Министерство науки и высшего образования Российской Федерации
Федеральное государственное автономное образовательное
учреждение высшего образования
«Московский физико-технический институт (национальный
исследовательский институт)»
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

Построение мультимодальной рекомендательной системы

Выпускная квалификационная работа
(бакалаврская работа)

Направление подготовки: 03.03.01 Прикладные математика и физика

Выполнил:
студент группы 574
Кислинский Вадим Геннадьевич

(подпись обучающегося)

Научный руководитель:
Доктор физико-математических наук
Воронцов Константин Вячеславович

(подпись научного руководителя)

Москва 2019

Аннотация

Работа посвящена алгоритму, на базе которого предлагается построить совместную рекомендательную систему двух различных онлайн-платформ. Рассматривается мультимодальная тематическая модель, являющаяся гибридным алгоритмом в области рекомендательных систем. Предложенный подход предоставляет аппарат для работы с различными типами объектов и транзакций.

Ключевые слова: *рекомендательные системы, коллаборативная фильтрация, тематическое моделирование, мультимодальные модели.*

Содержание

1 Введение

Рекомендательные система — это неотъемлемая часть крупных онлайн-платформ, она помогает пользователям находить релевантный контент в огромном объеме информации и представляет персональный интерес. Рекомендательные системы имеют две основные группы методов: коллаборативная фильтрация и контентно-основанные(англ. content-based) методы. Контентно-основанные методы используют свойства объектов для генерации рекомендаций, например, цена, тип, цвет товара в онлайн-магазине или автор и жанр в онлайн-журнале. Создают профиль пользователя, опираясь на свойства объектов, с которыми пользователь взаимодействовал ранее, и рекомендуют объекты наиболее похожие на полученный профиль. Коллаборативная фильтрация использует историю действий пользователей в системе для получения новых рекомендаций. Методы коллаборативной фильтрации основаны на предположении о том, что те, кто одинаково оценивал какие-либо объекты в прошлом, склонны давать похожие оценки в будущем. Такие методы устойчивы и способны выявлять скрытые свойства объектов, что улучшает релевантность рекомендаций.

В данной работе постановка задачи отличается от стандартной задачи *top-k* рекомендаций тем, что предлагается построить рекомендательную систему сразу для двух онлайн-платформ: интернет-магазина и интернет-журнала, имеющих разные типы объектов и транзакций. Учет пользовательских действий с различных онлайн-платформ позволяет увеличить базу пользователей на каждой платформе в отдельности. Онлайн-магазин и журнал являются одной экосистемой, в журнале пишут обзоры на товары и рекомендации покупок, магазин продает те самые товары. Совместная рекомен-

дательная система будет являться связующим звеном между этими платформами. В работе предложен и исследован алгоритм решения нестандартной задачи в области рекомендательных систем с помощью мультимодальной тематической модели. Кроме того, предложенный алгоритм является гибридным алгоритмом, он одновременно использует и свойства объектов, и историю действия пользователя для генерации рекомендаций. Гибридный алгоритм позволяет решать проблему холодного старта для объектов, которая заключается в том, что объект впервые появляется в системе и не имеет истории, тем самым теряет возможность оказаться в рекомендациях.

Одна из проблем контентно-основанных и коллаборативных методов — это неспособность предоставлять персональные рекомендации объектов пользователям-новичкам. Вместо этого пользователям-новичкам рекомендуются самые популярные или имеющие высокий рейтинг объекты. Предложенный в работе алгоритм частично решает эту проблему. Так, например, пользователю интернет-магазина, который ранее не читал статей в интернет-журнале, алгоритм составит списки рекомендаций статей на основе его покупок и наоборот.

2 Анализ литературы

В [1] рассмотрены основные подходы, применяемые в рекомендательных системах, и способы их комбинирования. Выделяются основные недостатки методов коллаборативной фильтрации: проблема холодного старта для пользователей и объектов, необходимость больших объемов данных для получения высокого качества. Предложены методы их решения путем комбинирования с контентно-основанными подходами. В работе [2] классический подход, основанный на SVD разложении матрицы рейтингов, сравнивается с гради-

ентными подходами разложения матрицы рейтингов. Полученный результат говорит о том, что этот метод не уступает градиентным подходам. В [4] предложен алгоритм решения проблемы холодного старта для объектов с помощью совместной матричной факторизации матрицы рейтингов и матрицы признакового описания объектов.

