УДК 026.06:004.62.032.2
DOI 10.25205/1818-7900-2018-16-2-62-69

А. А. Князева [1], **О. С. Колобов** [2], **И. Ю. Турчановский** [1], **А. М. Федотов** [1]

[1] *Институт вычислительных технологий СО РАН*
пр. Академика Лаврентьева, 6, Новосибирск, 630090, Россия

[2] *Институт сильноточной электроники СО РАН*
пр. Академический, 2/3, Томск, 634055, Россия

aknjazeva@ict.nsc.ru, okolobov@hcei.tsc.ru, tur@hcei.tsc.ru, fedotov@sbras.ru

КОЛЛАБОРАТИВНАЯ ФИЛЬТРАЦИЯ ДЛЯ ПОСТРОЕНИЯ РЕКОМЕНДАЦИЙ НА ОСНОВЕ ДАННЫХ О ЗАКАЗАХ *

Рассматривается возможность применения методов коллаборативной фильтрации в процессе создания рекомендательной системы на основе данных о заказах документов из библиотечного фонда. Приводится сравнительный экспериментальный анализ трех методов коллаборативной фильтрации: на основе документов, на основе пользователей и на основе гибридного метода, являющегося комбинацией первых двух методов.

Ключевые слова: рекомендательная система, коллаборативная фильтрация, унарные данные, бинарные данные.

Введение

Рекомендательные системы открывают новые возможности навигации в процессе информационного поиска. Очевидно, что одной из областей их применения могут быть библиотечные фонды [1]. Учет поведения пользователей для ранжирования документов, с которыми они взаимодействуют, ведет к установлению новых взаимосвязей между этими документами, выходящих за пределы традиционной рубрикации и ключевых слов. Такой учет позволяет связывать документы из смежных областей знания в условиях, когда в них используется различная терминология. В рамках данной работы рассматривалась возможность применения методов коллаборативной фильтрации для создания рекомендательной системы на основе данных о заказах документов в электронном каталоге Научно-технической библиотеки Томского политехнического университета (НТБ ТПУ). Задачи, поставленные в рамках исследования: 1) предварительная оценка качества рекомендаций на основе документов и на основе пользователей по сравнению с базовым методом без персонализации; 2) оценка качества работы гибридной рекомендательной системы, объединяющей описанные выше методы; 3) подбор некоторых параметров будущей системы.

Описание данных

В работе использовались данные о заказах читателей НТБ ТПУ за 2015 г., представленные в виде таблицы из двух столбцов: в первом столбце содержатся идентификаторы пользовате-

* Работа выполнена при частичной поддержке фонда РФФИ (проект № 18-07-01457).

Князева А. А., *Колобов О. С.*, *Турчановский И. Ю.*, *Федотов А. М.* Коллаборативная фильтрация для построения рекомендаций на основе данных о заказах // Вестн. НГУ. Серия: Информационные технологии. 2018. Т. 16, № 2. С. 62–69.

ISSN 1818-7900. Вестник НГУ. Серия: Информационные технологии. 2018. Том 16, № 2

лей, зашифрованные с помощью хеш-функции для обеспечения анонимности, а 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 времени заказа. Строго говоря, описанные данные являются унарными. Это означает, что мы знаем лишь о положительном отклике пользователя: факте заказа. При этом мы не знаем, насколько высоко пользователь оценил данный документ (рейтинги неизвестны), а также не обладаем сведениями об отрицательном отклике. Если пользователь не заказал конкретный документ, то причин может быть несколько:

- данный документ не является релевантным;
- документ релевантен, известен пользователю и, следовательно, не должен быть рекомендован;
- документ релевантен, 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. Технически это означает замену всех неопределенных значений на нули и переход от унарного типа данных к бинарному [3]. Если пользователь заказывал документ, то на пересечении соответствующих строки и столбца стоит единица, в противном случае – ноль. Такой подход позволяет формировать группу нерелевантных документов для проверки без оценки пользователем каждого документа в коллекции, что, как правило, 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, и также приведены к бинарному виду. Временной период: одна неделя в феврале 1998 г. 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-мя пользователями;
2) исключение пользователей, которые заказали менее 4-х документов.

