Social network analysis and reference system construction on Bilibili

Xiao Chen
School of Data Science
Fudan University
16307100066

Cheng Zhong
School of Data Science
Fudan University
16307110259

Lulu Zhou
School of Data Science
Fudan University
15307110349

***Abstract*—Online social networks have provided a great amount of information about the relationship between users. Constructing effective recommendation systems produce huge profits by improving the user experience. In our research, we crawled the social relation data from the video sharing website Bilibili, extracted useful features of this network and finally applied them to the construction of a recommendation system. Biased local random work algorithm was used and the precision we achieved on test set was 26%. Moreover, as we found that the social relationship of Bilibili was not significant, we considered adding the similarity of the uploaders' videos to the algorithm, and this could improve the accuracy by 8%.**

***Index Terms*—Social network analysis, web crawler, Reference system, Local random walk, Bilibili.**

1. Introduction

Social network analysis tries to describe the features of network and predict personal behavior using data mining method [1]. With the development of online social networks, a large amount of data has been generated, which gives us a unique opportunity to know the structure of social network and personal behavior patterns. How to make full use of data in online social network is of both theoretical interest and practical significance. One feasible and promising application of social network data is recommendation system.

In our research, the online social network was represented by a directed graph. Directed graph is a data structure which includes a collection of nodes and a collection of edges that relates one node to another. With the graph data structure, we abstracted the relation of users from the rich information on the website and focused on the interaction of users.

In this paper, we analyzed the data of Bilibili, a leading video sharing website in China, to get insights into the structure of social network on this website. According to the features of the network, we constructed a reference system to recommend channels to users who are likely to follow them.

2. Related Work

Quite a few researches have been carried out in the field of social network analysis and reference system of video website. Covington[2] successfully used deep learning to improve the recommendation system of YouTube. However, only search and view history were used to train the model, which indicates that the users were not put in their social context. A search-based method is applied to learning the features of network[3]. The search method could also be used in the process of crawling data from the website directly, which enables that the feature of the network could be studied without downloading all the data from the website. Susarla et al. studied the rate of diffusion of information on YouTube[4] and discovered that the rate of diffusion was significantly influenced by channels with many subscribers outside the local community. The importance of weak ties in the process of information propagation was confirmed by Bakshy et al.[5]. In our research, we will pay special attention to users who are 'bridges' between communities. Biased local random walk algorithm was applied to link prediction on social networks and showed good performance[8,10].

3. Data Collection and visualization

In this section, we introduce the Bilibili website and our data collection method. A brief description and visualization of our dataset are also provided.

3.1. Bilibili

Bilibili is a video sharing website where users could watch, like, share, comment and upload their own videos. Suppose a user, Amy, decides to upload her first video. Once the video is uploaded, Amy creates her own channel as a video maker. Users who are interested in Amy's videos could 'follow' her and they will be notified whenever a new video is uploaded by Amy. Meanwhile, some users just want to be a video watcher and never upload videos. By March 2019, Bilibili has had over 100 million monthly active users. As a leading video sharing website for young people in China, Bilibili has great commercial value. To our knowledge, no thorough research on the social network

features of Bilibili has been carried out. That is why we chose Bilibili as our data resource.

3.2. Data collection method

A web crawler program was used to collect users' data from Bilibili. The basic idea of this crawler is breadth first search (BFS): initially, 140 seed users was added into the waiting list, and then the information of these users was downloaded while all the users they followed were added into the waiting list (Figure 1). The data crawled from Bilibili was stored in a Sqlite3 database. The database contained 2 tables, which saved attributes of users and relation of users respectively.

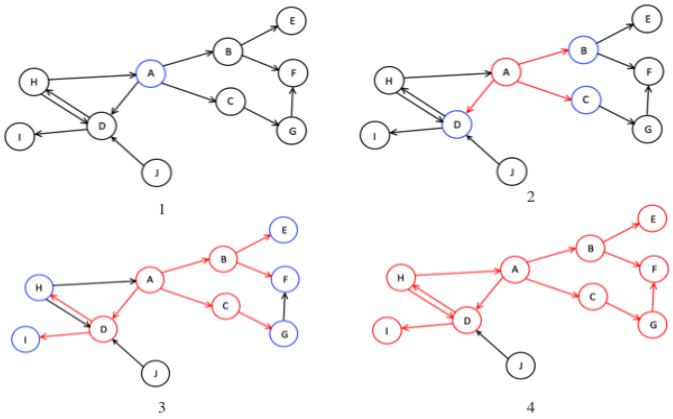

Figure 1. Example of a web crawler based on BFS algorithm. Red circles represent the nodes which has been crawled and blue circles represent the nodes in the waiting list.

3.3. Data description and visualization

By May 10, 2019, Information of 81399 users was downloaded from Bilibili, as well as 324827 relations among these users. Picture of the whole network and top 3000 nodes with highest degree (Figure 2) are drawn by Gephi [7], a data visualization software. Each point represents one user and each line represents 'follow' relation of users. We deduced that there was potential community structure in this graph.