Работа [3] посвящена тематическому моделированию. Рассмотрены основные методы этой области, подробно описан подход, использующий идею аддитивной регуляризации. Предложены подходы для моделирования мультимодальных и транзакционных данных. Описаны возможности применения тематических моделей в различных областях, в том числе в рекомендательных системах.

В работах [5], [6] рассматриваются подходы к коллаборативной фильтрации на основе матричной факторизации. В статье [5] описывается подход для выборок, в которых собраны явные предпочтения пользователей, оценки объектов от пользователей по некоторой шкале. В работе [6] предложен алгоритм для выборок с неявными предпочтениями от пользователя, такие выборки содержат информацию только о том, какие объекты предпочел пользователь.

3 Постановка задачи

Целью работы является построение рекомендательной системы для двух онлайн-платформ: интернет-магазина и интернет-журнала. Перед рекомендательной системой ставятся задачи подбора персональных рекомендаций товаров и статей для пользователей, а также подбора товаров к статьям.

Заданы:

- множество пользователей U и множество товаров I;
- коллекция статей D со словарем W;
- транзакции пользователей в интеренет-магазине и интернет-журнале;
- связи между статьями и товарами.

Каждый товар i имеет категорию c, которая объединяет товары в группы по некоторым характеристикам. Транзакцией в интернет-магазине является покупка пользователем u товара i, транзакция в интернет-журнале — это информация о том, что пользователь u перешел на страницу статьи d, связь пары (d, i) означает, что в статье d есть ссылка на товар i.

Требуется для каждого пользователя $u \in U$ построить список из k наиболее релевантных статей и список из k наиболее релевантных товаров, каждый из которых отсортирован по убыванию релевантности.

4 Критерии качества

Оценка качества рекомендательных систем сводится к сравнению двух списков: списка рекомендаций и списка релевантных для пользователя объектов. Список релевантных для пользователя объектов имеет нефиксированную длину, которая меняется от пользователя к пользователю. Поэтому, сравнивая его со списком рекомендаций, который имеет постоянную длину k, важно учесть не только совпадение элементов в списках, но и порядок объектов в списке рекомендаций: релевантные должны находиться выше нерелевантных.

Введем основные критерии качества. Метрика $recall@k$, вычисляемая по формуле (1), считает долю верно угаданных рекомендаций, среди объектов релевантных пользователю. Метрика $map@k$ – формула (2), помимо оценки релевантности рекомендаций, учитывает порядок объектов в списке рекомендаций, штрафует, если нерелевантный объект находится выше, чем релевантный.

$$recall@k = \frac{\sum_{u \in U} \sum_{i=1}^{k} r_u(i)}{\sum_{u \in U} min(l_u, k)} \quad , \tag{1}$$

$$map@k = \frac{1}{|U|} \sum_{u \in U} \frac{1}{min(l_u, k)} \sum_{i=1}^{k} r_u(i) precision_u@i \quad , \tag{2}$$

$$precision_u@k = \frac{1}{k} \sum_{i=k}^{k} r_u(i) \quad ,$$

где l_u — количество релевантных для пользователя объектов, $r_u(i)$ —

релевантный для пользователя объект или нет (1 или 0).

Кроме того для оценки качества рекомендаций товаров было введено два критерия «жесткий» и «мягкий».
«Жесткий» критерий:

$$r(i) = \begin{cases} 0, & \text{если товар } i \text{ не релевантен пользователю;} \\ 1, & \text{если товар } i \text{ релевантен пользователю.} \end{cases}$$

«Мягкий» критерий:

$$\hat{r}(i) = \begin{cases} 0, & \text{если категория товара } i \text{ не релевантна пользователю;} \\ 1, & \text{если категория товара } i \text{ релевантна пользователю.} \end{cases}$$

Критерии введены по причине того, что в интернет-магазине имеется широкий выбор товаров — несколько миллионов различных предложений, но при этом большинство товаров объединяются в группы, отличаясь лишь некоторыми не значимыми свойствами. Поэтому угадать точное предпочтение пользователя достаточно сложно, но при этом модель может хорошо предсказывать общие интересы пользователя.

5 Мультимодальная тематическая модель

При составлении списков рекомендаций необходимо выявлять какой объект из пары (i_1, i_2) наиболее релевантен для пользователя u. Сопоставим каждой паре пользователь-объект (u, i_1) и (u, i_2) вероятность транзакции, и будем считать, что чем выше вероятность транзакции, тем релевантнее объект. Таким образом для составления списков рекомендаций достаточно восстановить распределения

товаров и статей в контексте пользователя, т.е. восстановить распределения $p(i|u)$ и $p(d|u)$.