Описанная фильтрация позволяет существенно сократить объем данных для работы (табл. 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 и иметь как минимум один документ для проверки.

Таблица 1

Количественное описание данных

Данные	До фильтрации		После фильтрации	
	НТБ	MSWeb	НТБ	MSWeb
Записи о заказах / просмотрах	98 341	98 653	51 513	57 497
Уникальные пользователи	9 619	32 710	4 786	9 544
Уникальные документы	37 718	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 следующие варианты 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*);

4) гибридный подход (*Hybrid*).

Первый способ, при котором всем пользователям рекомендуются наиболее популярные документы, был использован в качестве базового метода для сравнения. Рекомендации, полученные с его помощью, не являются персонализированными.

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 привлечь 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.

Гибридный метод подразумевает комбинацию двух списков рекомендаций с заданными весовыми коэффициентами. В данной работе комбинировались два варианта коллаборативной фильтрации: на основе документов и на основе пользователей.

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 три документа, на основе которых строились рекомендации. Размер списка рекомендаций описывается параметром N. 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» (рис. 1).

Как видно из рис. 1, коэффициент Жаккара и мера Дайса дают очень близкие результаты, тогда как коэффициент Пирсона им несколько проигрывает. Значения полноты и точности для косинусной меры настолько малы, что вся кривая выглядит как одна точка рядом с началом координат. Такая аномалия проявляется за счет особенностей реализации построения рекомендаций с помощью косинусной меры в библиотеке *recommenderlab*. Использование

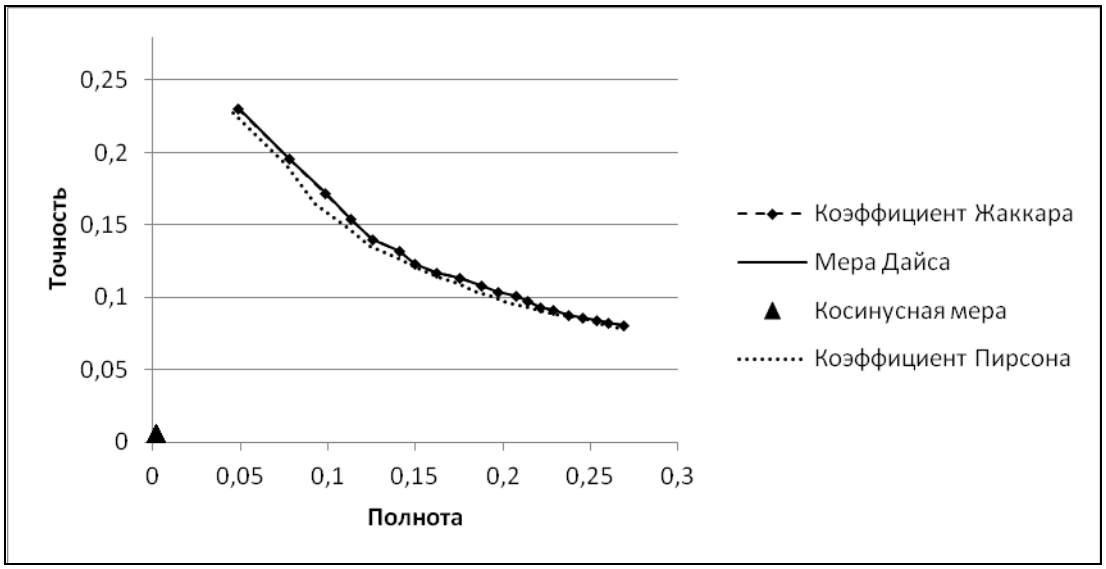

Рис. 1. Кривые «полнота-точность» для рекомендаций на основе документов (IBCF) в зависимости от меры схожести (параметр $k = 30$; количество рекомендаций N изменяется от 1 до 20)

данной меры схожести приводит к тому, что слишком многие документы получают максимально возможное значение меры схожести. В случае, когда для некоторого документа количество максимально схожих с ним документов больше значения параметра k, выбор k ближайших соседей становится проблематичным. Используемый инструментарий в этом случае возвращает пустое множество вместо списка рекомендаций. Таким образом, рекомендации на основе косинусной меры были сформированы менее чем для 4 % пользователей тестовой выборки. Для оставшихся пользователей были приняты нулевые значения показателей полноты и точности.

