СРАВНЕНИЕ РАЗЛИЧНЫХ МЕТОДОВ ПОСТРОЕНИЯ РЕКОМЕНДАЦИЙ НА ОСНОВЕ ДАННЫХ О ЗАКАЗАХ НАУЧНО-ТЕХНИЧЕСКОЙ БИБЛИОТЕКИ НИ ТПУ

-

Князева А.А.[(1)], Колобов О.С.[(2)], Турчановский И.Ю.[(1)]

[(1)] Институт вычислительных технологий, г. Новосибирск
[(2)] Институт сильноточной электроники, г. Томск

В работе рассматривается возможность создания рекомендательной системы для электронного каталога библиотеки на основе данных о заказах. Приводится сравнительный анализ двух подходов коллаборативной фильтрации: на основе документов и на основе пользователей.

Рекомендательная система, коллаборативная фильтрация, бинарные данные.

COMPARISON OF DIFFERENT METHODS FOR RECOMMENDATIONS CREATION ON BASE OF ORDER DATA FROM SCIENTIFIC AND TECHNICAL LIBRARY NR TPU

Knyazeva A.A. [(1)], Kolobov O.S. [(2)], Turchanovsky I.Yu. [(1)]

[(1)] Institute of Computational Technologies of the Siberian Branch of the Russian Academy of Science

[(2)] Institute of High Current Electronics of the Siberian Branch of the Russian Academy of Science, Tomsk

An opportunity of the recommender system creation for an online library catalog on base of order data is considered in this work. A comparison analysis of two collaborative filtering approaches is given: item-based and user-based.

Recommender system, collaborative filtering, binary data.

Введение. В настоящее время рекомендательные системы все более и более распространены во многих областях жизни. Очевидно, что они могут приносить пользу и для библиотек. Рекомендации пользователям новых книг (ссылок, файлов и пр.) позволяют создать новые возможности для пользователей библиотек и потенциальных читателей [1]. Цель данной работы – оценка возможности создания рекомендательной системы для электронного каталога НТБ ТПУ на основе данных о заказах документов. Задачи, поставленные для достижения цели: выбор наиболее перспективной модели построения рекомендаций и подбор некоторых параметров будущей системы.

Описание данных. В работе использовались данные о заказах читателей Научно-технической библиотеки Томского политехнического университета (НТБ ТПУ) за 2015 г., представленные в табличном виде. В первом столбце таблицы указаны идентификаторы пользователей (зашифрованные с помощью хеш-функции для обеспечения анонимности), во втором – идентификаторы документов. В качестве документа может выступать любой объект, библиографическое описание которого присутствует в электронном каталоге (книга, статья, цифровой носитель и т.д.). Каждая строка отражает факт заказа читателем документа без указания метки времени. Описанные данные являются бинарными. Это означает, что можно с уверенностью лишь о факте заказа. Рей-

тинги или оценки пользователями документов нам неизвестны. При этом отсутствие заказа может означать различные ситуации:

- данный документ не является релевантным;
- документ релевантный, он известен пользователю и, следовательно, не должен быть рекомендован;
- документ релевантен и неизвестен пользователю.

Очевидно, при построении рекомендаций необходимы документы последней группы. Однако выделить их на основе имеющихся данных не представляется возможным. Для создания тестовой выборки при оценке качества работы рекомендательной системы мы вынуждены использовать допущение, что все документы, которые не были заказаны, являются нерелевантными. Такой подход позволяет формировать группу нерелевантных документов для проверки без оценки пользователем каждого документа в коллекции, что, как правило, технически невозможно.

Дополнительно в работе был задействован набор данных под названием MSWeb, предоставляемый в рамках используемого инструментария. Данные получены путем выборочного анализа лог-файлов сайта *www.microsoft.com*. Они представляют собой записи об обращениях к различным областям сайта анонимных пользователей, выбранных случайным образом. Временной период: одна неделя в феврале 1998 года. В роли документа выступает область сайта. Набор данных MSWeb является вспомогательным. Его использование в данной работе обусловлено стремлением выделить особенности данных о заказах НТБ.