Определим тематическую модель в данной задаче. Имеем четыре типа транзакций:

- пользователь u купил товар i;
- пользователь u прочитал статью d;
- слово w входит в статью d;
- товар i связан со статьей d.

Построим граф связи между различными типами данных:

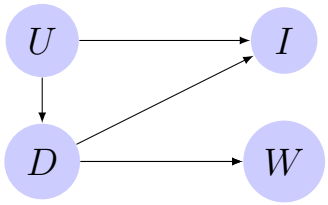

Рис. 1: Граф, описывающий связи между типами данных

Роль документов, в данной задаче, играют пользователи и статьи. Пользователи связаны с *модальностями*, образуемыми статьями и товарами, статьи — с *модальностями* слов и товаров. Объединение множеств U и D обозначим H, объединение множеств I, W и D — V. Введем множество типов транзакций K, транзакционные данные типа k — это выборка E_k независимых наблюдений $(h, v) \in H \times V$ размера n_k. Каждый элемент $(h, v) \in E_k$ входит в выборку n_{khv} раз. Будем считать, что появление элемента $v \in V$ вместе с элементом $h \in H$ связано с темой t из некоторого множества T и выполнена гипотеза условной независимости, т.е. $p(v|h, t) = p(v|t)$.

Восстановив распределения тем в каждом типе «документов» $p(t|h)$ и распределение «слов» по темам $p(v|t)$, получим необходимые распределения товаров по пользователям

$$p(i|u)|_{i=v,u=h} = \sum_{t\in T} p(t|h)p(v|t) \quad ,$$

статей по пользователям

$$p(d|u)|_{d=v,u=h} = \sum_{t\in T} p(t|h)p(v|t) \quad ,$$

а также товаров по документам

$$p(i|d)|_{i=v,d=h} = \sum_{t\in T} p(t|h)p(v|t).$$

Выпишем функцию правдоподобия выборки E_k от параметров модели $\theta_{th} = p(t|h)$ и $\varphi_{vt} = p(v|t)$ по каждому типу k транзакционных данных.

$$p_k((h_i, v_i)_{i=0}^{n_k}; \Phi, \Theta) = \prod_{i=0}^{n_k} p(v_i, h_i) = \prod_{hv\in E_k} p(v|h)^{n_{khv}} \underbrace{p(d)^{n_{khv}}}_{const}$$

Для оптимизации параметров модели воспользуемся принципом максимума правдоподобия, будем максимизировать взвешенную сумму логарифмов правдоподобия модели по каждому типу транзакций и регуляризатора $R(\Phi, \Theta)$.

$$\sum_{k\in K}\tau_k\sum_{hv\in E_k} n_{khv}\ln\sum_{t\in T}\theta_{th}\varphi_{vt} + R(\Phi,\Theta)\to \max_{\Phi,\Theta} \tag{3}$$

$$\sum_{v\in V_m}\varphi_{vt}=1,\quad \varphi_{vt}\geqslant 1;\quad \sum_{t\in T}\theta_{th}=1,\quad \theta_{th}\geqslant 1$$

Воспользуемся EM-алгоритмом для решения задачи (3), предложенном в работе [3]. На первом шаге случайным образом зададим матрицы Φ и Θ. Далее на E-шаге будем пересчитывать вспомогательные переменные p_{tdv} при фиксированных матрицах Φ и Θ, по формулам:

$$p_{thv} = \underset{t\in T}{norm}(\theta_{th}\varphi_{vt}). \tag{4}$$

На M-шаге обновляются параметры θ_{th} и ϕ_{vt} по правилам:

$$\varphi_{vt} = \underset{v\in V_m}{norm}(n_{vt} + \varphi_{vt}\frac{\partial R}{\partial \varphi_{vt}});\quad n_{vt} = \sum_{k\in K}\sum_{hv\in E_k} n_{khv}p_{thv}, \tag{5}$$

$$\theta_{th} = \underset{t\in T}{norm}(n_{th} + \theta_{th}\frac{\partial R}{\partial \theta_{th}});\quad n_{th} = \sum_{k\in K}\sum_{hv\in E_k} n_{khv}p_{thv}. \tag{6}$$

