

Подход к построению рекомендательной системы для библиотек

Ю. В. Леонова*, О. С. Колобов, А. А. Князева, И. Ю. Турчановский

ФГБНУ «Федеральный исследовательский центр информационных и вычислительных технологий», г. Новосибирск, Российская Федерация

Адрес: 630090, Российская Федерация, г. Новосибирск, пр. Академика Лаврентьева, д. 6

* ju1ia.ict@yandex.ru

Аннотация

В работе рассмотрена возможность построения гибридной рекомендательной системы для электронного каталога библиотеки в форме рекомендательного сервиса. Были рассмотрены следующие варианты построения рекомендаций: методы коллаборативной фильтрации, рекомендации на основе контента. Полученные результаты использовались при создании рекомендательного сервиса, в котором предлагается задействовать два рекомендательных алгоритма: коллаборативную фильтрацию на основе документов и рекомендации на основе контента. В качестве входных данных для рекомендательной системы используются анонимизированные данные о выполненных заказах (история заказов) и данные электронного каталога библиотеки. Рассмотрен пример построения рекомендательной системы на основе контента.

Ключевые слова: персонализированные библиотечные услуги, электронные каталоги библиотек, рекомендательные системы, методы коллаборативной фильтрации, рекомендации на основе контента

Конфликт интересов: авторы заявляют об отсутствии конфликта интересов.

Для цитирования: Подход к построению рекомендательной системы для библиотек / Ю. В. Леонова [и др.] // Современные информационные технологии и ИТ-образование. 2023. Т. 19, № 3. С. 676-683. <https://doi.org/10.25559/SITITO.019.202303.676-683>

© Леонова Ю. В., Колобов О. С., Князева А. А., Турчановский И. Ю., 2023



Контент доступен под лицензией Creative Commons Attribution 4.0 License.
The content is available under Creative Commons Attribution 4.0 License.



Approach to Building a Recommendation System for Libraries

Yu. V. Leonova*, O. S. Kolobov, A. A. Knyazeva, I. Yu. Turchanovskyi

Federal Research Center for Information and Computational Technologies, Novosibirsk, Russian Federation

Address: 6 Academician M.A. Lavrentiev Ave., Novosibirsk 630090, Russian Federation

* ju1ia.ict@yandex.ru

Abstract

The paper considers the possibility of building a hybrid recommendation system for the electronic library catalog in the form of a recommendation service. The following options for building recommendations were considered: collaborative filtering methods, content-based recommendations. The results obtained were used to create a recommendation service, in which it is proposed to use two recommendation algorithms: collaborative filtering based on documents and recommendations based on content. Anonymized data on completed orders (order history) and data from the library's electronic catalog are used as input data for the recommendation system. An example of building a recommendation system based on content is considered.

Keywords: personalized library services, electronic library catalogs, recommendation systems; collaborative filtering methods, content-based recommendations

Conflict of interests: The authors declare no conflict of interest.

For citation: Leonova Yu.V., Kolobov O.S., Knyazeva A.A., Turchanovskyi I.Yu. Approach to Building a Recommendation System for Libraries. *Modern Information Technologies and IT-Education*. 2023;19(3):676-683. <https://doi.org/10.25559/SITITO.019.202303.676-683>



Введение

Вопрос о создании персонализированных библиотечных услуг на основе адаптивного подхода, который позволяет учитывать предпочтения читателя с течением времени, является интересным для библиотечных специалистов. Рекомендательная система библиотеки может взять на себя рутинную часть работы в этом вопросе и продемонстрировать возможности современных технологий. Для рекомендательной системы библиотеки необходим расширенный подход к процессам сбора и обработки данных. Основная сложность, с которой сталкивается библиотека на начальном этапе, — это понимание какие данные нужны, как их получать и хранить, а также как обеспечить безопасность данных на всех этапах работы. Успешное решение этого вопроса позволит библиотеке сильно продвинуться на пути к рекомендательной системе. В данной работе мы предлагаем рассматривать рекомендательную систему в форме *рекомендательного сервиса*, который может работать наряду с существующей библиотечной системой и адаптивно

учитывать предпочтения читателя. Для этого в работе рассматриваются вопросы сбора и обработки входных данных, методы и стратегия работы рекомендательной системы, а также архитектура рекомендательного сервиса, пригодного для интеграции с автоматизированной библиотечной системой. Для удобства восприятия материала в данной работе будем рассматривать рабочий пример на основе современной библиотеки. Фонд библиотеки содержит более 2 миллионов экземпляров. В библиотеке имеется автоматизированная библиотечная система и электронный каталог, который содержит более 500 тысяч библиографических записей. В библиотеке зарегистрировано более 10 тысяч читателей, для которых библиотека в период с 2014 по 2021 г. сделала более 680 тысяч выдач. Учетные данные читателей надежно хранятся в системе и доступны только для авторизованных пользователей. Также нам потребуются некоторые дополнительные данные для проведения экспериментов. Сводное описание имеющихся исходных данных приведены в таблице 1.