Коллаборативная фильтрация на основе пользователей

Для вычисления сходства между пользователями использовались уже перечисленные меры схожести (рис. 2).

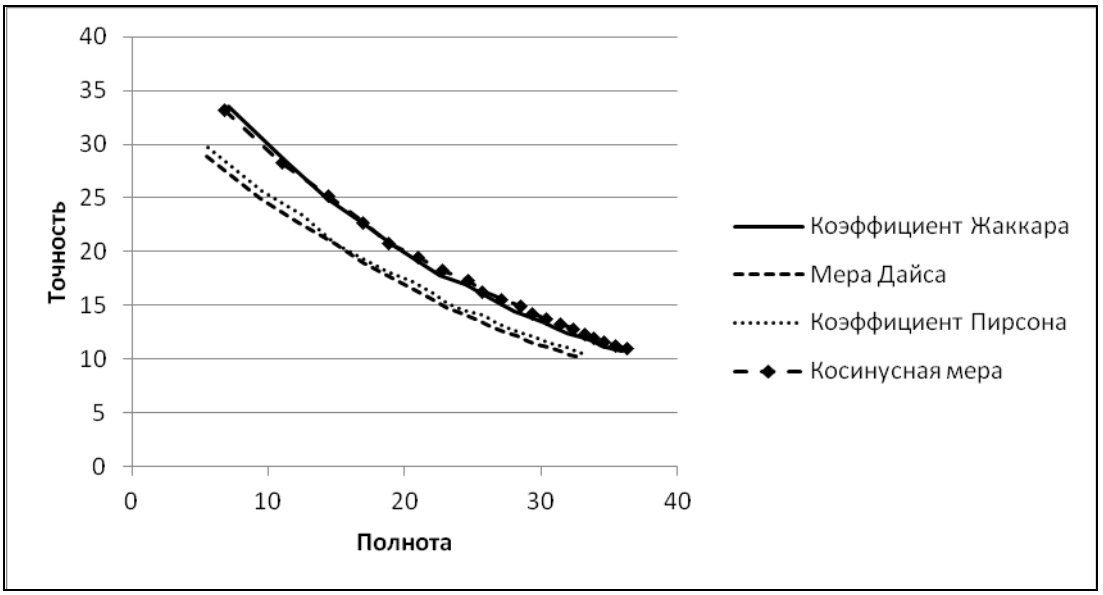

Рис. 2. Кривые «полнота-точность» для рекомендаций на основе пользователей (UBCF) в зависимости от меры схожести (параметр $k = 50$; количество рекомендаций N изменяется от 1 до 20)

Лучшие 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 гистограмме (рис. 3). Из рассмотренных значений параметров рекомендательной системы наиболее качественные рекомендации дает использование параметра $k = 50$ в сочетании с косинусной мерой. По результатам аналогичного анализа метода построения рекомендаций на основе документов параметр k был выбран равным 30. В табл. 2 приведены показатели качества для двух вариантов коллаборативной фильтрации.

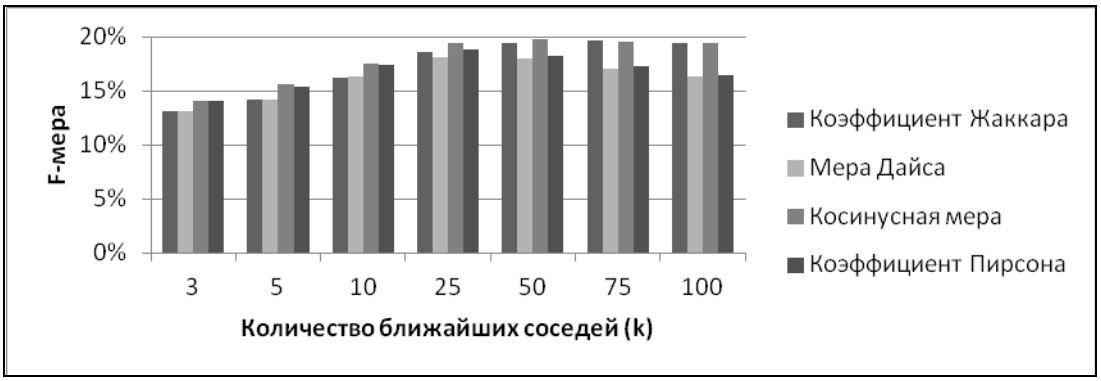

Рис. 3. F-мера для рекомендаций на основе пользователей (UBCF) в зависимости от меры схожести и значения параметра k ($N = 10$)

