

TRƯỜNG ĐẠI HỌC SƯ PHẠM KỸ THUẬT TP. HỒ CHÍ MINH
KHOA ĐÀO TẠO CHẤT LƯỢNG CAO
NGÀNH CÔNG NGHỆ THÔNG TIN



BÁO CÁO TIẾN ĐỘ
MÔN: ĐỒ ÁN 3
ĐỀ TÀI: TÌM HIỂU VỀ THUẬT TOÁN
RECOMMENDATION

GVHD : Thầy Huỳnh Xuân Phụng

SVTH :

Nguyễn Thành Như

17110202

Võ Ngọc Thuận

17110234

TP. Hồ Chí Minh, tháng 12 năm 2020

I. Báo cáo đánh giá hiệu quả của thuật toán collaborative filtering

Với thuật toán Collaborative Filtering chúng ta sử dụng similarity hay “độ tương đồng” để đưa ra gợi ý.

Với user-based Collaborative Filtering thì độ tương có thể tính bằng phương pháp đo lường cosine, Pearson, Euclidean, ...

Similarity Measures	Computational formulae's
Pearson Correlation(PCC)	$\text{sim}(u, u')^{\text{PCC}} = \frac{\sum_{i \in I} (r_{u,i} - \bar{r}_u) \cdot (r_{u',i} - \bar{r}_{u'})}{\sqrt{\sum_{i \in I} (r_{u,i} - \bar{r}_u)^2} \cdot \sqrt{\sum_{i \in I} (r_{u',i} - \bar{r}_{u'})^2}}$ <p>Where I is the set of items, $r_{u,i}$ rating of i given to item i by user u, \bar{r}_u average rating of user u</p>
Cosine (COS)	$\text{sim}(u, u')^{\text{COS}} = \frac{\sum_{i \in I} (r_{u,i}) \cdot (r_{u',i})}{\sqrt{\sum_{i \in I} (r_{u,i})^2} \cdot \sqrt{\sum_{i \in I} (r_{u',i})^2}}$
Adjusted Cosine (ACOS)	$\text{sim}(i, i')^{\text{ACOS}} = \frac{\sum_{u \in U} (r_{u,i} - \bar{r}_i) \cdot (r_{u,i'} - \bar{r}_{i'})}{\sqrt{\sum_{u \in U} (r_{u,i} - \bar{r}_i)^2} \cdot \sqrt{\sum_{u \in U} (r_{u,i'} - \bar{r}_{i'})^2}}$ <p>Where U is the set of users rated both items i.e. i and i'</p>
Spearman's Rank Correlation similarity	$\text{sim}(u, u')^{\text{SRCC}} = 1 - \frac{\sum_{i \in I} (\text{rank}(r_{u,i}) - \text{rank}(r_{u',i}))^2}{ I \cdot (I - 1)}$ <p>Where I is the cardinality of co-rated items.</p>

Hình 1. Trích - A New Similarity Measure Based on Mean Measure of Divergence for Collaborative Filtering in Sparse Environment

Hiệu quả và độ chính xác của các dự đoán được xác định dựa trên tham số đó chính là RMSE (Root Mean Squared Error) hoặc MAE (Mean Absolute Error).
MAE sẽ được tính bằng công thức sau:

$$MAE = \frac{1}{|S|} \sum_{i=1}^S |Pred_i - r_i|$$

RMSE sẽ được tính bằng công thức sau:

$$RMSE = \sqrt{\frac{1}{|S|} \sum_{i=1}^S (Pred_i - r_i)^2}$$

$Pred_i$ là dự đoán điểm rating cho đối tượng i , r_i là rating thực tế của đối tượng i và $|S|$ là bản chất của test ratings.

RMSE là một số thực không âm sẽ có giá trị từ 0.0 và 0.0 là trường hợp tốt nhất khi mà dự đoán không có sai lệch.

Để thực hiện tính toán 2 tham số này trong python chúng ta có thể sử dụng 2 thư viện là **from sklearn.metrics import mean_absolute_error** để tính MAE và **from sklearn.metrics import mean_squared_error** và set `squared = False` để tính RMSE

Bằng cách áp dụng vào bài demo chúng ta thu được kết quả như sau:

```
[('0002005018', 5)]
Dự đoán
-----
0002005018    2.500000
1881320189    1.470814
0452264464    1.072051
0671870432    0.981014
1558746218    0.846022
dtype: float64
RMSE: 0.9696363370563034
MSE: 0.9062682574444774
```

Với mỗi rating được dự đoán sẽ sai lệch khoảng 0.96 nếu người dùng chỉ đánh giá một cuốn sách điều này cho chúng ta thấy hiệu suất ở lần dự đoán này là thấp.
Tương tự nếu chúng ta thực hiện dự đoán tiếp:

```
[('0002005018', 5), ('0060973129', 5), ('0061076031', 2)]
```

Dự đoán

1881320189	2.885838
------------	----------

0060973129	2.456209
------------	----------

0002005018	2.445146
------------	----------

0393045218	1.978672
------------	----------

0609804618	1.971701
------------	----------

dtype: float64

RMSE: 0.5553340533523824

MSE: 0.4550762929914415

Với mỗi rating được dự đoán sẽ sai lệch khoảng 0.55, đây là một con số tốt đối với giải thuật Collaborative Filtering cho thấy hiệu suất của giải thuật là tương đối cao.

Kết luận: Collaborative Filtering là một giải thuật tốt trong một hệ thống recommend vừa và nhỏ.