Meta Path based KNN Recommendation

Tong Zhou

Overview

Users and movies are connected in a variety of ways, such as a friend of the target user has viewed the movie, and the user used to view some other similar movies. Different ways can be found in a heterogeneous network, which includes users, groups, movies, and the others properties of both users and items. For different paths, which is meta paths, that connect user and movie, we make a prediction based on KNN model and we calculate a weighted sum of predictions from different paths. Our project is an extension of KNN recommendation. The core of the algorithm is to learn the weight of different paths. In essence, meta path is a way to capture discriminating properties of both users and items.

General Term

Theory

Key Words

Rating prediction, recommender system, meta path, heterogeneous network.

Heterogeneous Information Networks

Traditionally, the user-item interaction information (e.g., rating information) are extensively exploited in collaborative filtering. Recently, the social recommendation methods illustrate that the social relations (e.g., following and trusting relations) among users can improve recommendation performances.

In this case, the objects and their relations in recommender system constitute a Heterogeneous Information Network (HIN) where there are different types of objects or relations. An example on movie recommender system is illustrated in Fig. 1.

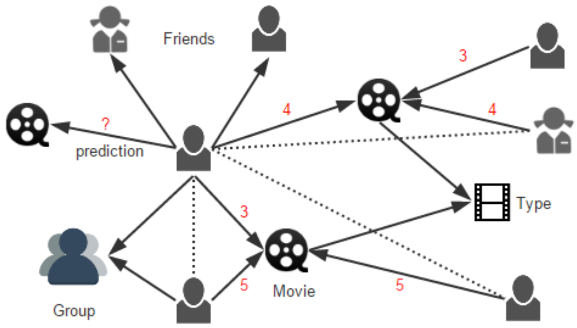

Figure 1: an example of Heterogeneous Information Network in movie prediction

In Fig. 1, the network not only contains different types of objects in movie recommendation but illustrates all kinds of relations among objects, such as viewing information between users and items, social relations among users, and attribute information of users or items(e.g., type of movies).

Meta path

Proposed by Sun[1], meta path provides a variety kinds of link between a user and a movie when applied to recommendation. For example, meta path UU (i.e., User, User), indicates that two users are linked by friendship; meta path UGU (i.e., User, Group, User) means that two users are connected because of a group they are both in.

Framework

- Collect data from social website. For a user u , in addition to the rating records of u , we collect people who are in a same interest group with u . For a movie m , we collect its type, such as fiction, action or romantic.
- Split the data set into two sets, namely training set and test set.
- Predict a score for each record in training set, the predicted score of u on m is affected by the rating on m of his similar users from different meta paths. Based on a loss function, the weight of each path is learned through iteration.
- Evaluate the performance on test set.
- Compare the model with other common model, like matrix factorization, SVD plus plus, and baseline method.

Methodology

Data collection

We collect our data from Douban (http://www.douban.com/). Aimed at sharing opinion on movies, Douban also contain social relationship, users can be friends with other users and can join in interest groups. Figure 1 illustrates the heterogeneous information network in Douban community.

Meta path

No	Meta path	Semantic meaning
1	UUM	Movies viewed by friends of the target user
2	UGUM	Movies viewed by users in the same group of the target user
3	UMUM	Movies viewed by users who view the same movies with the target user
4	UMTMUM	Movies seen by users who view the movies having the same types with that of the target user
5*	UMTM	Movies that are of the same types as movies viewed by the target user

Table 1: the meanings of different meta paths.

Note that the first four paths corresponds to user KNN while the 5th path corresponds to item KNN.

Prediction Algorithm

We employ basic K-nearest neighbor model to make prediction, i.e., rating of a target user is represented by the weighted sum of the ratings of a group of similar users on the same item. Similarly, the predicted rating can also be represented by the weighted sum of the ratings of a group of movies that are similar to target and are viewed by target user. In this model, we discover similar users and similar movies through meta path.

Training Algorithm

- Prediction
  a. Given the pair (u,m)
  b. Find different similar user sets for the target from different meta paths. For example, for user u, we can find four different user groups, i.e., friends of u, users who are in a same group as u, users who have watched the same movies as u.
  c. For each group, a meta path based rating is calculated. The prediction of user i on movie m is described as below.

$$r(i,m\,|\,P_l)=\frac{1}{N_1}\sum_{j\in G_l} r(j,m) \qquad (1)$$

In which P_l denotes the one of the meta path, G_l denotes a group of similar users of i based on the meta path *l*. N_1 denotes the number of users in G_l that have rated the movie m.

  d. Considering all meta paths, the prediction of i on m is described as below.

$$r(i,m)=\frac{1}{N_2}\sum_{l} w_l\cdot r(i,m\,|\,P_l) \qquad (2)$$

In which w_l denotes the weight of meta path *l*. N_2 denotes the number of meta paths.