Таблица 2

Оценки качества (%) для списка из 10 рекомендаций

Мера сходства	Рекомендации					
	основанные на документах (IBCF, $k = 30$)			основанные на пользователях (UBCF, $k = 50$)		
	точность	полнота	F-мера	точность	полнота	*F*-мера
Жаккара	10,85	18,75	13,74	15,15	27,01	19,41
Пирсона	10,46	18,33	13,32	14,53	24,27	18,18
Косинусная	0,65	0,18	0,28	15,58	27,06	19,78
Дайса	10,84	18,73	13,73	14,26	24,28	17,97

В результате проведенных экспериментов для рекомендаций на основе документов был выбран коэффициент Жаккара, а для рекомендаций на основе пользователей – косинусная мера. При этом результаты метода на основе пользователей заметно превосходят качество рекомендаций на основе документов.

Гибридный метод

При создании гибридного подхода были использованы два метода: построение рекомендаций на основе пользователей с применением косинусной меры и на основе документов с использованием коэффициента Жаккара. Пропорция для комбинирования методов задавалась с помощью параметра a:

$$R_{hybrid} = aR_{UBCF} + (1-a)R_{IBCF}.$$

Зависимость *F*-меры от параметра a проиллюстрирована на рис. 4.

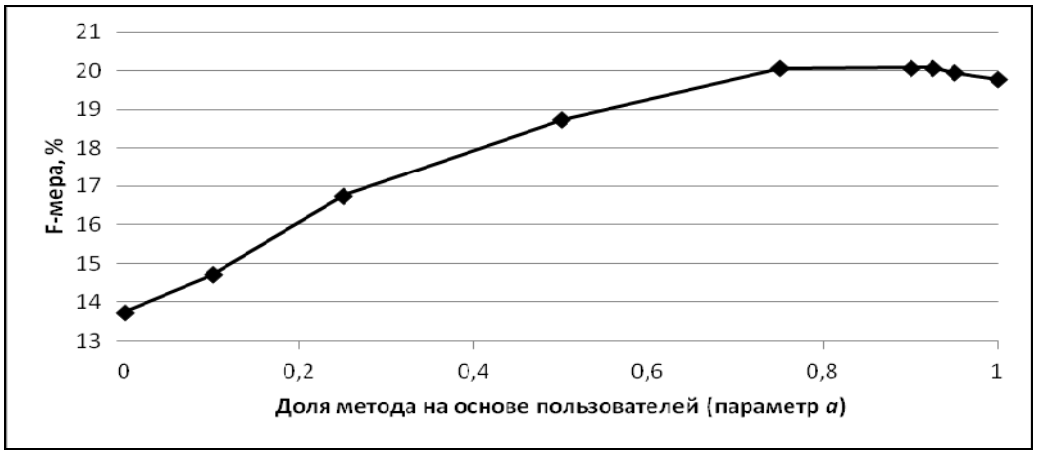

Рис. 4. F-мера для гибридных рекомендаций в зависимости от значений параметра *a* (количество рекомендаций равно 10)

Среди рассмотренных значений лучшим оказалось значение $a = 0{,}925$. При этом достигнутое значение *F*-меры превосходит значение, полученное методом на основе пользователей. Таким образом, комбинация превосходит по качеству каждый метод в отдельности.

Сравнение описанных методов

В табл. 3 приведены оценки качества рекомендаций рассмотренных выше подходов. Для всех методов параметр $N = 10$. 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. Выигрыш в *F*-мере для гибридного метода составляет всего 1,5 % от значения для метода на основе пользователей, в то же время он является значительно более трудоемким (табл. 4).