Таблица 1. Исходные данные
Table 1. Source data

№	Название	Формат	Кол-во записей	Краткое описание
1.	Электронный каталог	RUSMARC	675 830	Библиографические записи на материалы библиотечного фонда
2.	Электронные библиотечные системы	RUSMARC, MARC21, ONIX	199 770	Библиографические записи лицензионных электронных ресурсов
3.	Коллекция научных публикаций по теме – электронная библиотека	ISIS	38 962	Записи из БД ВИНТИ
4.	История пользователей	CSV	680 000	Журнал выполненных заказов

Источник: здесь и далее в статье все таблицы и рисунки составлены авторами.
Source: Hereinafter in this article all tables and figures were made by the authors.

Краткий обзор

Обзоры, посвященные созданию рекомендательных систем на основе библиографических данных, можно найти в работах [1-3]. По данным [3] более половины таких рекомендательных систем применяют контентную фильтрацию [4-6]. Наша цель — исследовать рекомендательные системы на основе гибридного подхода [5, 7], при котором возможно применять не один, а несколько методов для получения рекомендаций. Исследование основано на данных, полученных из автоматизированной библиотечной системы, и включает в себя: историю читателя (информация о выполненных заказах) и библиографические описания материалов различных фондов библиотеки. В случае коллаборативной фильтрации мы имеем дело с унарной рейтинговой матрицей данных не явной обратной связи, но природа наших данных другая. В [8] авторы имели дело с предпочтениями в телепередачах и

имели данные о количестве и продолжительности просмотров. В нашем случае мы имеем данные, содержащие информацию только о факте выполнения удаленного заказа на литературу [9]. Для решения проблемы «холодного старта» могут применяться различные методы семантического анализа текстового описания документов. В настоящее время разработано несколько методов, выполняющих смысловой анализ текста в автоматическом режиме. Процесс автоматического определения темы в документах называется тематическим моделированием¹. Одним из наиболее эффективных методов тематического моделирования является метод LDA [10, 11]. Основные преимущества LDA — это быстрота, интуитивно понятная концепция, возможность работы с длинными документами. К основным недостаткам LDA можно отнести тонкую настройку моделей, необходимость привлечения эксперта для интерпретации тем.

¹ Воронцов К. В. Вероятностное тематическое моделирование [Электронный ресурс] // MachineLearning.ru, 2013. URL: <http://www.machinelearning.ru/wiki/images/2/22/Voron-2013-ptm.pdf> (дата обращения: 19.09.2023).



Описание подхода

Рекомендательная система (далее система) — это комплекс программ, который анализирует предпочтения читателя (далее пользователь) и пытается предсказать публикации, представляющие персональный интерес. Систему можно рассматривать как *архитектуру агента*, которая определяет некоторый план для включаемых в нее программных агентов: *агент наблюдения*, *агент агрегирования* и *агент рекомендаций* [12]. Агент наблюдения наблюдает за действиями пользователей в рамках его рабочей сессии с библиотечной системой и передает полученные сигналы агенту агрегирования. Агент агрегирования выполняет обработку входных данных на основе одного или нескольких алгоритмов, составляет списки рекомендаций, которые в свою очередь передаются агенту рекомендаций. Роль агента наблюдения играет библиотечная система. Библиотечная система обрабатывает все запросы на выдачу литературы, а также хранит историю выполненных заказов в административной базе данных. Выполненный заказ читателя, а точнее запись о нем в административной базе данных, можно рассматривать как *сигнал*, который после предварительной анонимизации данных передается агенту агрегирования. Важно отметить, что агентам агрегирования и рекомендаций нет необходимости использовать персональные данные пользователя в явном виде, для того чтобы создать список рекомендаций.

