**Лабораторная работа №4. Коллаборативная фильтрация**

**ЗАДАНИЕ НА ЛАБОРАТОРНУЮ РАБОТУ**

1. **Таблица рейтингов**

Создать таблицу, где строки – пользователи (users), столбцы – объекты (items), которые оцениваются (фильмы, песни и т.п.). Значения ячеек таблицы – рейтинги пользователей, поставленные объектам. Шкала рейтингов выбирается от 1 до 5, где 5 – «отлично», 4 – «хорошо» и т.д. Размер таблицы должен быть не менее, чем 5х5. Кроме того, должны быть объекты без оценок.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | i1 | i2 | i3 | i4 | i5 | i6 |
| u1 | 2 | 4 |  | 5 |  |  |
| u2 | 3 | 4 |  |  | 5 | 1 |
| u3 |  |  | 5 | 5 | 5 |  |
| u4 |  | 4 |  | 4 | 5 | 4 |
| u5 | 2 |  | 5 | 5 | 2 |  |
| u6 | 2 | 4 |  | 5 | 4 |  |

Рис. 1 – Таблица рейтингов

1. **Коэффициент схожести**

Необходимо произвольно выбрать пользователя ***u***, у которого имеются не оцененные им объекты, и на основе метода коллаборативной фильтрации определить список рекомендуемых объектов для этого пользователя. Для этого предлагается вычислить коэффициент схожести. Это может быть коэффициент корреляции между оценками двух пользователей (*user-based collaborative filtering)* – *PC(u,v)*, коэффициент корреляции между оценками двух объектов (*item-based collaborative filtering*) – *PC(i,j)*, или нормированный по средней оценке пользователя косинусный коэффициент (*item-based collaborative filtering*) – *AC(i,j)*:

где – рейтинг объекта *i*, поставленный пользователем *u*; – рейтинг объекта *i*, поставленный пользователем *v*; , – средние рейтинги пользователей *u* и *v.*

где – рейтинг объекта *i*, поставленный пользователем *u*; – рейтинг объекта *j*, поставленный пользователем *u*; , – средние рейтинги объектов *i* и *j.*

где – рейтинг объекта *i*, поставленный пользователем *u*; – рейтинг объекта *j*, поставленный пользователем *u*; , – средний рейтинг пользователя *u*.

При вычислении коэффициентов следует учесть, что коэффициенты изменяются в пределах [-1;1], где значения близкие к 1 означают схожесть между выставленными рейтингами. Кроме того, при расчете средних рейтингов участвуют только выставленные оценки.

1. **Формирование списка рекомендаций**

Для того чтобы сформировать итоговый список рекомендаций, необходимо:

1. В случае *user-based collaborative filtering* определяется пользователь с наибольшей корреляцией оценок с выбранным пользователем ***u***, и имеющий оцененный объект ***i***, который не оценен у ***u*.** Если его оценка выше, например, 3, то такой объект может быть включен в итоговую выдачу рекомендованных к ознакомлению объектов.
2. В случае *item-based collaborative filtering* для всех неоцененных пользователем ***u*** объектов рассчитывается корреляция с уже оцененными, которая учитывает оценки всех пользователей (см. формулы 2 и 3). Далее из оцененных выбирается объект с наибольшей корреляцией с не оцененным, и в зависимости от оценки, которую ему поставил пользователь ***u***, принимается решение о включении в итоговую выдачу (по аналогии с пунктом 1).

Можно пойти и иным путем, и предсказать оценки которые может поставить пользователь ***u***:

1. В случае *user-based collaborative filtering* использовать общую формулу:

где – предсказываемый рейтинг для объекта *i*, поставленный пользователем ***u***; – *k* пользователей (с индексом *v*), оценивших объект *i*, и имеющих наибольшую корреляцию с пользователем ***u***; – корреляция между пользователями *u* и *v*.

Можно уточнить расчет, нормализовав шкалы оценок пользователей (*mean-centering*), т.к., например, у пользователя u3 (см. рис. 1) шкала состоит только из оценки «5».

1. В случае *item-based collaborative filtering*:

где – предсказываемый рейтинг для объекта *i*, поставленный пользователем ***u***; – *k* объектов (с индексом *j*), оцененных пользователем ***u*** и имеющих наибольшую корреляцию с объектом *i*; – корреляция между объектами *i* и *j*.

Формула для *mean-centering:*

**4. Варианты**

1 Вариант – реализовать *user-based collaborative filtering*

2 Вариант – реализовать *item-based collaborative filtering*