Для того, чтобы исключить из работы пользователей и документы, о которых слишком мало информации, были применены следующие фильтры (в указанном порядке):

1. Исключение документов, которые были заказаны менее чем 4 пользователями;
2. Исключение пользователей, которые заказали менее 4 документов.

Описанная фильтрация позволяет существенно сократить объем данных для работы (таблица 1). Кроме того, она позволяет составлять тестовую выборку из тех пользователей, кто заказал 4 и более документов. Это означает, что мы можем строить рекомендации на основе трех документов, а остальные использовать для проверки.

Таблица 1. Количественное описание данных.

Количество	**До фильтрации**		**После фильтрации**	
	НТБ	**MSWeb**	**НТБ**	**MSWeb**
Записей о заказах / просмотрах	98341	98653	51513	57497
Уникальных пользователей	9619	32710	4786	9544
Уникальных документов	37718	285	3764	231

Краткое описание инструментария и моделей. В работе была использована библиотека *recommenderlab* [2] для вычислительной среды R project. С помощью данной библиотеки для исходных данных были построены три варианта рекомендательных систем:

1. Рекомендации по популярности (*Popular*);

2. Коллаборативная фильтрация на основе документов (*Item-based collaborative filtering, IBCF*);

3. Коллаборативная фильтрация на основе пользователей (*User-based collaborative filtering, UBCF*).

Первый способ, при котором всем пользователям рекомендуются наиболее популярные документы, можно считать точкой отсчета для сравнения методов. Рекомендации, полученные с его помощью, не являются персонализированными, в отличие от двух других методов.

Модели на основе сходства документов используют предположение, что похожие между собой документы будут оцениваться пользователями сходным образом. Таким образом, производится вычисление меры схожести для каждой пары документов, и задействуются те документы, для которых значения меры наибольшие.

Модели на основе пользователей базируются на аналогичной идее: похожие между собой пользователи оценивают документы приблизительно одинаково. Для того, чтобы спрогнозировать оценку данным пользователем конкретного документа, можно использовать оценки других пользователей, похожих на данного пользователя [3]. Количество похожих документов или пользователей может варьироваться. В данной работе оно задается с помощью значения параметра *k*.

Используемые меры схожести. Для оценки того, насколько документы или пользователи похожи между собой были использованы следующие меры.

1. Коэффициент Жаккара [4];
2. Мера Дайса [5];
3. Косинусная мера [3];
4. Коэффициент корреляции Пирсона [3].

Описание экспериментов. Данные, используемые в работе, были случайным образом разбиты на обучающую (70 % пользователей) и тестовую (30%) выборки. Для пользователей из тестовой выборки, в свою очередь, производилось разделение документов. Для каждого пользователя были выбраны по 3 документа, на основе которых строились рекомендации. Размер списка рекомендаций описывается параметром *N*. Затем полученные рекомендации сравнивались с остальными «скрытыми» документами пользователя. По результатам сравнения были вычислены оценки качества работы системы.

Качество рекомендаций оценивалось с помощью показателей, традиционно используемых для оценки качества информационного поиска: полноты, точности и F-меры [6].

Коллаборативная фильтрация на основе документов. Результаты применения различных мер схожести для построения рекомендаций на основе документов можно проиллюстрировать с помощью так называемых кривых «полнота-точность» (рисунок 1).

Прежде всего, в приведенном рисунке привлекает внимание необычная форма кривой для косинусной меры. Нестандартное поведение косинусной меры связано с тем, как данная метрика подобия используется в библиотеке *recommenderlab*. Для того

чтобы разобраться в проблеме, была рассмотрена оценка качества рекомендаций для каждого пользователя из тестовой выборки (1436 пользователей). Показатель точности был определен только для 6 % пользователей тестовой выборки. Для остальных пользователей количество рекомендованных релевантных документов и рекомендованных нерелевантных документов равнялось нулю. В результате в знаменателе показателя точности для таких пользователей стояло нулевое значение. Причиной было то, что рекомендации для них не были сформированы.