6 Составление списков рекомендаций

Полученные матрицы Φ и Θ имеют блоки, каждый из которых описывает распределение тем в соответствующем типе данных. Матрица Φ имеет три блока: первый блок порождает распределение товаров по темам, второй — распределение статей по темам, третий описывает распределение слов по темам. Матрица Θ состоит из бло-

ка распределения тем по пользователям и блока распределения тем по статьям.

$$\Phi = \begin{Vmatrix} \Phi_I \\ \Phi_D \\ \Phi_W \end{Vmatrix}; \quad \Theta = \begin{Vmatrix} \Theta_U & \Theta_D \end{Vmatrix}$$

Скалярно умножая столбец матрицы Θ на строку матрицы Φ, получим вероятность соответствующей транзакции. Например, взяв из матрицы Θ столбец, отвечающий пользователю u, и скалярно умножив его на строку матрицы Φ, соответствующую товару i, получим вероятность покупки пользователем u товара i.

Опишем процесс генерации списков персональных рекомендаций для пользователя u. Пусть I_u множество товаров, с которыми пользователь уже взаимодействовал, D_u — список, прочитанных статей. Тогда список персональных рекомендаций товаров определим, как:

$$\underset{v \in I \setminus I_u}{argtop_k} \sum_{t \in T} \theta_{tu} \varphi_{vt}. \tag{7}$$

Список персональных рекомендаций статей определим, как:

$$\underset{v \in D \setminus D_u}{argtop_k} \sum_{t \in T} \theta_{tu} \varphi_{vt}, \tag{8}$$

где top_k — операция нахождения k максимальных элементов и сортировка их в порядке убывания.

Получим распределение $p(d|t)$ для нового документа d. По определению условной вероятности

$$p(d|t) = \frac{p(d,t)}{p(t)} = \frac{p(t|d)p(d)}{p(t)} \quad , \tag{9}$$

распределение $p(t|d)$ найдем с помощью блока Φ_W матрицы Φ. Для этого случайно зададим параметры θ_{td}, после чего по формулам (4), (6) пересчитаем их по известным параметрам φ_{vt}. Предполагая, что темы порождаются дискретным равномерным распределением, по формуле (9) найдем искомое распределение $p(d|t)$. Таким образом, при появлении новой статьи в журнале нет необходимости пересчитывать всю модель, достаточно дополнить блок матрицы Φ, отвечающий за статьи.

7 Вычислительный эксперимент

7.1 Описание данных

Для проведения экспериментов предоставлены данные за год активности пользователей на сайтах интернет-магазина и интернет-журнала, тексты всех прочитанных статей и информация о том, какие-то товары привязаны к различным статьям.

За имеющийся период 2,5 миллиона уникальных пользователей совершили покупки в интернет-магазине, 270 тысяч из них прочитали хотя бы одну статью, куплено около 1,5 миллиона различных товаров, прочитано 700 статей. На сайте онлайн-магазина совершено 65 миллионов транзакций и 500 тысяч транзакций в интернет-журнале. Таким образом разреженность матрицы частот очень высокая, около $2 \cdot 10^{-6}$. Такая высокая разреженность является проблемой для большинства алгоритмов коллаборативной фильтрации, чтобы ее решить были предложены методы матричной факторизации, которые при обучение используют только известные транзакции пользователей [5]. Тематическая модель также используют только известные транзакции.

Модальность	Объем
Пользователи	2,5млн
Товары	1,5млн
Статьи	700
Слова	25000

Таблица 1: Объемы модальностей

Вид	Количество
Пользователь-товар	65млн
Пользователь-статья	0,5млн
Статья-товар	3500
Статья-слово	0.47млн

Таблица 2: Количество транзакций

7.2 Стратегия валидации

При оценки качества рекомендательных систем используют две стратегии: онлайн и оффлайн. Онлайн оценка качества заключается в том, что пользователям в реальном времени показываются рассчитанные рекомендации, после чего снимаются показатели различных метрик. Оффлайн стратегия используют исторические данные для подсчета метрик, поэтому в этом случае необходимо учесть специфику данных. В данной работе были проведены только оффлайн эксперименты.

Проводя оффлайн эксперимент оценки качества рекомендательной системы, попробуем максимально воссоздать идею онлайн-стратегии оценки качества. Учтем две особенности, первая заключается в том, что рекомендательная система должна предсказывать будущий интерес пользователя на основе исторических данных, вторая особенность — это непохожесть пользователей друг на друга.