- Iteration
  a. Loss function is defined as below

$$\Phi=(r-r(i,m))^2+\lambda\sum_{l} w_l{}^2 \qquad (3)$$

  b. In which r denotes the real rating and λ is a regularization factor to avoid overfitting.
  c. In the model we employ stochastic gradient descent to iterate.

$$\Phi'=2(r(i,m)-r)\frac{\partial r(i,m)}{\partial w_l}+2\lambda\sum_{l} w_l \qquad (4)$$

$$w_l \leftarrow w_l+\alpha\Phi' \qquad (5)$$

In which α is the learning rate.

pseudo-code

```
Initialize $w,\lambda,\alpha$.
For user-item pairs (u,m) in training set:
    Score=0
    For different meta paths l:
        Calculate $r(u,m\,|\,P_l)$
        Score+= $w_l\cdot r(u,m\,|\,P_l)$
    $r$ =Score
    For l:
        $w_l \leftarrow w_l+\alpha\{r(i,m\,|\,P_l)(r(i,m)-r)+\lambda w_l\}$
```

Evaluation and Discussion

We collect 600,000 rating records from Douban, and 3267 users who are from 5 different groups. Among the users, there are 1743 pairs of friends. We planned to divide the rating records into several groups. Current result is described as below, which is tested on the whole set.

Method	Performance
Matrix Factorization	RMSE=0.652 MAE=0.505
Meta path based KNN	RMSE=0.645 MAE=0.544
Biased Matrix Factorization	RMSE=0.665 MAE=0.529
SVD++	RMSE=0.802 MAE=0.639

Table 2: performance on the whole set

Note that further experiments are under way.

Group selection

To develop a denser network of users, we select 5 interest groups that are likely to have overlapping users and skip inactive users. The number of users in different groups is described as below.

Group	Number of users
French movie group	765
Japanese movie group	570
British group	843
Romantic movie group	770
Drama group	319

Table 3: the name of groups and the total number of users in different groups

Initialization

Due to the sparseness of social networks, a large amount of users do not have friends or never join in any interest group. In other words, not all kinds of meta path lead these users to a certain movie and thus we have no idea of how these invisible meta paths affect them. Still, we allocate the same weight on all paths and assume that all the paths affect the prediction of the user to the same extent in the beginning.

Smooth

In iteration, the weight of a certain path is updated in the way described in (5).

$$w_l \leftarrow w_l + \alpha\{r_l \cdot (\hat{r} - r) + \lambda w_l\} \tag{6}$$

In which r_l denotes the predicted rating on path *l*, $\hat{r}$ denotes the ratings considering all the paths. In the formula, the update of w_l involves r_l. However, in many cases, r_l is unknown for several paths. In this situation, we use known rating of other paths to represent the unknown ratings.

$$for\ k \in S_2:$$

$$\begin{cases} r = \frac{1}{N_1} \sum_{p \in S_1} r_p \ (if\ \ S_1 \neq \phi) \\ r = \bar{r}_m \ (if\ \ S_1 = \phi) \end{cases} \tag{7}$$

In which *p,k* denotes meta path, and S_1 represents the a group of paths that can not connect the target user to the target movie and thus can not predict a rating while S_2 is the complement of set S_1. N_1 denotes the size of S_1 *and* $\bar{r}_m$ denotes average ratings on movie *m*.

Further experiments

To go a step further, we plan to divide our users into several distinct groups and test our model upon these certain small groups.

No	Characteristics of the group
1	A group of users who average rating is over 4
2	A group of users who average rating is below 3
3	A group of users who are highly active and have rated more movies than average

Table 4: groups to be added and their characteristics

In terms of meta path UMUM, which implies movies viewed by users who view the same movies with the target user. Empirically, we set the threshold 10 and that means users who have viewed more than 10 same movies as target user are considered as similar users. In further experiments, we plan to take the rating into account and we only consider users who average rating is similar to that of target user as a possible neighbor. For example, if user *v* has viewed lots of same movies as user *u*, but their average score of these movies are disparate, we still do not consider user *v* a neighbor of *u*.

In addition, we plan to add more kinds of meta path (e.g., some users can be connected by tags or 5th path in table 1) and measure their importance in prediction.

Conclusion

Our project applies KNN recommendation in heterogeneous information network, and thus make use of discriminating properties of both users and items.

Reference

[1] Y. Sun, J. Han, X. Yan, P. Yu, and T. Wu. Pathsim: Meta path-based top-k similarity search in heterogeneous information networks. In VLDB, pages 992–1003, 2011.