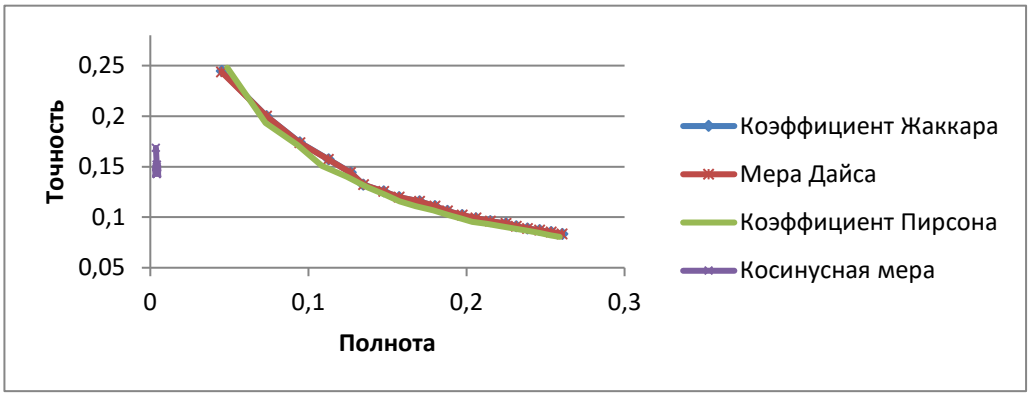

Рис. 1. Кривые «полнота-точность» для рекомендаций на основе документов в зависимости от используемой меры схожести; параметр k=50; количество рекомендаций изменяется от 1 до 20.

В качестве примера описанной ситуации рассмотрим одного из пользователей тестовой выборки. Данный пользователь заказал 13 документов, три из которых были случайным образом выбраны в качестве базы для создания рекомендаций. Для каждого из этих трех документов были рассчитаны значения косинусной меры схожести с каждым документом коллекции. Количество документов с ненулевым значением меры схожести хотя бы с одним из трех документов составило 874. При этом для 55 из них схожесть с каждой из трех книг оказалась максимальной (т.е. значение меры было равно 1). Параметр k в описанном примере был равен 50, то есть для каждого документа следовало выбрать 50 наибольших значений меры схожести. Возникла дилемма выбора 50 документов из 55, имеющих одинаковые оценки схожести. Соответствующая функция из библиотеки *recommenderlab* в этом случае не позволяет сделать выбор. В результате получаем пустое множество рекомендаций, неопределенную точность и нулевую полноту. При увеличении параметра k до 100 описанная коллизия разрешается, и для всех пользователей тестовой выборки формируются рекомендации. Однако и в этом случае качество рекомендаций для косинусной меры уступает остальным мерам схожести, использованным в данной работе.

Кривые, отображающие качество рекомендаций, построенных на основе коэффициента Жаккара и меры Дайса, практически совпадают. Кривая для коэффициента Пирсона проигрывает им, но незначительно.

Коллаборативная фильтрация на основе пользователей. Для вычисления сходства между пользователями использовались уже перечисленные меры схожести. Результаты построение рекомендаций проиллюстрированы на рисунке 2.

Наилучшие результаты для данных НТБ ТПУ показала косинусная мера, которой незначительно уступает коэффициент Жаккара. Иллюстрация того, как на качество рекомендаций влияет количество ближайших соседей, приведена на рисунке 3.

