Big Data Final Project Checkpoint Report

TIANYU DU and NUO LEI*, New York University, Center for Data Science

Additional Key Words and Phrases: Big Data, Recommendation System, Spark

1 PRE-PROCESSING

In the pre-processing part, we mainly focus on checking for null values and duplicated values. A null value may represent a mistake in data input, and a duplicated value may represent a duplicated data record in the data collection.

After careful checking, we found that there is no missing values in the table. But still, there are 6779 duplicated rows in the interaction table, which probably shows that we have duplicated listening records (same user listening to same recording at the same time). So we drop all these duplicated rows to prevent possibly corrupted interaction data.

2 TRAIN TEST PARTITION

we partitioned the `interaction_train_small` table into a training table and a validation table. The data for each user (represented by `user_id`) is sequentially ordered by timestamp and partitioned to the training table up to the first 99%, while the remaining 1% is assigned to the validation table (a 99:1 train-validation ratio). Our first attempt was to partition with ratio 70 30, but the resulting validation table had too many rows and our popularity baseline algorithm ran out of memory (our spark job was killed). Then we decided to use 99:1 training-validation ratio and were able to run the spark job.

3 BASELINE POPULARITY MODEL

We implemented the popularity baseline by finding the top 100 played music (represented by `recording_msid`) in the training table, and used metrics mAP @ 100 (mean average precision at 100) and ndcg @ 100 (normalized discounted cumulative gain at 100) to evaluate the baseline on the train and validation data. Since we are calculating the top 100 over all the interactions in the training, so we don't need other hyperparameters such as the damping factor. But of course in the next coming part, we are going to try different hyperparameters and find the best one by cross-validation on the validation set.

4 PERFORMANCE

The result is at below:

Train mAP: 0.00009019, Train ndcg: 0.018460609

Validation mAP: 0.00011476, Validation ndcg: 0.00129512

Authors' address: Tianyu Du, sebastiandu2000@gmail.com; Nuo Lei, nuo.lei@nyu.edu, New York University, Center for Data Science.

Manuscript submitted to ACM