In order to make a better observation, we reduced the number of nodes to 500. Community discovery algorithm was also applied to this network (Figure 3). The community colored with blue mainly consisted of game video maker and the community colored with red focused on cooking and everyday life. Users in pink community favored the topic of film and photography, while users in green were more official. According to this graph, we concluded that:

Feature 1: similar users tend to be friends with each other and form a community.

Finally, we visualized the communities detected network of top 3000 users based on k-clique:

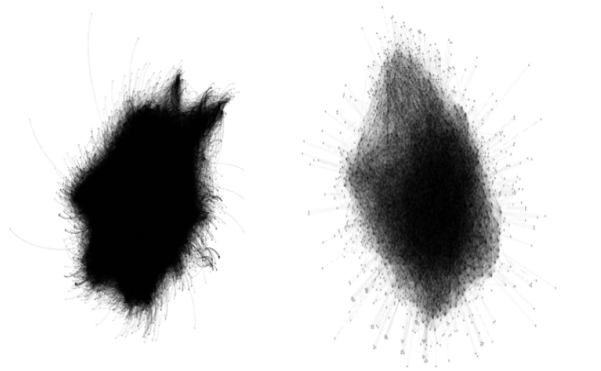

Figure 2. visualization of network

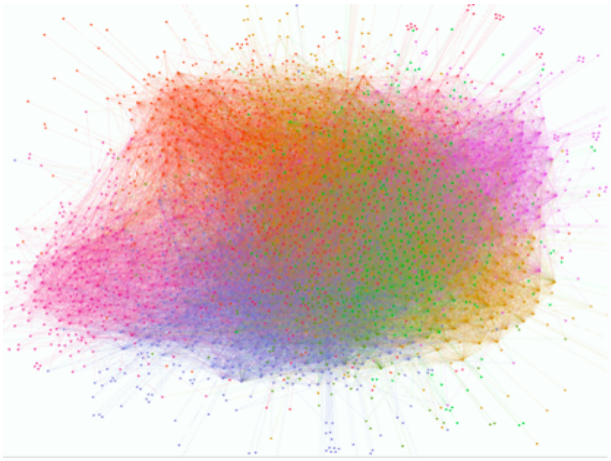

Figure 3. Top 500 network based on Modularity

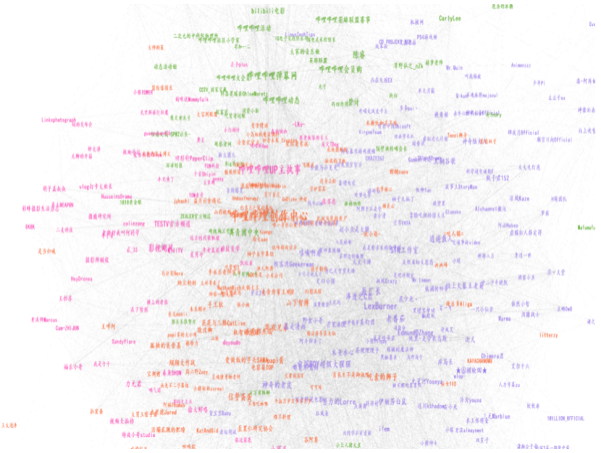

Figure 4. Top 3000 network based on k-clique

We set k=4 and got the community discovery result with 101 communities and we could find big colored parts in the figure showing that these big communities are well classified which strongly support our Feature 1.

3.4. Limitation

Although the data provide us with a unique opportunity to gain insights into the community of Bilibili, it has several limitations. First, Bilibili has over 100 million users and the data we collected was only a small part of it. Second, Breadth first search creates a biased sample which is more likely to include nodes with high centrality [6]. Finally, we could only download 100 relations of each user due to the regulation of the website. However, since the objective of our research is to analyze the relation of video makers with large numbers of fans and recommend these popular video makers to users, this biased dataset is just what we need.

4. Features of network and recommendation system

We measured the features of the network to know more about the overall structure of the network. Meanwhile, we tested the hypothesis that influential users are only a small portion of the total network.

4.1. Measurement on network

Different measurement methods were applied to the network to describe the features of it with the help of python package NetworkX [8]. First, we computed the distribution of degree (Figure 4). The number of edges which starts from the node is defined as the in-degree of this node. Similarly, the number of edges which ends in the node is defined as the out-degree of this node. We deduced that the distribution was exponential, which implies that users with high degree are only a small portion of all the users in the whole network. Moreover, the average degree is 11.32. However, the graph density is 0.004, the adjacency matrix is really sparse.

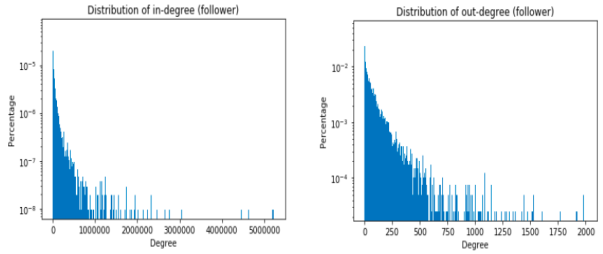

Figure 5. Degree distribution

Second, we computed eigen-vector centrality and page rank centrality of all the nodes and drew the distribution of them (Figure 5). Both eigen-vector centrality and PageRank centrality measures the importance of a node in the network. The definition of these two centralities are given in the documents of NetworkX [8]. When centrality grows, percentage of users drops at least at the speed of exponential function.

