CS5344 Big Data Analysis -- Lab 2 Report

Zuo Shuman (E0336049)

Luo Yun (E0267766)

**Dataset**

“Musical Instrument” category has been choosen from <http://jmcauley.ucsd.edu/data/amazon/links.html>.

Frequently browsed together by the customer algorithm (FP Growth):

“metadata” of the category has been used for training, which is a json file of 102MB.

Collaborative Filtering algorithm:

“ratings only” data has been used, which is a csv file of 20.4MB.

**Visualise Dataset**

1. FP growth algorithm

The following SQL queries have been called to visualize the dataset:

* SELECT COUNT(\*) as total\_count FROM musical
* SELECT COUNT(\*) as null\_count FROM musical WHERE related IS NULL OR related.bought\_together IS NULL

It has been observed that asin, related contains null data. Those data has been cleared later.

|  |  |
| --- | --- |
| # of metadata | # of empty data |
| 84901 | 58954 |

2. Collaborative Filtering algorithm

The following SQL queries have been called to visualize the dataset:

* SELECT COUNT(\*) FROM useritem
* SELECT COUNT(DISTINCT itemId) FROM useritem
* SELECT COUNT(DISTINCT userId) FROM useritem
* SELECT COUNT(\*) FROM (SELECT DISTINCT COUNT(\*) AS count, userId FROM useritem GROUP BY userId HAVING count == 1)

|  |  |  |  |
| --- | --- | --- | --- |
| # of ratings | # of products | # of users | # of user with 1 rating |
| 500176 | 83046 | 339231 | 270914 |

It has been observed that the rating matrix is very sparse. More than ⅔ of the ratings are from the user with one rating and ¾ of the users only rate one item.

**Preprocess**

1. FP growth algorithm

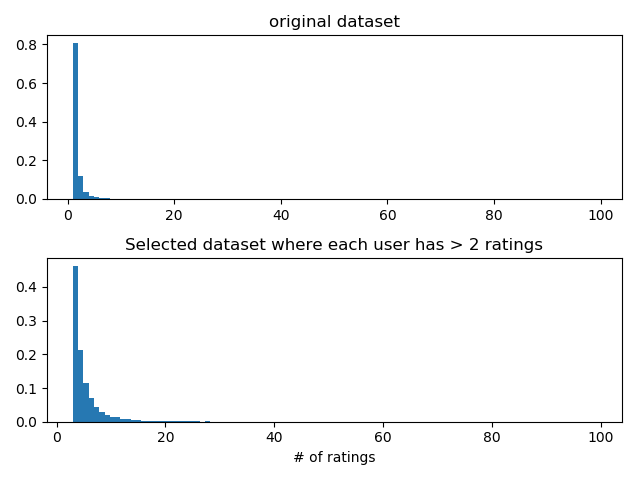
The dataset has been preprocessed as below:

* For the training dataset (metadata of musical instrument)
  + For those data rows which has null for asin, related and bought\_together has been removed.
  + The bought\_together has been set as >= 1, which has been tested to show good result. Bought\_together has been used as we believe it has most large association with asin. As only 1 to 5 items to be recommended, bought\_together contains enough data.
  + Asin and bought\_together has been grouped into one column as “bought”.

* For validation dataset (ratings of musical instrument)
  + A dictionary has been generated with user id as key and items bought as a whole as value. The dictionary is used later to validate if recommended item is efficient.
* For testing dataset (10% of ratings of musical instrument)
  + A list has been produced with (item, user). The testing dataset is the same as the one used in Collaborative Filtering algorithm for better comparison.

2. Collaborative Filtering algorithm

The dataset has been preprocessed, users with ratings < 3 has been removed. It could be observed that after removal, data become more balanced. The total data count become 150708.



After filtering, test data has been selected from filtered data frame in the two following ways:

1. Random split 10% ratings from the total dataset as test set
2. Select the top 10% ratings from each user(of he has > 10 ratings) sorted by timestamp as test set. This is to simulate process to recommend products to a user based on the history bought items

**Result**

1. FP Growth algorithm

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *K* | *1* | *2* | *3* | *4* | *5* |
| Conversion Rate (random) | 0.040139 | 0.0617416 | 0.0684091 | 0.0725430 | 0.0746766 |
| Convestion Rate (top 10%) | 0.040853 | 0.0624812 | 0.07089215 | 0.075097626 | 0.0772003 |

2. Collaborative Filtering algorithm

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| *K* | *1* | *2* | *3* | *4* | *5* |
| Conversion Rate (random) | 0.0004565 | 0.0007304 | 0.0008217 | 0.0009131 | 0.0010044 |
| Conversion Rate (top 10%) | 0.0 | 0.0001329 | 0.0002659 | 0.0003988 | 0.0007977 |

From the result above, a few points have been found:

* The Conversion Rate (CR) grows as more items have been recommended to the user. Moreover, the improvement is larger when K is small. As FP Growth as an example, when K increases from 2 to 1, CR increase 0.02 while CR increases 0.007 when K increases from 3 to 2.
* The CR of FP Growth is bigger than Collaborative Filtering when same K has been applied.
  + One of affecting factor for this might be that full metadata set has been used to train FP Growth while 90% rating data set has been used to train Collaborative Filtering, which means, there is overlap between test data and metadata when training FP Growth
  + Another reason might be, we are using ‘bought\_together’ metadata, since it will best capture which products are actually bought together by users .
* Collaborative Filtering performs better when the dataset has been splitted randomly.
  + One of the possible reason is that the
* FP Grouth performs better when dataset is chosen for the top 10%.

But in other opinion, in most of the cases, association mining should NOT perform better than Collaborative Filtering. It is less personalized, it does not matter who you are (your past history), what counts is what you are doing now.