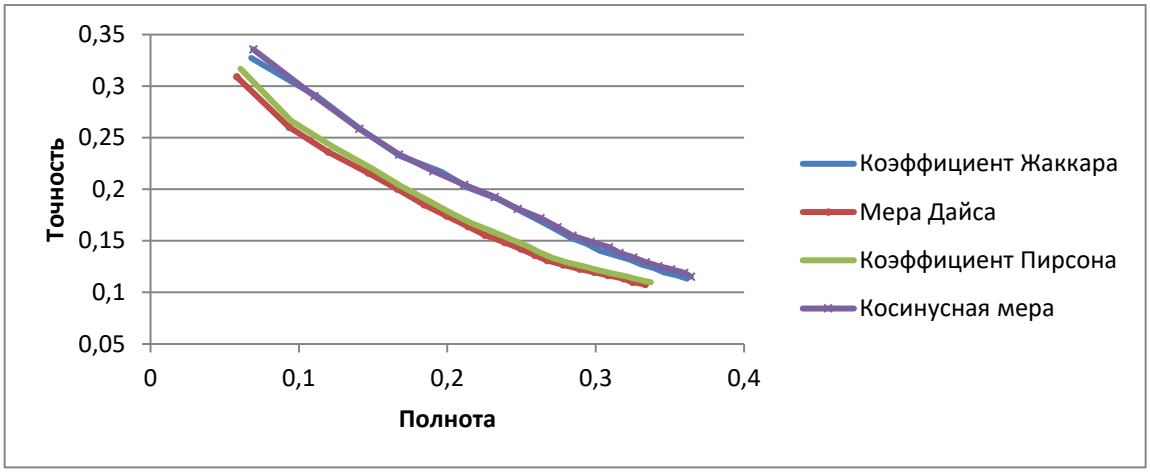

Рис. 2. Кривые «полнота-точность» для рекомендаций на основе пользователей в зависимости от используемой меры схожести; параметр k=50; количество рекомендаций изменяется от 1 до 20.

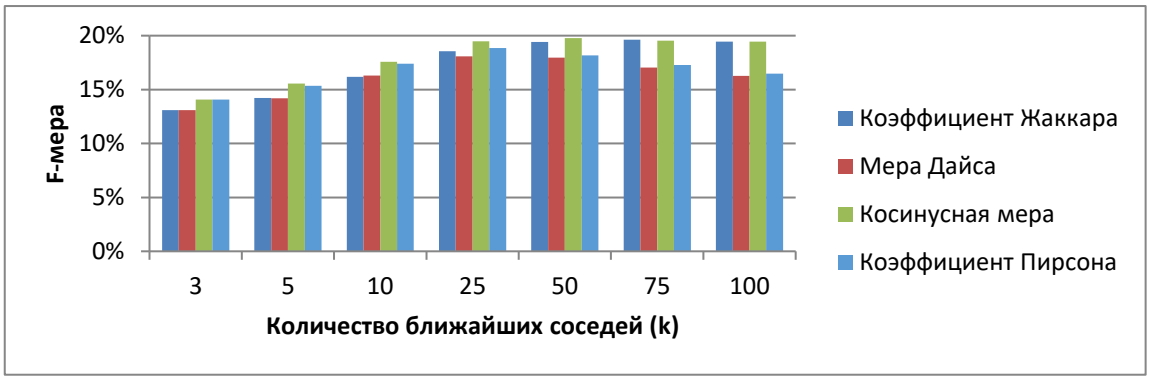

Рис. 3. F-мера для рекомендаций на основе пользователей в зависимости от меры схожести и значения параметра k (количество рекомендаций равно 10).

Таблица 2. Оценки качества (%) для списка из 10 рекомендаций (k=50).

Мера сходства	**Рекомендации, основанные на документах (item-based collaborative filtering)**			**Рекомендации, основанные на пользователях (user-based collaborative filtering)**		
	Точность	**Полнота**	**F-мера**	**Точность**	**Полнота**	**F-мера**
Жаккара	11,10	18,01	13,73	16,08	27,36	20,26
Пирсона	10,67	17,90	13,37	15,17	24,27	18,67
Косинусная	14,49	0,41	0,81	16,34	27,41	20,48
Дайса	11,09	18,01	13,73	14,87	23,86	18,32

Из рассмотренных значений параметров рекомендательной системы наиболее качественные рекомендации дает использование параметра k=50 в сочетании с косинусной мерой.