Получение рекомендаций на основе данных о выполненных заказах

Система рекомендует публикации на основе оценок самого пользователя, а также явных или неявных предпочтений других пользователей. Если пользователь интересовался книгой, то система предложит книги, которые смотрели другие пользователи, также интересовавшиеся данной книгой. Для получения списка рекомендаций можно применять алгоритм коллаборативной фильтрации [13] на основе метода kNN (k-ближайших соседей) [14]. В общем случае возможны два варианта применения алгоритма коллаборативной фильтрации: *на основе документов (IBCF)* и *на основе пользователей (UBCF)*. В обоих случаях количество ближайших соседей определяется с помощью параметра k, который может быть задан как параметр конфигурации [15]. Из нашего рабочего примера видно, что количество читателей библиотеки значительно меньше количества материалов библиотечного фонда, поэтому разумно применять вариант *коллаборативной фильтрации на основе документов*. В случае когда коллаборативная фильтрация не работает, можно применять рекомендации на основе контента.

Решение проблемы «холодного старта»

В основе системы лежит подход коллаборативной фильтрации на основе документов, который опирается на информацию о выполненных заказах. Однако при появлении новых пользователей или внесении новых документов возникает так называемая проблема *холодного старта*, которую можно разделить на «холодный старт» для пользователей (что рекомендовать новым пользователям) и «холодный старт» для документов

(кому рекомендовать добавленные в систему документы). Для решения данной проблемы предлагается применять подход, который ориентирован на получение рекомендаций на основе контента и не требует данных из истории пользователя.

Получение рекомендаций на основе контента

Для рекомендаций на основе контента применяется тематическое моделирование на основе LDA [11]. Схему работы LDA можно представить следующим образом:

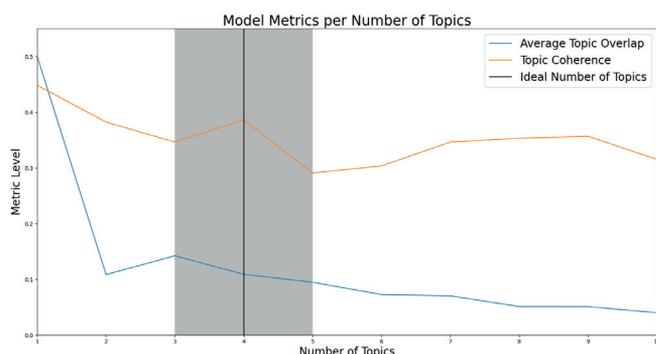
1. Подготовка данных. Слова в тексте проходят предварительную обработку, т. е. отбрасываются стоп-слова и проводится лемматизация. Далее составляется единый словарь слов и таблица встречаемости слов в текстах публикаций. Слова взвешиваются по TF-IDF [16, 17] и для слишком длинных текстов оставляется только топ N самых весомых слов.
2. Определение количества тем. Для применения LDA необходимо заранее определить количество тем, в которые будут группироваться данные. Выбор оптимального количества тем является сложной задачей и зависит от конкретных характеристик данных.
3. Обучение модели. Обработка текстов и определение LDA тематического распределения тем и слов в документах является итеративным процессом. Цель алгоритма — максимизировать вероятности наличия тем в каждом документе и появления слов в каждой теме.

Оптимальное количество тем

Для проведения тематического моделирования LDA необходимо задать оптимальное количество тем, наиболее точно отражающих содержание коллекции документов. Для этого используются различные методы. Один из наиболее распространенных подходов — это использование *перплексии* [18]. Перплексия является оценкой качества тематической модели и характеризует то, насколько хорошо модель предсказывает новые документы. Меньшее значение перплексии означает лучшую предсказательную способность модели. Поэтому можно вычислить перплексию для разного количества тем и выбрать количество тем с минимальной перплексией. К недостаткам перплексии можно отнести плохую интерпретируемость численных значений. Кроме того, перплексия зависит и от таких параметров: мощность словаря, длина документов. Другим подходом является использование *когерентности*² [19–22] тем, которая оценивает степень семантической связности слов внутри каждой темы. Чем более согласованы (когерентны) темы, тем лучше модель. Преимуществом когерентности по сравнению с другими методами оценивания тематических моделей является высокая корреляция с экспертными оценками. Также можно использовать методы визуализации для выбора оптимального количества тем. Для этого применяют метод называемый *метод локтя* [23], который заключается в графическом представлении значения метрики (когерентности или перплексии) в зависимости от количества тем. На графике обычно можно видеть «локоть» — точку, в которой значение метрики перестает значительно улучшаться (см. рис. 1). Такое количество тем можно считать оптимальным.

² Cutting J. Pragmatics and Discourse. London; N. Y.: Routledge, 2002. 187 p.