Таблица 3

Сравнение качества рекомендаций для описанных подходов ($N = 10$), %

Показатель качества	По популярности		На основе документов (коэф. Жаккара, $k = 30$)		На основе пользователей (косинусная мера, $k = 50$)		Гибридный метод ($a = 0{,}925$)	
	НТБ	MSWeb	НТБ	MSWeb	НТБ	MSWeb	НТБ	MSWeb
Точность	3,79	16,16	10,85	17,46	15,58	17,27	15,84	18,50
Полнота	4,78	57,63	18,75	64,37	27,06	65,19	27,43	68,60
F-мера	4,23	25,24	13,74	27,47	19,78	27,31	20,08	29,14

Таблица 4

Среднее время на итерацию вычислений

Рекомендации	Время, с		
	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
Гибридный метод	13529,17	848,61	14377,79

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 значительно, матрица схожести имеет большую размерность, что затрудняет вычисления. При этом время формирования рекомендаций для пользователей значительно меньше, чем для подхода на основе пользователей. Что касается гибридного метода, выигрыш в качестве, который он обеспечивает, вряд ли может компенсировать его временные затраты.

Заключение

Проведенные эксперименты позволяют утверждать о возможности построения рекомендательной системы методами 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, а также по сравнению с рекомендациями на основе документов. Гибридный подход с использованием двух методов коллаборативной фильтрации позволил несколько улучшить показатели качества, но при этом потребовал значительного времени как на этапе моделирования, так и на этапе формирования рекомендаций. В рамках дальнейшей работы планируется исследовать возможности сбора более качественной информации о предпочтениях пользователей (например, в виде рейтингов документов), а также оценить возможности привлечения информации из библиографических описаний документов, хранящихся в фонде библиотеки.

Список литературы

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

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

3. *Aggarwal C.* Recommender Systems: The Textbook. Springer International Publishing, Switzerland, 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. Vol. 26 (3). P. 297–302.

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

Материал поступил в редколлегию 17.03.2018

A. A. Knyazeva [1], **O. S. Kolobov** [2], **I. Yu. Turchanovsky** [1], **A. M. Fedotov** [1]

[1] *Institute of Computational Technologies SB RAS*
6 Academician Lavrentiev Ave., Novosibirsk, 630090, Russian Federation

[2] *Institute of High Current Electronics SB RAS*
2/3 Akademichesky Ave., Tomsk, 634055, Russian Federation

aknjazeva@ict.nsc.ru, okolobov@hcei.tsc.ru, tur@hcei.tsc.ru, fedotov@sbras.ru

COLLABORATIVE FILTERING FOR CREATION OF RECOMMENDATIONS ON BASE OF ORDER DATA

In the article an opportunity of the collaborative filtering methods application in a process of creating a recommender system on the base of order data of documents from library fund is consid-

ered. A comparison experimental analysis of three collaborative filtering methods is provided: item-based, user-based and hybrid method, which is a combination of first two methods.

Keywords: recommender system, collaborative filtering, unary data, binary data.

References

1. Karaush A. S. Recommender system in a public library. *Bibliosphere*, 2009, no. 1, p. 41–43. (in Russ.).

2. Hahsler M. Recommenderlab: A Framework for Developing and testing Recommender Algorithms. 2011. URL: https://cran.r-project.org/web/packages/recommenderlab/vignettes/recommenderlab.pdf (access: 19.09.2017).

3. Aggarwal C. Recommender Systems: The Textbook. Springer International Publishing, Switzerland, 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, vol. 26 (3), p. 297–302.

6. Manning C. D. Introduction to Information Retrieval. URL: http://www-nlp.stanford.edu/IR-book/ (access: 19.09.2017).

For citation:

Knyazeva A. A., Kolobov O. S., Turchanovsky I. Yu., Fedotov A. M. Collaborative Filtering for Creation of Recommendations on Base of Order Data. *Vestnik NSU. Series*: *Information Technologies*, 2018, vol. 16, no. 2, p. 62–69. (in Russ.)

DOI 10.25205/1818-7900-2018-16-2-62-69