|  |  |
| --- | --- |
| Gerb-BMSTU_01 | **Министерство науки и высшего образования Российской Федерации**  **Федеральное государственное бюджетное образовательное учреждение**  **высшего образования**  **«Московский государственный технический университет**  **имени Н.Э. Баумана**  **(национальный исследовательский университет)»**  **(МГТУ им. Н.Э. Баумана)** |

ФАКУЛЬТЕТ «Информатика и системы управления»\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

КАФЕДРА «Программное обеспечение ЭВМ и информационные технологии»\_\_\_\_\_\_\_\_\_\_\_\_\_\_

**Лабораторная работа №5**

**по курсу «Проектирование рекомендательных систем»**

|  |  |
| --- | --- |
| **Тема: «Гибридные рекомендательные системы»**  **Студент Горячев В. Г.**  **Группа ИУ7-33М**  **Преподаватель Быстрицкая А. Ю.** |  |

Москва

2023 г

## Описание алгоритмов

В данном случае гибридный алгоритм составлен из двух использовавшихся ранее – копирую их описания.

***Фильтрация по пользователям***

Алгоритм основан на использовании мер близости. В матрице оценок пользователями товаров ищутся пользователи, схожие друг с другом. Подразумевается, что людям со схожими покупками/вкусами можно предположить те объекты, которые один из похожих пользователей ещё не попробовал.

***Фильтрация по контенту c TF-IDF***

Алгоритм основан на анализе текста, переводящем естественный язык описания продукта в некое векторное представление, которое может быть сравнено любыми разумными в данной задаче способами. TF-IDF — статистическая мера, используемая для оценки важности слова в контексте документа, являющегося частью коллекции документов или корпуса. Вес некоторого слова пропорционален частоте употребления этого слова в документе и обратно пропорционален частоте употребления слова во всех документах коллекции.

***Объединение результатов***

Реализовано посредством простого смешивания результатов: коллаборативная фильтрация находит похожего пользователя, результат сокращается до 5 самых лучших (имеющих наивысшую оценку) непросмотренных целевым пользователем фильмов, после чего для каждого из этих 5 фильмов с помощью контентной фильтрации ищется по одному похожему фильму.

***Библиотеки***

В качестве языка программирования был выбран язык Python вместе с интерактивной средой Jupyter Notebook, поскольку они предоставляют удобный инструментарий для исследования, в частности, для выполнения лабораторных работ. Это определило выбор библиотек — нужно было найти совместимые с языком программирования.

В качестве источника алгоритмов использовались материалы из интернета для написания алгоритмов, подлежащих тестированию, поскольку не получилось найти библиотеки, предоставляющие их в готовом виде. Все реализации используют стандартные библиотеки – numpy, pandas и sklearn.

***Данные***

Для сравнения алгоритмов был использован учебный вариант набора данных MovieLens (<https://grouplens.org/datasets/movielens/>). Он содержит оценки пользователей по фильмам и специально составлен для подобного рода учебных задач.

***Сравнение***

В результате совмещения двух разнотипных алгоритмов пропала возможность явного прогнозирования оценки, которую оставит пользователь, а следовательно, и возможность оценки качества работы алгоритма с помощью метрик, сравнивающих правильный ответ с предсказанием модели. Кроме того, малый размеров текстов в наборе данных ограничивает возможность работы фильтрации по контенту. Но потенциально гибридизация улучшает качество рекомендаций, поскольку пользователь может получить как фильмы похожих на него пользователей, так и, как минимум, по два варианта похожих фильмов.

Из-за изменений в логике работы коллаборативной фильтрации (ограниченные предсказания на 5 фильмов) гибридный алгоритм стал работать даже чуть быстрее, чем одна из его составляющих, хотя и время на инициализацию выросло.

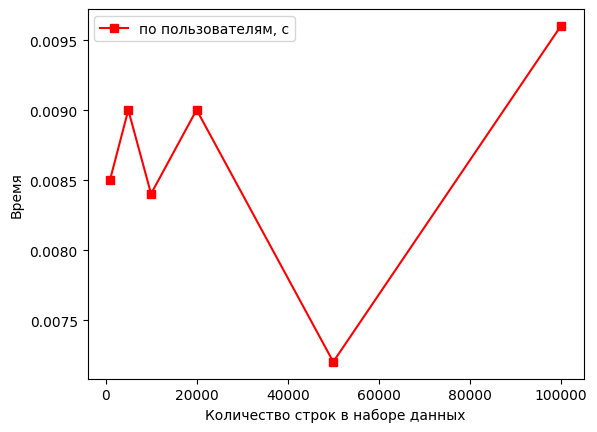


Рисунок 1 — Зависимость времени работы от размера набора у гибридного алгоритма

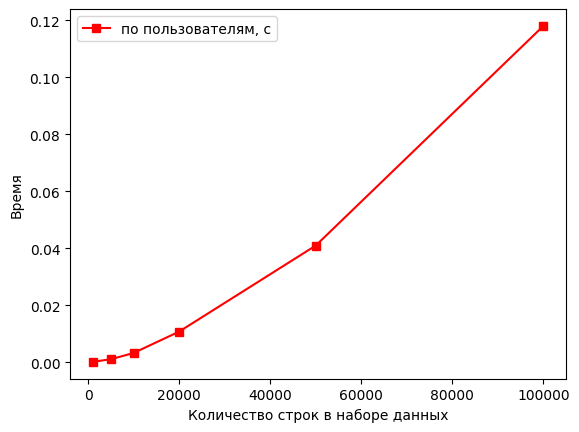


Рисунок 1 — Зависимость времени работы от размера набора у коллаборативной фильтрации по пользователям

**Вывод**

В ходе лабораторной работы был реализован гибридный алгоритм рекомендаций. Сам факт его реализации показывает жизнеспособность такого подхода, однако совмещение именно коллаборативной и контентной фильтрации может быть не всегда оправданным, поскольку эти алгоритмы лучше бы показали себя по-отдельности, но в одном продукте (например, если речь идёт о сайте с фильмами, то коллаборативная фильтрация подошла бы общей ленте рекомендаций, а контентная – когда пользователь открывает какой-либо фильм и видит конкретно к нему список похожих).