В таблице 2 приведены показатели качества для двух вариантов коллаборативной фильтрации.

Построение рекомендаций на основе пользователей позволяет добиться более хороших результатов, по сравнению с рекомендациями на основе документов. В таблице 3 приведены результаты лучших вариантов для трех описанных подходов.

Таблица 3. Сравнение качества рекомендаций для трех описанных подходов (N=10, k=50).

Показатель качества	**По популярности**		**На основе документов (коэф. Жаккара)**		**На основе пользователей (косинусная мера)**	
	НТБ	**MSWeb**	**НТБ**	**MSWeb**	**НТБ**	**MSWeb**
Точность, %	3,79	16,16	11,10	17,46	16,34	17,27
Полнота, %	4,78	57,63	18,01	64,37	27,41	65,19
F-мера, %	4,23	25,24	13,73	27,47	20,48	27,31

Для всех методов параметр N=10. Для методов коллаборативной фильтрации параметр k=50. Значения показателей заметно различаются для двух наборов. Набор данных MSWeb можно назвать более «предсказуемым», поскольку он позволяет добиться более высоких показателей качества (значение F-меры достигает 27,47%). Набор данных НТБ ТПУ показывает более скромные результаты. Но при этом он характеризуется значительной разницей между рекомендациями по популярности и коллаборативной фильтрацией.

В таблице 4 приведено среднее время в секундах, потраченное на одну итерацию вычислений. Метод, основанный на документах, требует значительно больше времени для моделирования. Это связано с особенностями данного подхода, а также с реализацией данного алгоритма в библиотеке *recommenderlab*. Поскольку количество документов в нашем случае достаточно большое, матрица схожести имеет большую размерность, что затрудняет вычисления. При этом время формирования рекомендаций для пользователей значительно меньше, чем для альтернативного подхода.

Таблица 4. Среднее время на итерацию вычислений

Рекомендации	**Время, сек.**		
	моделирование	**формирование рекомендаций**	**Всего**
По популярности	0,004	3,59	3,594
На основе документов	2467,38	2,58	2469,96
На основе пользователей	0,007	59,36	59,367

Заключение. Проведенные эксперименты позволяют утверждать о возможности построения рекомендательной системы на основе методов коллаборативной фильтрации на основе данных о заказах НТБ ТПУ. Использование подхода, основанного на пользователях, позволило добиться более качественных рекомендаций по сравнению с

базовым методом: рекомендациями по популярности. Дальнейшую работу можно разделить на три направления. Во-первых, возможно улучшение качества рекомендаций на основе рассмотренных данных за счет применения гибридных методов построения рекомендаций. Во-вторых, необходимо провести оценку возможностей снижения времени моделирования и построения рекомендаций. И наконец, необходимо исследовать возможности сбора более качественной информации о предпочтениях пользователей, например, в виде рейтингов документов.

ЛИТЕРАТУРА

[1] *Карауш А.*С. Рекомендательные системы в публичных библиотеках // Библиосфера. 2009. № 1. С. 41-43.

[2] *Hahsler M.* recommenderlab: Lab for Developing and testing Recommender Algorithms. **https://cran.r-project.org/web/packages/recommenderlab/vignettes/recommenderlab.pdf** (дата обращения 20.09.2017).

[3] *Aggarwal C.* Recommender Systems: The Textbook. Switzerland: Springer International Publishing, 2016. 498 p.

[4] *Leskovec J., Rajaraman A., Ullman J.D.* Mining of Massive Datasets (2nd ed.). New-York: Cambridge University Press, 2014. 476 p.

[5] *Dice L.* Measures of the amount of ecologic association between species // Ecology. 1945. V 26 (3). P. 297–302.

[6] *Manning C.D.* Introduction to Information Retrieval. **http://www-nlp.stanford.edu/IR-book/** (дата обращения 20.09.2017).