Разобьем транзакциионные данные на две части по временной отметке, на обучение возьмем 11 месяцев, 1 месяц отложим в качестве тестовой выборки. Тестовую выборку поделим на две части по пользователям, на одной из них будем подбирать гиперпараметры, с помощью второй будем отслеживать результаты, чтобы не подстроиться под конкретных пользователей.

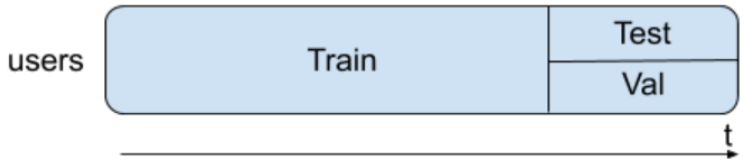

Рис. 2: Стратегия валидации

Помимо этого разобьем тестовую выборку на несколько групп по пользователям на основе их активности на сайте интернет-магазина. С помощью полученных подвыборок исследуем вопрос зависимости качества модели от количества известных транзакций.

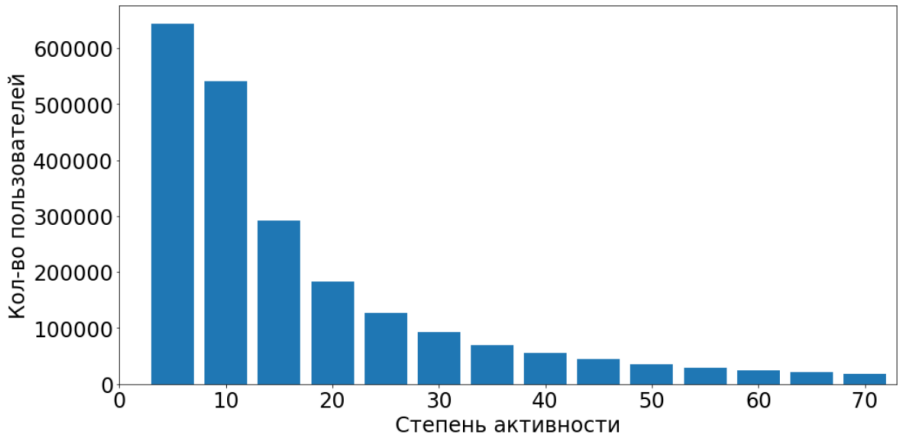

Рис. 3: Распределение пользователей по кол-ву транзакций

7.3 Оптимизация количества тем

Количество тем является одним из важнейших гиперпараметров тематической модели. Он значимо влияет на качество и в стандартных задачах тематического моделирования текстов, и в остальных приложениях тематического моделирования.

В данной работе предлагается подобрать оптимальное количество тем по внешним критериям качества рекомендательной системы $map@k$ и $recall@k$. Вторым важным критерием является вычислительная нагрузка, которая растет линейно от количества тем по времени и квадратично по памяти. Ниже на рис.4 и рис.5 представлены графики зависимости качества модели от количества тем. Тесты были проведены на валидационной выборке.