Р и с. 1. Оптимальное количество тем на основе когерентности

F i g. 1. Optimal number of topics based on coherence

Оценка качества рекомендаций и эксперименты

Оценка качества рекомендаций на основе данных о заказах
Для оценки качества рекомендаций был проведён ряд экспериментов на основе данных о заказах читателей в библиотеке за 2015 г. (см. табл. 2). В качестве инструмента использовалась библиотека программ recommenderlab [24], программный пакет slimrec [25] для платформы R. Для проведения экспериментов к данным о заказах были применены фильтры: (1) — исключение документов, которые были заказаны менее чем четырьмя пользователями; (2) — исключение пользователей, которые заказали менее четырех документов.

Т а б л и ц а 2. Подготовка данных о заказах читателей
библиотеки за 2015 г.

T a b l e 2. Preparation of data on library readers' orders for 2015

Наименование	До фильтрации	После фильтрации
Записей о заказах	98 341	51 513
Уникальных пользователей	9619	4786
Уникальных документов	37718	3764

Затем уникальные пользователи были случайным образом разбиты на пользователей обучающей и тестовой выборки в пропорции 0,7:0,3. В работе использовались два протокола проведения экспериментов, реализованных в recommenderlab [24]: 3-given, all-but-last. В рамках этого протокола для каждого пользователя из тестовой выборки было случайным образом выбрано 3 документа. Эти документы использовались в качестве базы для построения рекомендаций. А остальные «скрытые» документы были использованы для оценки качества. Чем больше «скрытых» документов совпало с рекомендованными, тем выше показатели качества. Этот протокол относится к так называемым офлайн-методам оценки качества. Они не лишены недостатков. В частности, все документы, не

попавшие в список «скрытых» документов, считаются нерелевантными, хотя в реальности это не обязательно так. Однако в условиях отсутствия возможностей явно оценить степень удовлетворенности пользователей офлайн-методы позволяют сравнивать различные алгоритмы между собой.

На основе протокола 3-given были вычислены следующие показатели качества:

- 1) точность — процент релевантных рекомендаций;
- 2) полнота — процент «угаданных» скрытых документов;
- 3) F-мера — комбинация показателей полноты и точности.

В таблице 3 приведены показатели качества рекомендаций, основанных на документах (kNN, количество ближайших соседей $k=30$).

Т а б л и ц а 3. Показатели качества рекомендаций (k=30)

T a b l e 3. Recommendation quality indicators (k=30)

Мера сходства	Точность, %	Полнота	F-мера, %
Жаккара	10,85	18,75	13,74
Пирсона	10,46	18,33	13,32
Косинусная мера	0,65	0,18	0,28
Дайса	10,84	18,73	13,73

Итак, для получения списка рекомендаций рекомендательная система может применять коллаборативную фильтрацию на основе метода kNN, а в том случае, если таких рекомендаций нет (например, система находится в состоянии «холодный старт»), может применяться алгоритм получения рекомендации на основе контента. Также необходимо отметить, что в первом случае рекомендации могут быть подготовлены в режиме офлайн, а во втором случае рекомендации могут быть получены только в режиме онлайн. Последнее связано с тем, что рекомендации на основе контента основаны на текущей сессии пользователя и на его истории выполненных заказов.

Эксперимент для получения рекомендаций на основе контента

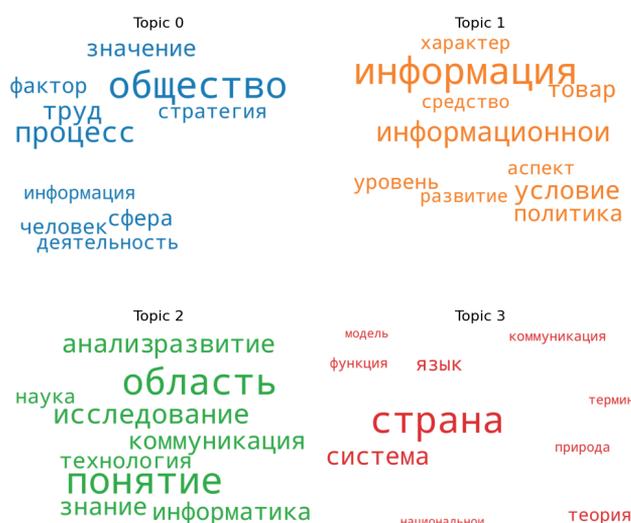
Для эксперимента был использован корпус научно-технических статей ВИНТИ по предметной рубрике «Электронные библиотеки». Всего было выбрано 38 962 документа на русском языке по трем направлениям исследований (20.01.01, 20.01.07 и 20.01.09) согласно рубрике ГРНТИ. Документы содержат следующую информацию: код (ГРНТИ), заглавие, аннотацию. Для построения тематической модели был разработан модуль на языке программирования Python с использованием библиотек NLTK³, spaCy⁴ и Gensim⁵. На рисунке 2 представлены 10 самых характерных терминов для каждой тематики. Из рисунка видно, что некоторым темам можно дать осмысленное название, например, тематика 0 — «Общество и деятельность». В теме 2 термины связаны между собой, но дать определенное название теме сложнее. Темы 1, 3 характеризуют термины, не позволяющие интерпретировать их однозначно.

