Machine Discovering Assignment 2

Tung-Chun, Chiang R05922027,
Yen-Chen, Fu R05922006,
Han-Hao, Chen R05922021

November 20, 2016

1 Problem description

- Given:
  - Aggregative statistics for target link data
  - Attributes of nodes and types of links
- Goal:
  - Existence of user-item links

2 Methods

2.1 Graphical Model

2.1.1 Probability Inference

We regard that there are three possible paths in user-item links: user-friend-item(U-F-I), owner-item(O-I), user-category-item(U-C-I), which are shown in Figure 1. And the following is the derivations:

- U-F-I:

$$\begin{aligned} &P(user, item) = P(user) \cdot P(item|user) \\ &= P(user) \cdot \sum_{friend_{user}} \sum_{item_{friend}} P(item|item_{friend}) * P(item_{friend}|friend) * P(friend|user) \end{aligned}$$

- O-I:

$$\begin{aligned} &P(user, item) = P(user) \cdot P(item|user) \\ &= P(user) \cdot \sum_{item_{user}} P(item|item_{user}) \cdot P(item_{user}|user) \end{aligned}$$

- U-C-I:

$$\begin{aligned} P(user, item) &= P(user) \cdot P(item|user) \\ &= P(user) \cdot \sum_{c}^{|C|} P(item|c) \cdot P(c|user) \\ &= P(user) \cdot \sum_{c}^{|C|} \frac{P(item, c)}{P(c)} \cdot \frac{P(c, user)}{P(user)} \\ &= \sum_{c}^{|C|} \frac{P(item, c)}{P(c)} \cdot P(c, user) \end{aligned}$$

- $item_{user}$: the item owned by $user$
- $friend_{user}$: the user that is $user$'s friend
- $|C|$: number of categories

2.1.2 Mixed Model

In this task, we use two different methods to mix probabilities of the above three paths

- max path: We assume that the user-item link depends on only one of the three paths.
  $P(user, item) = max(P_{U-F-I}, P_{O-I}, P_{U-C-I})$
- weighted path: We assume that the user-item link comes from a mixed probability of the three paths.
  $P(user, item) = w_0 \cdot P_{U-F-I} + w_1 \cdot P_{O-I} + w_2 \cdot P_{U-C-I}$
  - $w_0 : \frac{(\frac{T}{|V|})^{\alpha}}{Z}$
  - $w_1 : \frac{(\frac{T}{|I|})^{\alpha}}{Z}$
  - $w_2 : \frac{(\frac{T}{|C|})^{\alpha}}{Z}$

  α is the smoothing parameter, and Z is the normalization factor.
  We think that the less of the number of users means the data of users is dense, it might be more informative in users' relationship, so it will has higher w_0. And so do w_1, w_2.

To generate a robust model, we combine three models to capture different information attributes, which are modeled by $User-Friend-Item$, $Owner-Item$, and $User-Category-Item$ probability paths respectively. We use linear combination to combine three models and given different weights according to the input data feature. Figure 1 shows our models.

2.1.3 Dataset Analysis

We analyze three given data set to see if they have some distribution. In user-category analysis, we found that the number of categories owned by some user have long-tail attribute. It means that most users own only one category. So using U-C-I model to model valid dataset may be an efficient way. Figure 2 shows the distribution of user-categories.

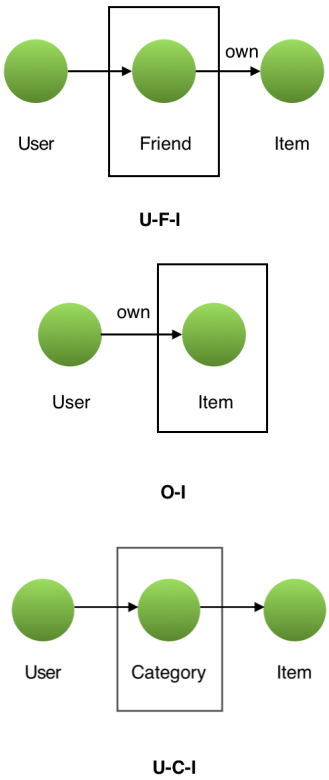

Figure 1: Graphical Model

	$\|U\|$	$\|I\|$	$\|C\|$	w_0	w_1	w_2
valid	105912	77792	100	0.0009	0.0012	0.9977
test1	71634	180684	16947	0.1778	0.0705	0.7516
test2	199377	246407	100	0.0005	0.0004	0.9990

