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 5x5. Кроме того, должны быть объекты без оценок.

	$\mathbf{i_1}$	$\mathbf{i_2}$	$\mathbf{i_3}$	$\mathbf{i_4}$	$\mathbf{i_5}$	$\mathbf{i_6}$
$\mathbf{u_1}$	2	4		5		
$\mathbf{u_2}$	3	4			5	1
$\mathbf{u_3}$			5	5	5	
$\mathbf{u_4}$		4		4	5	4
$\mathbf{u_5}$	2		5	5	2	
$\mathbf{u_6}$	2	4		5	4	

Рис. 1 – Таблица рейтингов

2. 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)*:

$$PC(u,v) = \frac{\sum_{i \in I_{u,v}} (r_{ui} - \overline{r_u})(r_{vi} - \overline{r_v})}{\sqrt{\sum_{i \in I_{u,v}} (r_{ui} - \overline{r_u})^2 \sum_{i \in I_{u,v}} (r_{vi} - \overline{r_v})^2}}, \qquad (1)$$

где r_{ui} – рейтинг объекта *i*, поставленный пользователем *u*; r_{vi} – рейтинг объекта *i*, поставленный пользователем *v*; $\overline{r_u}$, $\overline{r_v}$ – средние рейтинги пользователей *u* и *v*.

$$PC(i,j) = \frac{\sum_{u \in U_{i,j}} (r_{ui} - \overline{r_i})(r_{uj} - \overline{r_j})}{\sqrt{\sum_{u \in U_{i,j}} (r_{ui} - \overline{r_i})^2 \ \sum_{u \in U_{i,j}} (r_{uj} - \overline{r_j})^2}}, \qquad (2)$$

где r_{ui} – рейтинг объекта *i*, поставленный пользователем *u*; r_{uj} – рейтинг объекта *j*, поставленный пользователем *u*; $\overline{r_i}$, $\overline{r_j}$ – средние рейтинги объектов *i* и *j*.

$$AC(i,j) = \frac{\sum_{u \in U_{i,j}} (r_{ui} - \overline{r_u})(r_{uj} - \overline{r_u})}{\sqrt{\sum_{u \in U_{i,j}} (r_{ui} - \overline{r_u})^2 \ \sum_{u \in U_{i,j}} (r_{uj} - \overline{r_u})^2}}, \qquad (3)$$

где r_{ui} – рейтинг объекта *i*, поставленный пользователем *u*; r_{uj} – рейтинг объекта *j*, поставленный пользователем *u*; $\overline{r_u}$, – 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.

3. 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:

$$\hat{r}_{ui} = \frac{\sum_{v \in N_i(u)} w_{uv} r_{vi}}{\sum_{v \in N_i(u)} |w_{uv}|}, \qquad (4)$$

где $\hat{r}_{ui}$ – предсказываемый рейтинг для объекта i, поставленный пользователем ***u***; $N_i(u)$ – k пользователей (с индексом v), оценивших объект i, и имеющих наибольшую корреляцию с пользователем ***u***; w_{uv}– корреляция между пользователями u и v.

Можно уточнить расчет, нормализовав шкалы оценок пользователей (*mean-centering*), т.к., например, у пользователя u_3 (см. рис. 1) шкала состоит только из оценки «5».

$$\hat{r}_{ui} = \overline{r_u} + \frac{\sum_{v \in N_i(u)} w_{uv}(r_{vi} - \bar{r}_v)}{\sum_{v \in N_i(u)} |w_{uv}|}, \quad (5)$$

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

$$\hat{r}_{ui} = \frac{\sum_{j \in N_u(i)} w_{ij} r_{uj}}{\sum_{j \in N_u(i)} |w_{ij}|}, \qquad (6)$$

где $\hat{r}_{ui}$ – предсказываемый рейтинг для объекта *i*, поставленный пользователем ***u***; $N_u(i)$ – *k* объектов (с индексом *j*), оцененных пользователем ***u*** и имеющих наибольшую корреляцию с объектом *i*; w_{ij}– корреляция между объектами *i* и *j*.

Формула для *mean-centering:*

$$\hat{r}_{ui} = \overline{r_i} + \frac{\sum_{j \in N_u(i)} w_{ij}(r_{uj} - \bar{r}_j)}{\sum_{j \in N_u(i)} |w_{ij}|}, \quad (7)$$

4. Варианты

1 Вариант – реализовать *user-based collaborative filtering*

2 Вариант – реализовать *item-based collaborative filtering*