³ Natural Language Toolkit [Электронный ресурс] // NLTK Project, 2023. URL: <https://www.nltk.org> (дата обращения: 19.09.2023).

⁴ Introducing spaCy [Электронный ресурс] // Explosion, 2023. URL: <https://explosion.ai/blog/introducing-spacy> (дата обращения: 19.09.2023).

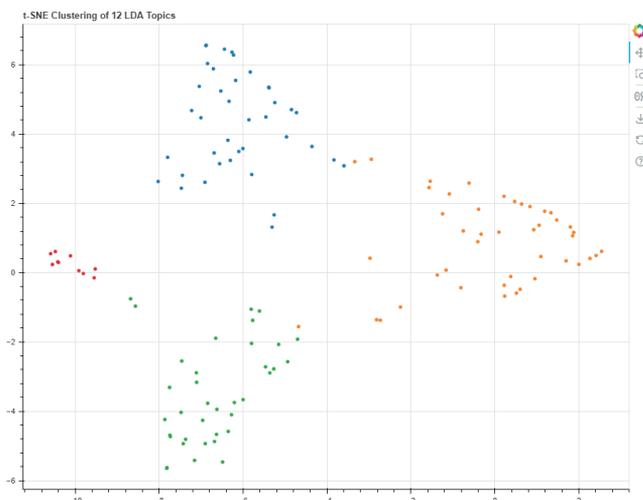
⁵ Gensim [Электронный ресурс] // Python Software Foundation, 2023. URL: <https://pypi.org/project/gensim> (дата обращения: 19.09.2023).





Р и с. 2. Пример выделения ключевых слов, описывающих тему
F i g. 2. An example of highlighting keywords describing a topic

Для получения визуального представления о виде плотного представления тематик была сделана проекция на двумерное пространство с помощью библиотеки Vokeh⁶. Полученное представление кластеров тематик приведено на рисунке 3.



Р и с. 3. Проекция плотного представления тематик с сохранением расстояний
F i g. 3. Projection of a dense representation of topics preserving distances

⁶ Дронов В. Библиотека Vokeh в Python [Электронный ресурс] // Обучение Python, 2023. URL: <https://tonais.ru/library/biblioteka-bokeh-v-python> (дата обращения: 19.09.2023).

⁷ Колобов О. Рекомендательный сервис для электронного каталога библиотеки [Электронный ресурс] // Bitbucket, 2023. URL: https://bitbucket.org/oleg_kolobov/Lf/src/master/ (дата обращения: 19.09.2023).

⁸ Run JavaScript Everywhere [Электронный ресурс] // OpenJS Foundation, 2023. URL: <https://nodejs.org/en> (дата обращения: 19.09.2023).

Реализация

Гибридный подход

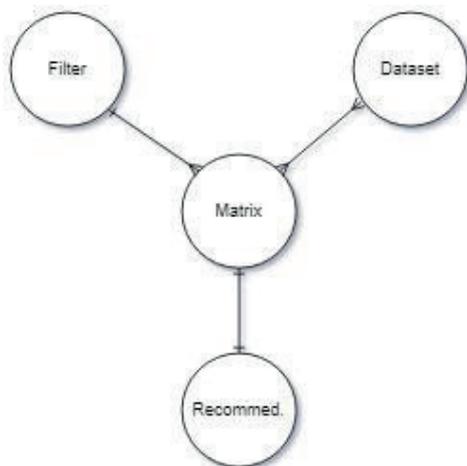
Предварительный анализ исходных данных показал, что процесс получения рекомендаций должен основываться на *гибридном подходе*, т.е. включать в себя комбинацию нескольких подходов (подход на основе контента применяется в том случае, когда не работают рекомендации на основе заказа). Реализация такого процесса является отдельной задачей, так как существует не менее шести вариантов для комбинирования различных подходов [1, 2]. В нашем случае применяется вариант *переключение*, т.е. для получения рекомендаций применяется один из двух подходов, в зависимости от некоторого условия.