Table 1: Weights in different datasets with $\alpha = 0.8$

The weights of mixed path in different datasets are shown in Table 1. The number of categories in valid and test2 are only 100, so the information strongly focuses on those 100 categories and w_2 should be very large. In fact, we use path U-C-I as score and get accuracy 0.9787 in valid, but it is not general enough.

2.2 Social Network Embeddings

We assume that users will have similar interest with their friends, and users will have interest with items in the categories which they own items there. So we propagate scores from item to its owner and the owner's friends, and also propagate scores to the owners in the same category with this item. Figure 3 shows our model.

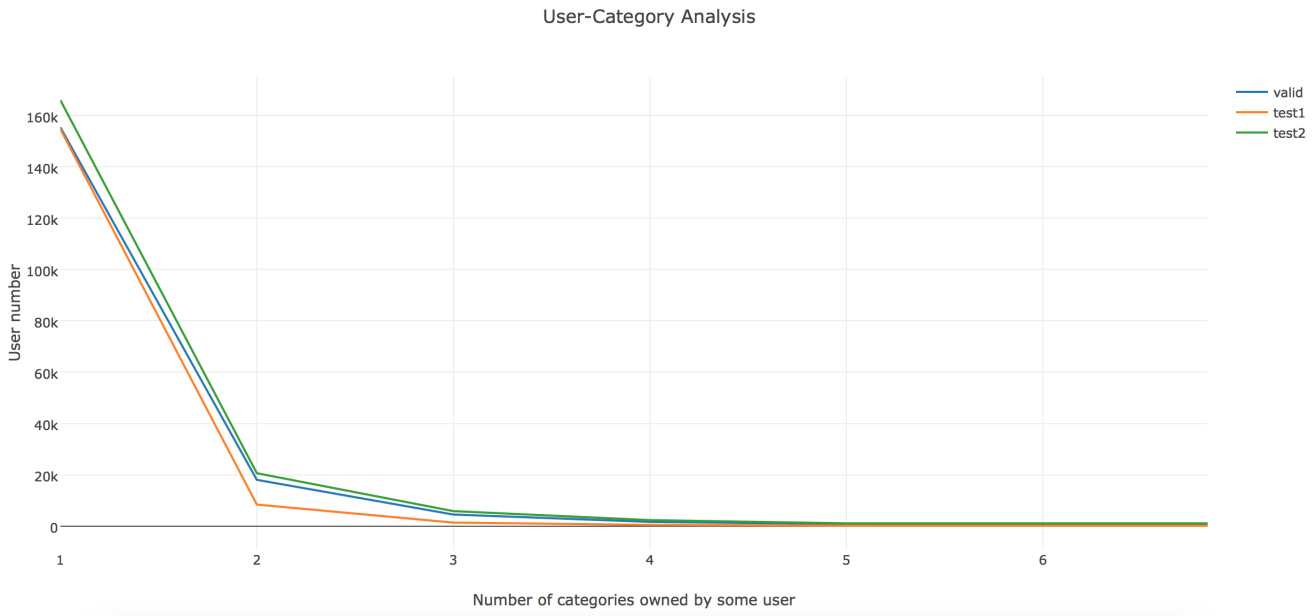

Figure 2: User-Category Analysis

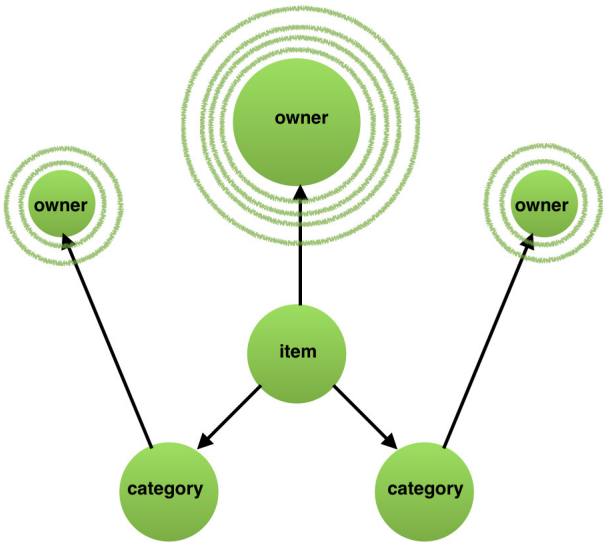

Figure 3: Social Network Model

2.2.1 Algorithm

We use embeddings to represent every user and item. First, we use user's relationship to train user's embeddings, so similar users will have similar embedding, then we use those pre-trained user embeddings to train item embeddings and category embeddings.