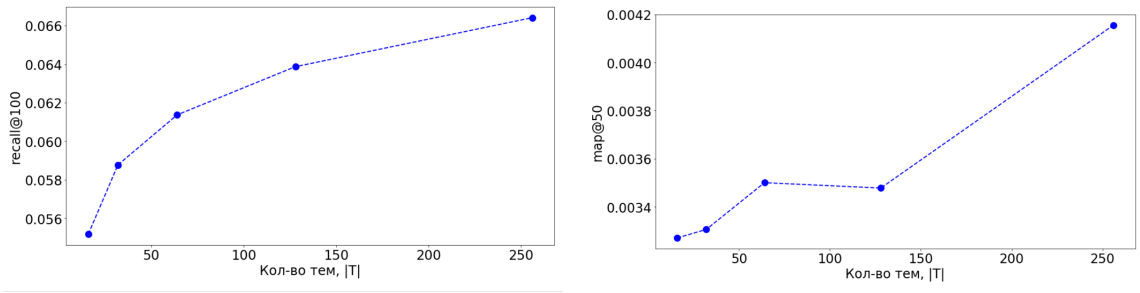

Рис. 4: Качество рекомендаций товаров

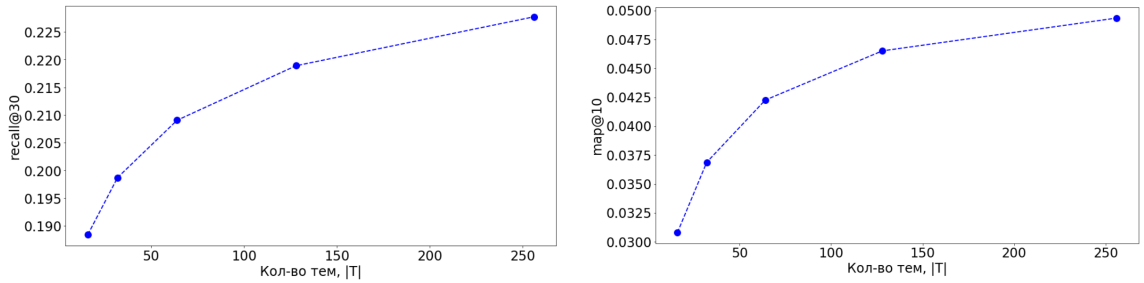

Рис. 5: Качество рекомендаций статей

Из графиков видно, что качество рекомендательной системы растет при увеличении количества тем, но темп роста метрик падает, при этом одновременно с качеством растут и вычислительные нагрузки модели. Исходя из этих рассуждений, количество тем было выбрано равным 256.

7.4 Зависимость качества от активности пользователя

Одной из стандартных проблем основанных на моделях (англ. model-based) подходов к коллаборативной фильтрации является то, что они переобучаются в сторону активных пользователей и популярных объектов. Проявляется такой эффект в том, что рекомендации активных пользователей имеют более высокие метрики качества, а

популярные товары чаще остальных попадают в списки рекомендаций.

Проведем эксперимент, проверяющий описанный эффект. Возьмем группы пользователей, имеющие разное количество транзакций в обучающей выборке, и рассчитаем для каждой группы пользователей метрики качества. На рис.6 представлены графики, иллюстрирующие результат такого эксперимента, по оси абсцисс отложены средние значения количества товаров у группы пользователей.

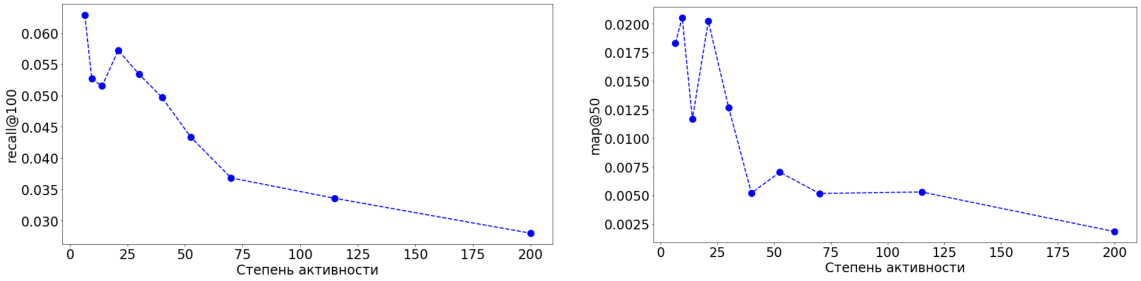

Рис. 6: Зависимость качества от активности пользователя

Полученные результаты говорят о том, что в данной задаче имеется обратный эффект — качество падает с увеличением количества транзакций. Объяснение нестандартного эффекта, предположительно, лежит в природе задачи. Активные пользователи уже сформировали собственную корзину и совершают транзакции известные в прошлом, а полученная модель фильтрует такие объекты, стараясь предложить пользователю что-то новое.

7.5 Результаты

При оценки качества рекомендательных систем важно оценить, как ведут себя метрики при различных длинах списков рекомендаций. Связано это с тем, что пользователи видят рекомендации блоками:

сначала они видят блок из 3-5 объектов, если ничего из начала списка не заинтересовало пролистывают дальше, пока не остановятся.

Ниже на рис.7 и рис.8 представлены графики с конечными метриками на тестовой выборке. В качестве модели взята модель с параметрами выбранными на валидационной выборке. В качестве алгоритма для сравнения выбран алгоритм матричной факторизации iALS [6], который также находит некоторые векторные представления пользователей и товаров.