Демонстрационная программа Lf

Для демонстрации подхода была создана программа Lf⁷. Эта программа рассматривается в контексте имеющегося программного окружения библиотеки. Такой подход позволяет повторно использовать имеющуюся инфраструктуру и соответствующий API (идентификация пользователя, авторизация запросов, личный кабинет читателя, поиск и извлечение документов в электронном каталоге и др.) для создания рекомендательного сервиса. И более того, собственное API рекомендательного сервиса в этом случае становится частью общего API. Программа является автономным сервером с поддержкой собственного API, который работает в экосистеме Node.js⁸. Входные данные могут быть загружены в сервер на основе соответствующего вызова API, в пакетном режиме (инкрементальное пополнение данных не поддерживается в данной версии программы). Для получения рекомендаций в программе применяется подход коллаборативной фильтрации на основе документов, а для оценки близости документов используется усредненно-вероятностная метрика.

В основе программы Lf лежит простая модель данных (см. рис. 4), включающая в себя основные объекты и их отношения. Эта модель необходима для поддержки процессов создания рекомендаций. Объект «*фильтр*» (Filter) содержит все необходимые параметры для коллаборативной фильтрации. Объект «*набор данных*» (Dataset) содержит входные данные для фильтрации. Объект «*матрица*» (Matrix) содержит связь с используемыми наборами данных и применяемым фильтром, а также текущее собственное состояние. Для объекта матрица определены две команды, которые меняют его состояние — «создать матрицу» или «опустошить матрицу». В результате успешного выполнения команды «создать матрицу» создается связанный объект «*рекомендации*» (Recommendations), и наоборот, в результате успешной команды «опустошить матрицу» удаляется связанный объект рекомендации. Объект рекомендации содержит идентификатор документа и список идентификаторов рекомендованных документов. Данная модель позволяет создавать на одном наборе данных различные варианты списков рекомендаций в зависимости от входных параметров фильтра.





Р и с. 4. Модель данных для программы Lf
F i g. 4. Data model for the Lf program

Заключение

В работе рассмотрен подход к созданию рекомендательной системы для библиотек в форме рекомендательного сервиса. Главным преимуществом такого подхода является возможность создавать новые современные приложения без изменения существующей библиотечной системы. В работе показано, что для получения списка рекомендаций можно применять коллаборативную фильтрацию на основе документов, с минимальным набором входных данных (история читателя о выполненных заказах). В качестве обоснования выбора приведен результат сравнения оценки качества рекомендаций. Также показано, что для решения проблемы «холодного старта» можно применять методы математического моделирования для получения рекомендаций на основе контента. В обоих случаях все необходимые данные потенциально присутствуют в библиотечной системе. В работе дано краткое описание рекомендательного сервиса Lf в форме API-сервера, который реализован в экосистеме Node.js. Для этого был применен сервис-ориентированный подход на основе шаблонов проектирования микросервисов.

