PARADE

This is a recommendation system using large-scale collaborative filtering. Key modules involved are user clustering, anti-spam, long tail recommendation, ranking strategy etc.
System complexity is carefully considered to satisfy real-time requirement.

System overview

- User Clustering. Since the computation of similarity between users is not applicable, a user clustering algorithm should be implemented for collaborating filtering. For generalization, the distance between clusters should be quantifiable in a certain way.
- Voted Targeting Items. Targeting items are the spam-free items for recommendation. The voting behaviors from user to item should be tracked and quantized without cheating.
- Long Tail Solution. Different methods for low-frequent user recommendation could be used to guarantee the recall.
- Ranking Strategy. More sophisticated off-line feature of user and item could be used in ranking stage.

User Clustering

A simple clustering strategy is now used based on user’ s interest and Euclidean distance is used to estimate the similarity between clusters. Other algorithms like Min-Hash are to be forked in no time.

Anti-spam, Anti-cheating and Voting

- Anti-spam strategy is not revealed to me and I just use the result.
- Anti-cheating strategy is simple. We just use the data with reasonable amount from highly-trusted user.
- Voting is just a simple model using linear regression with experienced arbitrary weight of each type of behavior.

Long Tail Solution

We use something like Gaussian Mixture Model to handle long tail problem.
Traditionally, the items we push to user is taken from the candidate pool of the exact cluster he belongs to. Once a user belongs to a small cluster with few other users, the items in this candidate pool would be few or even in bad quality. To solve this problem we try to take items from the neighbour cluster as well.
When we import an item from the candidate pool of neighbour clusters, the voting weight of this item should be modified (decrease). Assuming item i originally in cluster s is imported to cluster t, the modified voting weight goes to

$$w_{it} = \sum_{s} w_{is} \times \alpha_{st} exp(-\beta \cdot d_{st})$$

- d_{st} is the distance between s and t
- β controls the decay due to distance
- α_{st} is called invasion coefficient which means the invation power that cluster s has on t. A cluster with large size actually does not need the items imported from other clusters while a small cluster needs to refer much to its neighbours so α_{st} could be defined as $\alpha_{st} = \frac{size(s)}{size(t)}$

Generally, one cluster should has distance value with all the others but in practice we do not record all the distance between pairs. When the distance is large enough, we view it as infinite to make the distance matrix a sparse one.

Ranking Strategy

There’ s no ranking strategy for now and the mixed voting result is directly used. There could be a logistic regression or other popular methods on sophisticated feature space.

Real time implementation suggestion

To be continued