Interim Progress Report

Guangmo Tong Xinwen Zhu Ruili Yao

1. Topic

In this project, we apply big data techniques and machine learning algorithms to twitter dataset. In particular, we consider two problems of friend recommendation.

2. Dataset

The considered dataset reveals the who-follow-who relationship on Twitter. The whole dataset forms a directed graph $G = (V, E)$ which contains $N = 11,316,811$ nodes and $85,331,846 edges$. An edge from node i to j implies user i follows user j.

3. The problem

In a online social network, people are always looking for new friends. The k-friend recommendation problem aims to reveal k users who should be recommended to a certain user on Twitter. Intuitively, a user prefer to follow anther user who has the similar interests. Therefore, given a user i, the users recommended to i are supposed to be similar to i. In this project, we develop the measure of similarly based on the who follow who relationship from the dataset. For a user i , we define its attribute vector A_i as $A_u = (A_i^1, ..., A_i^N) \in \{0, 1\}^N$ where $A_i^j = 1$ (resp. $A_i^j = 0$) implies i follows (resp. does not follow) j. We define $A(i, i)$ to be 1.

4. Basic Recommendation

As introduced in the last section, associate with each user there is an attribute vector. Thus, according to the Jaccard similarity measure, the similarity $S(i, j)$ of two users i and j can be defined as

$$S(i,j) = \frac{\sum_{l=1}^{N} min(A_i^l, A_j^l)}{\sum_{k=1}^{N} max(A_i^l, A_j^l))}. \quad (1)$$

Therefore, a natural approach is to recommend to j the k users that have the largest $S(i, j)$.

5. Clustering-enhanced Recommendation

In this section, we discuss another recommendation approach, namely clustering-enhanced recommendation. On twitter, a node can be referred to several topics, i.e., sports, news and music. The basic recommendation proposed in the last section has the drawback that the similarity defined in Eq. (1) may magnify a certain topic too much. For example, suppose there are ten users $u_1, ..., u_{10}$ with the following who-follow-who relationship.

$$\begin{array}{c} \\ u_1 \\ u_2 \\ u_3 \end{array} \begin{array}{c} \begin{array}{cccccccccc} u_1 & u_2 & u_3 & u_4 & u_5 & u_6 & u_7 & u_8 & u_9 & u_10 \end{array} \\ \left[\begin{array}{cccccccccc} 1 & 0 & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ 0 & 1 & 0 & 1 & 1 & 1 & 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 1 & 1 & 1 & 1 \end{array}\right] \end{array}$$

Now let us consider select one user from u_2 and u_3 to recommend to u_1. One can check that $S(u_2, u_1) > S(u_3, u_1)$, according to Eq. (1). However, if we have known that u_4, u_5, u_6 and u_7 are all music pages, and u_8, u_9 and u_{10} are related to different other topics, then u_1 has two common interests with u_2 while four with u_3. Therefore, in this case, u_3 is a better choice for recommendation. To address this problem, we first clustering the users into several clusters and then redefine the attribute vector of each users based on the clusters. These two steps are shown as follows.

Topic clustering. We first clustering the users into p clusters by the k-means algorithm, where the distance of two users is calculated by the length of shortest path.

Friend Recommendation. Suppose the clusters are $C_1, ..., C_p$ where each C_i is set of users. One can see that each cluster servers as a group of similar users. Now we redefine the attribute vector $\overline{A}_i$ of user i as follows.

$$\overline{\mathbf{A}}_i = (\overline{A}_i^1, ..., \overline{A}_i^p), \quad (2)$$

where $\overline{A}_i^j$ is the number of users in cluster C_j followed by user i. According to the new attribute vector, we compute the similarity of two users by the cosine similarity measure.

$$\overline{S}(i,j) = \frac{\overline{\mathbf{A}}_i \cdot \overline{\mathbf{A}}_j}{|\overline{\mathbf{A}}_i| \cdot |\overline{\mathbf{A}}_j|}. \quad (3)$$

6. Current Progress

Our goal is to implement both the basic and clustering-enhanced recommendation approaches via Spark. We have done the data collection and will focus on the implementations in the following weeks.