References

- [1] Franke M., Geyer-Schulz A., Neumann A.W. Recommender Services in Scientific Digital Libraries. In: Tsihrintzis G.A., Jain L.C. (eds.) *Multimedia Services in Intelligent Environments. Studies in Computational Intelligence*. Vol. 120. Berlin, Heidelberg: Springer; 2008. p. 377-417. https://doi.org/10.1007/978-3-540-78502-6_15
- [2] Smeaton A.F., Callan J. Personalisation and recommender systems in digital libraries. *International Journal on Digital Libraries*. 2005;5(4):299-308. <https://doi.org/10.1007/s00799-004-0100-1>
- [3] Beel J., Gipp B., Langer S., Breiting C. Research-paper recommender systems: a literature survey. *International Journal on Digital Libraries*. 2016;17:305-338. <https://doi.org/10.1007/s00799-015-0156-0>
- [4] Risch J., Krestel R. What Should I Cite? Cross-Collection Reference Recommendation of Patents and Papers. In: Kamps J., Tsakonas G., Manolopoulos Y., Iliadis L., Karydis I. (eds.) *Research and Advanced Technology for Digital Libraries. TPD L 2017. Lecture Notes in Computer Science*. Vol. 10450. Cham: Springer; 2017. p. 40-46. https://doi.org/10.1007/978-3-319-67008-9_4
- [5] Tsuji K., Takizawa N., Sato S., Ikeuchi U., Ikeuchi A., Yoshikane F., Itsumura H. Book Recommendation Based on Library Loan Records and Bibliographic Information. *Procedia – Social and Behavioral Sciences*. 2014;147:478-486. <https://doi.org/10.1016/j.sbspro.2014.07.142>
- [6] Naak A., Hage H., Aïmeur E. A Multi-criteria Collaborative Filtering Approach for Research Paper Recommendation in Papyrus. In: Babin G., Kropf P., Weiss M. (eds.) *E-Technologies: Innovation in an Open World. MCETECH 2009. Lecture Notes in Business Information Processing*. Vol. 26. Berlin, Heidelberg: Springer; 2009. p. 25-39. https://doi.org/10.1007/978-3-642-01187-0_3
- [7] Fu S., Zhang Y., Seinmin N. On the recommender system for university library. In: *Proceedings of the IADIS International Conference e-Learning 2013*. Prague, Czech Republic; 2013. p. 215-222. Available at: <https://files.eric.ed.gov/fulltext/ED562315.pdf> (accessed 19.09.2023).
- [8] Hu Y., Koren Y., Volinsky C. Collaborative Filtering for Implicit Feedback Datasets. In: *2008 Eighth IEEE International Conference on Data Mining*. Pisa, Italy: IEEE Computer Society; 2008. p. 263-272. <https://doi.org/10.1109/ICDM.2008.22>
- [9] Pathan A.F., Prakash C. Unsupervised Aspect Extraction Algorithm for opinion mining using topic modeling. *Global Transitions Proceedings*. 2021;2(2):492-499. <https://doi.org/10.1016/j.gltp.2021.08.005>
- [10] Chauhan U., Shah A. Topic Modeling Using Latent Dirichlet allocation: A Survey. *ACM Computing Surveys*. 2021;54(7):145. <https://doi.org/10.1145/3462478>
- [11] Blei D.M., Ng A.Y., Jordan M.I. Latent Dirichlet allocation. *Journal of Machine Learning Research*. 2003;3(4-5):993-1022. Available at: <https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf> (accessed 19.09.2023).
- [12] Mönnich M., Spiering M. Adding Value to the Library Catalog by Implementing a Recommendation System. *D-Lib Magazine*. 2008;14(5/6). <https://doi.org/10.1045/may2008-monnich>
- [13] Fleder D., Hosanagar K. Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity. *Management Science*. 2009;55(5):697-712. <https://doi.org/10.1287/mnsc.1080.0974>
- [14] Piryonesi S.M., El-Diraby T.E. Role of Data Analytics in Infrastructure Asset Management: Overcoming Data Size and Quality Problems. *Journal of Transportation Engineering, Part B: Pavements*. 2020;146(2):04020022. <https://doi.org/10.1061/JPEODX.0000175>
- [15] Fareed A., Hassan S., Belhaouari S.B., Halim Z. A collaborative filtering recommendation framework utilizing social networks. *Machine Learning with Applications*. 2023;14:100495. <https://doi.org/10.1016/j.mlwa.2023.100495>



- [16] Jones K.S. A statistical interpretation of term specificity and its application in retrieval. *Journal of Documentation*. 2004;60(5):493-502. <https://doi.org/10.1108/eb026526>
- [17] Qaiser S., Ali R. Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *International Journal of Computer Applications*. 2018;181(1):25-29. Available at: <https://www.ijcaonline.org/archives/volume181/number1/qaiser-2018-ijca-917395.pdf> (accessed 19.09.2023).
- [18] Azzopardi L., Girolami M., van Risjbergen K. Investigating the relationship between language model perplexity and IR precision-recall measures. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR '03). New York, NY, USA: Association for Computing Machinery; 2003. p. 369-370. <https://doi.org/10.1145/860435.860505>
- [19] Rüdiger M., Antons D., Joshi A.M., Salge T.-O. Topic modeling revisited: New evidence on algorithm performance and quality metrics. *PLoS ONE*. 2022;17(4):e0266325. <https://doi.org/10.1371/journal.pone.0266325>
- [20] O'Callaghan D., Greene D., Carthy J., Cunningham P. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*. 2015;42(13):5645-5657. <https://doi.org/10.1016/j.eswa.2015.02.055>
- [21] Bicalho P., Pita M., Pedrosa G., Lacerda A., Pappa G.L. A general framework to expand short text for topic modeling. *Information Sciences*. 2017;393:66-81. <https://doi.org/10.1016/j.ins.2017.02.007>
- [22] Khamis I. The Use of Machine Learning in Libraries: How to Build a Book Recommender System. In: Holland B. (ed.) Handbook of Research on Advancements of Contactless Technology and Service Innovation in Library and Information Science. IGI Global; 2023. p. 23-44. <https://doi.org/10.4018/978-1-6684-7693-2.ch002>
- [23] Thorndike R.L. Who belongs in the family? *Psychometrika*. 1953;(18):267-276. <https://doi.org/10.1007/BF02289263>
- [24] Hahsler M. Recommenderlab: Lab for Developing and Testing Recommender Algorithms. *arXiv:2205.12371*. 2022. <https://doi.org/10.48550/arXiv.2205.12371>
- [25] Ning X., Karypis G. SLIM: Sparse Linear Methods for Top-N Recommender Systems. In: 2011 IEEE 11th International Conference on Data Mining. Vancouver, BC, Canada: IEEE Computer Society; 2011. p. 497-506. <https://doi.org/10.1109/ICDM.2011.134>

