подтверждают, что на данных по посещению радиостанций выполняется степенной закон, а для данных о посещаемости пользователей вероятность совершить ошибку отвергнув нулевую гипотезу (нет степенного закона) около 0.1. Мы учитываем степенной характер распределения для обоих наборов данных.

Проведенный анализ позволяет сделать полезное заключение согласно хорошо известному «80:20» правилу $W = P^{(\alpha-2)/(\alpha-1)}$, которое означает, что доля богатства W находится в руках части населения P. В нашем случае 50% пользователей

совершают 80% всех радиопрослушиваний, и 50% всех радиостанций привлекают 83% всех слушателей. Таким образом, если сервис стремится принимать во внимание только активных пользователей и радиостанции, то достаточно покрыть 80% всех посещений, рассматривая только 50% активных пользователей. Однако новые радиостанции заслуживают включения в списки рекомендаций, поэтому правило целесообразно применять только к набору данных о пользователях.

Оценка качества. Для оценки качества обслуживания (QoS) подсистемы IBRS (алгоритм RecBi3.1) мы вычисляем среднюю точность и полноту на множестве $R_N \subset R$, в котором N — доля «скрытых» радиостанций. Мы предполагаем, что для всех r из R_N и любого пользователя $u \in U$ алгоритм не знает была ли радиостанция им отмечена как понравившаяся, добавлена в фавориты или же просто посещена, мы также изменяем A_f и R соответсвующим образом. Затем алгоритм RecBi3.1 пытается рекомендовать Top-N радиостанций для измененной матрицы A_f.

Средняя Top-N точность и полнота вычисляются следующим образом:

$$\text{Precision} = \frac{\sum\limits_{u \in U} \frac{|R_u^I \cap L_u \cap R_N|}{|L_u \cap R_u^I|}}{|U|}, \ \text{Recall} = \frac{\sum\limits_{u \in U} \frac{|R_u^I \cap L_u \cap R_N|}{|L_u \cap R_N|}}{|U|}.$$

Для оценки CBRS мы используем модификацию скользящего контроля по схеме «исключай по одному». На каждом шаге процедуры для конкретного пользователя u, мы «скрываем» все радиостанции $r \in R_N$ посредством зануления $c_{fur} = 0$. Затем мы применяем RecBi3.2 в предположении, что $c_{fu'r}$ не изменилась для пользователя $u' \in U/u$. Далее мы находим среднюю точность и полноту для CBRS.

Для настройки системы ERS мы используем комбинацию двух указанных процедур, подбирая оптимальное β следующим образом

$$\beta^* = \arg\max_{\beta} \frac{2 \cdot \text{Precision} \cdot \text{Recall}}{(\text{Precision} + \text{Recall})}.$$

Мы полагаем, что статистики о месячной посещаемости 50% наиболее активной аудитории достаточно для подбора β и выбора подходящей меры сходства и расстояния, также как и порогов. Для оценки качества мы предполагаем вычислять Top-10 точность и полноту. Дополнительный контроль качества проводится путем опроса.

Заключение и дальнейшая работа

Мы надеемся, что разработанные алгоритмы помогут найти релевантные пользователю радиостанции для прослушивания. Мы рассмотриваем методы матричной факторизации как подходящий инструмент для повышения масштабируемости. Важным вопросом является работа с триадической природой данных (пользователи, радиостанции (треки), и теги). Как показано в [12] такие данные могут быть успешно проанализированы средствами трикластеризации, на основе которой мы планируем создать рекомендательную систему.

Литература

[1] *Hilliges O., Holzer P., Klüber R., Butz A.* Audioradar: A metaphorical visualization for the navigation of large music collections // Smart Graphics, LNCS. Springer, 2006. — Vol. 4073. 82–92

[2] *Gleich D. F., Zhukov L., Rasmussen M., Lang K.* The World of Music: SDP Embedding of High Dimensional data // Information Visualization, 2005.

[3] *Brandenburg K., Dittmar C., Gruhne M., Abeßer J., Lukashevich H., Dunker P., Gartner D., Wolter K., Grossmann H.* Music search and recommendation // Handbook of Multimedia for Digital Entertainment and Arts. Springer, 2009. — Pp. 349–384.

[4] *Celma Ò.* Music Recommendation and Discovery - The Long Tail, Long Fail, and Long Play in the Digital Music Space Springer, 2010.

[5] *Avesani P., Massa P., Nori M., Susi A.* Collaborative radio community // The Second International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems. Springer, 2002. — Pp. 462–465.

[6] *Nanopoulos A., Rafailidis D., Symeonidis P., Manolopoulos Y.* Musicbox: Personalized music recommendation based on cubic analysis of social tags // IEEE Transactions on Audio, Speech & Language Processing, 2010. — Vol. 18, No. 2. — Pp. 407–412.

[7] *Ignatov D., Poelmans J., Zaharchuk V.* Recommender System Based on Algorithm of Bicluster Analysis RecBi // Concept Discovery in Unstructured Data, 2011. — CEUR-WS, Vol. 757. — Pp. 122–126.

[8] *Ignatov D. I., Kuznetsov S. O.* Concept-based Recommendations for Internet Advertisement // CLA 2008, Palacky University, Olomouc, 2008. — Pp. 157–166.

[9] *Edwards W., Barron F.* SMARTS and SMARTER: Improved Simple Methods for Multiattribute Utility Measurement // Organizational Behavior and Human Decision Processes, 1994. — Vol. 60, No. 3. — Pp. 306–325.

[10] *Ignatov D. I., Poelmans J., Dedene G., Viaene S.* A New Cross-Validation Technique to Evaluate Quality of Recommender Systems // PerMIn, LNCS. Springer, 2012. — Vol. 7143. — Pp. 195–202.

[11] *Clauset A., Shalizi C. R., Newman M. E. J.* Power-law distributions in empirical data // SIAM Rev, 2009. — Vol. 51, No. 4. — Pp. 661–703.

[12] *Ignatov D. I., Kuznetsov S. O., Magizov R. A., Zhukov L. E.* From Triconcepts to Triclusters // RSFDGrC'11, Springer, 2011. — Pp. 257–264.