$$O_f = -\sum_{(u_j,u_i)} logP(u_j|u_i) \quad (1)$$

$$O_i = -\sum_{pairs} \{logP(u|item) + logP(item|c)\} \quad (2)$$

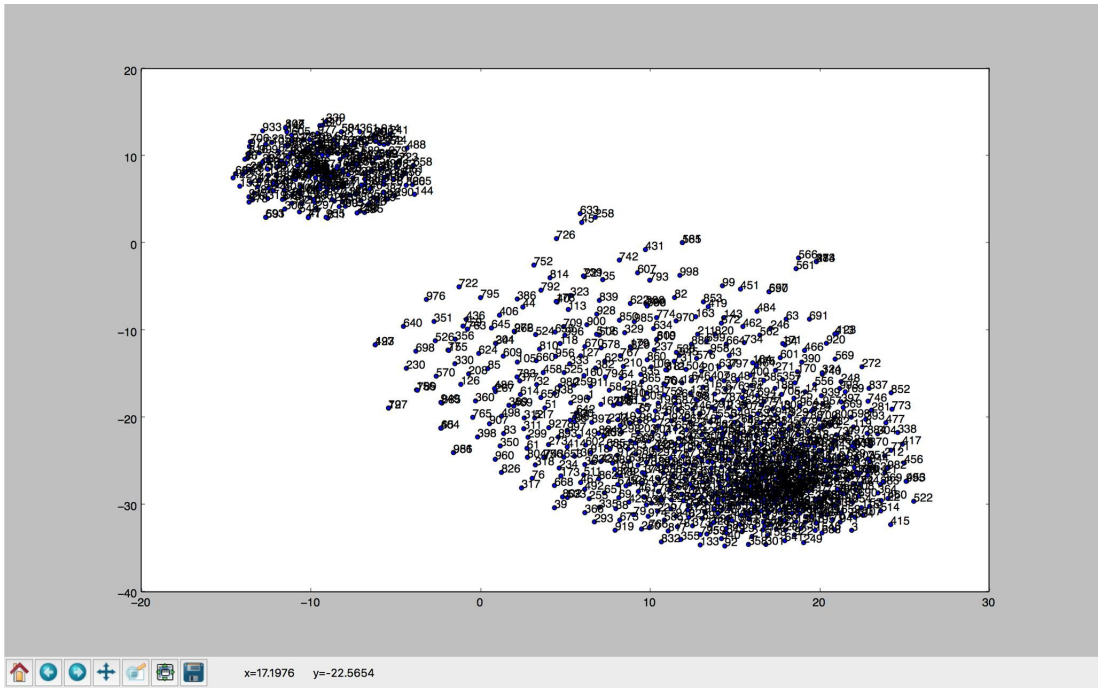

Figure 4: User Embeddings Visualization

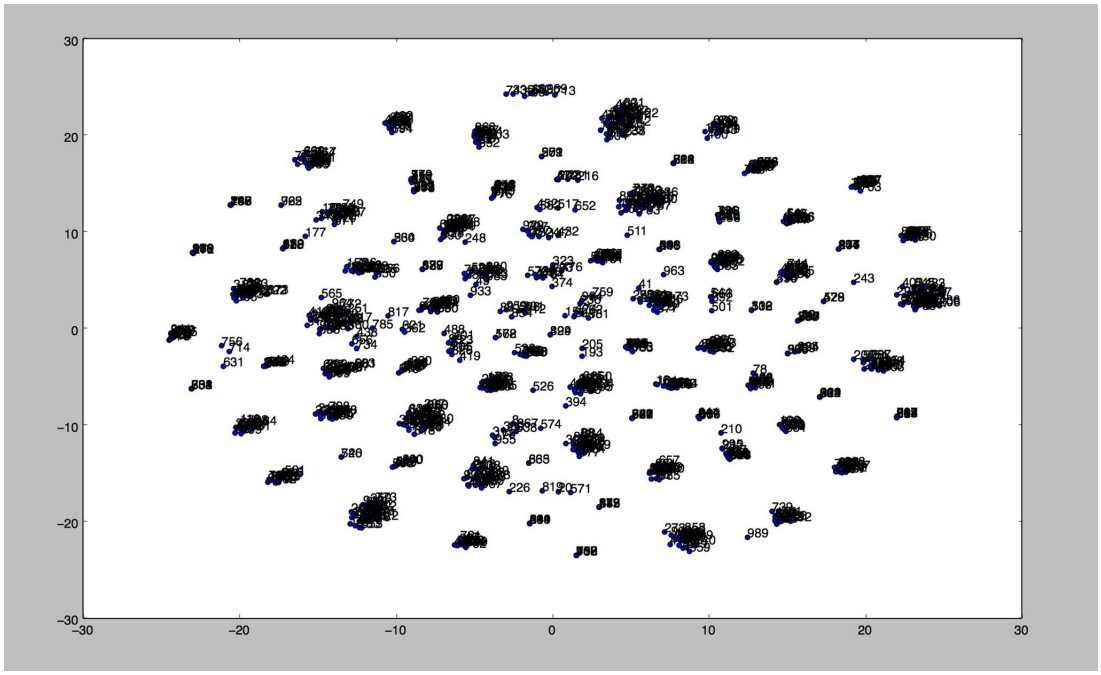

Figure 5: Item Embeddings Visualization

In Figure 4, we can see that users are divided into two groups, the user's data consist of two social networks. In Figure 5, items are divided into multiple groups, where the number of groups is close to the number of categories.

2.2.2 Inference

Owner's embedding are similar to his friends' embedding, so the product of owner's embedding and its item's embedding is similar to the product of his friends' embedding and item's embedding. The items in the same category will also have similar embedding.

For each pair in predict.id, we dot their embeddings as their score, and sort the scores in ascending order. Then we use the middle score as threshold, we predict the value of the pair from highest score to lowest score, if the score is larger than the threshold and the count of assigned pair is less than or equal to the item's link , then we predict this pair as 1, and vice versa.

3 Comparison and Analysis

Methods	Accuracy
Graphical Model (Max path)	0.93952
Graphical Model (Mixed path)	0.97651
Social Network Embeddings	0.67606
Social Network Embeddings(Bidirectional)	0.69654

Social Network embedding method mainly model the Owner-Item relationship. It might loss some category information in validation dataset. The bidirectional embedding is built by bi-direction friendship. The result shows that it gain more informative User embedding, so it has higher performance.

For Graphical Model, Max path means that probabilities may come from different paths, but those probabilities may not be comparable. Because the probability represention are actually $P(user, item|PATH1)$, $P(user, item|PATH2)$ and $P(user, item|PATH3)$, and they are not comparable. Mixed path combines the three paths and uses the same weights in a dataset, so it's comparable for all links in the same dataset.

4 Division of work

Tung-Chun, Chiang	Model Innovation and Implementation, Report
Yen-Chen, Fu	Model Innovation and Implementation, Report
Han-Hao, Chen	Model Innovation and Implementation, Report

References

[1] Jian Tang, Meng Qu, Mingzhe Wang, Ming Zhang, Jun Yan, Qiaozhu Mei. *LINE: Large-scale Information Network Embedding*. WWW, 2015.