Поступила 19.09.2023; одобрена после рецензирования 06.10.2023; принята к публикации 10.10.2023.

Submitted 19.09.2023; approved after reviewing 06.10.2023; accepted for publication 10.10.2023.

Об авторах:

Леонова Юлия Викторовна, научный сотрудник, ФГБНУ «Федеральный исследовательский центр информационных и вычислительных технологий» (630090, Российская Федерация, г. Новосибирск, пр. Академика Лаврентьева, д. 6), кандидат технических наук, ORCID: <https://orcid.org/0009-0008-1706-5237>, ju1ia.ict@yandex.ru

Колобов Олег Сергеевич, научный сотрудник, ФГБНУ «Федеральный исследовательский центр информационных и вычислительных технологий» (630090, Российская Федерация, г. Новосибирск, пр. Академика Лаврентьева, д. 6), кандидат технических наук, ORCID: <https://orcid.org/0000-0002-7697-7681>, okolobov@gmail.com

Князева Анна Анатольевна, научный сотрудник, ФГБНУ «Федеральный исследовательский центр информационных и вычислительных технологий» (630090, Российская Федерация, г. Новосибирск, пр. Академика Лаврентьева, д. 6), кандидат технических наук, ORCID: <https://orcid.org/0000-0002-2538-7278>, aknyazeva22@gmail.com

Турчановский Игорь Юрьевич, заместитель директора – директор Томского филиала Института вычислительных технологий Сибирского отделения Российской академии наук, ФГБНУ «Федеральный исследовательский центр информационных и вычислительных технологий» (630090, Российская Федерация, г. Новосибирск, пр. Академика Лаврентьева, д. 6), кандидат физико-математических наук, ORCID: <https://orcid.org/0000-0003-4335-6085>, tur@hcei.tsc.ru

Все авторы прочитали и одобрили окончательный вариант рукописи.

About the authors:

Yulia V. Leonova, Researcher, Federal Research Center for Information and Computational Technologies (6 Academician M.A. Lavrentiev Ave., Novosibirsk 630090, Russian Federation), Cand. Sci. (Eng.), ORCID: <https://orcid.org/0009-0008-1706-5237>, ju1ia.ict@yandex.ru

Oleg S. Kolobov, Researcher, Federal Research Center for Information and Computational Technologies (6 Academician M.A. Lavrentiev Ave., Novosibirsk 630090, Russian Federation), Cand. Sci. (Eng.), ORCID: <https://orcid.org/0000-0002-7697-7681>, okolobov@gmail.com

Anna A. Knyazeva, Researcher, Federal Research Center for Information and Computational Technologies (6 Academician M.A. Lavrentiev Ave., Novosibirsk 630090, Russian Federation), Cand. Sci. (Eng.), ORCID: <https://orcid.org/0000-0002-2538-7278>, aknyazeva22@gmail.com

Igor Yu. Turchanovskiy, Deputy Director – Director of the Tomsk Branch of the Institute of Computational Technologies, Siberian Branch of the Russian Academy of Sciences, Federal Research Center for Information and Computational Technologies (6 Academician M.A. Lavrentiev Ave., Novosibirsk 630090, Russian Federation), Cand. Sci. (Phis.-Math.), ORCID: <https://orcid.org/0000-0003-4335-6085>, tur@hcei.tsc.ru

All authors have read and approved the final manuscript.