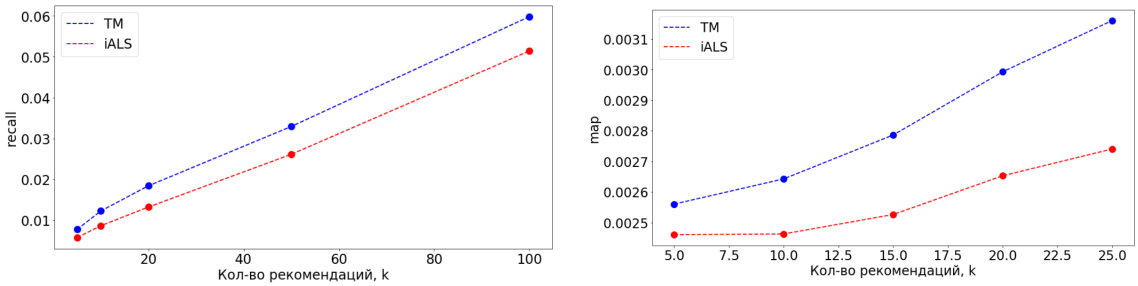

Рис. 7: «Жесткий» критерий

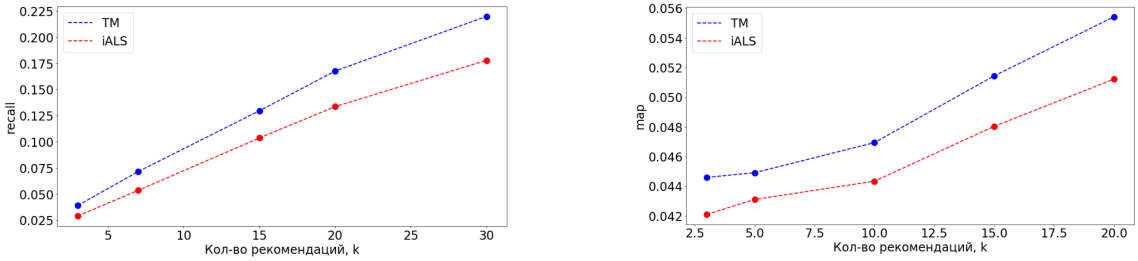

Рис. 8: «Мягкий» критерий

Тематическая модель превосходит алгоритм iALS. Интересно, что разница в качестве по «жесткому» и «мягкому» критериям очень высока. Например, по «жесткому» критерию тематическая модель достигает *recall* равный 6% только при $k = 100$, а по «мягкому» при $k = 5$. Такое различие говорит о том, что модель в целом выдает релевантный результат, но плохо выбирает между похожими

объектами. Эту проблему можно решить дополнительными операциями после составления длинного ($k \sim 1000$) списка рекомендаций, как часто делается на практике. Например, строят модель второго уровня, которая учитывает дополнительную информацию об объекте (цену, популярность, новизну и т.д.) и пользователе (пол, возраст, средний чек) и выбирает небольшой список рекомендаций из обширного списка, предложенного первой моделью.

8 Заключение

В работе рассмотрена задача построения совместной рекомендательной системы двух онлайн-платформ. Рассмотрен метод, основанный на мультимодальной тематической модели. Предложен алгоритм, с помощью которого решены задачи составления списков персональных рекомендаций пользователям интеренет-магазина и интернет-журнала, а также задача подбора товаров к статьям.

В ходе эксперимента были подобраны оптимальные параметры построенной мультимодальной тематической модели. Проведены сравнения с другими алгоритмами матричной факторизации. Выявлен нестандартный эффект, связанный с падением метрик качества у наиболее активных пользователей.

Список литературы

[1] Robin Burke. Hybrid Recommender Systems: Survey and Experiments. User Modeling and User-Adapted Interaction,12(4):331–370, 2002

[2] Cremonesi, P., Koren, Y., Turrin, R.. Performance of recommender algorithms on top-n recommendation tasks. Proceedings of the Fourth ACM Conference on Recommender Systems - RecSys '10. 2010

[3] Воронцов К.В. Вероятностное тематическое моделирование: обзор моделей и аддитивная регуляризация. 2019

[4] 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

[5] Y. Koren, R. Bell, C. Volinsky. Matrix factorization techniques for recommender systems, Computer 42(8), 30–37. 2009

[6] Y. Hu, Y. Koren, and C. Volinsky. Collaborative filtering for implicit feedback datasets. In Proc. IEEE ICDM (2008), pages 263–272, 2008.