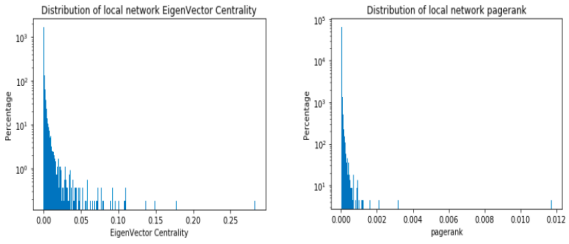

Figure 6. Eigen-vector centrality and page-rank centrality distribution

Finally, we computed the betweenness centrality (Figure 6). You could refer to documents of NetworkX to know more about the detail of the definition and computational algorithm of betweenness centrality [8]. Intuitively, nodes with high betweenness centrality frequently show in the shortest path of two nodes and are in the position of 'bridges' between communities. It is clear that most users have low degree centrality and only a few have high centrality.

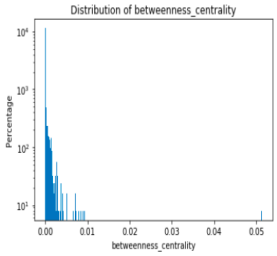

Figure 7. betweenness centrality distribution

To summarize the findings in the distributions of degree and centrality, we derived the following insight: Feature 2: Influence is unevenly distributed in the network of Bilibili. A few nodes are extremely influential while others are trivial.

4.2. Construction of LRW recommendation system

Biased local random walk algorithm [9] is deployed in the construction of our recommendation system. The possibility of walking from node i to node j is P_{ij}, which is determined by the following equations:

$$\begin{cases} S_{ij} = \dfrac{(V_i \cap V_j)}{d_{out}(V_i) \cdot d_{out}(V_j)} = \dfrac{|common(V_i, V_j)|}{|follow(V_i)| \cdot |follow(V_j)|} \\ \\ c_i = log(1 + betweeness(i)) \end{cases} \tag{1}$$

$$\Rightarrow P_{ij} = \gamma_{i,j}(\alpha \cdot S_{ij} + \beta \cdot c_j \cdot \frac{1}{d_{out}(V_i)}) \text{ , where} \gamma_{i,j} = \frac{1}{\sum_{j \in follow(V_i)}(\alpha \cdot S_{ij} + \beta \cdot c_j \cdot \frac{1}{d_{out}(V_i)})}$$

Biased local random walk is suitable in our context because of feature 1. The basic idea of local random walk is that a user is more likely to be related to the friends of their friends. As long as users who focus on similar topic tends to be friends with each other, recommending friend of friend to a user is reasonable. In the equations, is the weighted average of and . indicates the similarity of node and node . is the measure of centrality. Here we chose betweenness centrality because we thought that the nodes that bridge the communities are of more importance. If all the recommended video makers are from the same community, users could never be exposed to other topics. That is why we paid special attention to betweenness centrality, trying to solve this problem.

Due to the limitation of computational capability, only 3000 nodes with highest in-degree were sampled. However, since influential users are rare (Feature 2), sampling would not significantly affect the recommendation precision.

After the possibility matrix is computed, we multiplied it to the initial position vector for k times to get the position vector after k steps of random walk:

$P^k \cdot v_0 = v_k$

To get the final results, we sorted the elements of v_k to get the top 20 nodes with highest possibility, excluding the nodes in the initial position.

4.3. Performance of LRW reference system

We randomly sampled 500 users from the whole Bilibili network to test whether our algorithm can provide a good recommendation to a user based on the users he/she has already followed. The hyper-parameters were set as the following: $\alpha = 0.8$, $\beta = 0.2$, $k = 2$.

The sampled dataset was evenly split into 2 sets: train set and test set. The performance of the system is measured by the precision on the test set. The definition of precision is given below:

$$Recall = \frac{TP}{TP + FN}$$

, where TP is the number of users both followed by a user and recommended by the system and T is the number of users followed by a user. The average precision of the sampled data is 26%.

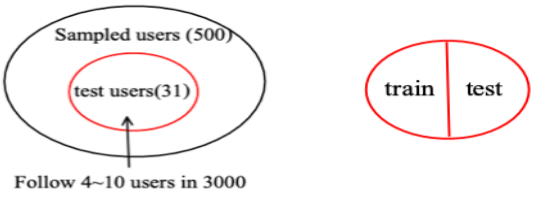

Figure 8. test sample

4.4. Discussion

The precision of biased local random walk algorithm on different data sets range from 13% to 99% [10]. As a matter of fact, the performance of our system is not perfect but it is much better than randomly recommend video makes to users. The precision of our could be further improved by modifying the hyper-parameters and combining other measures of centrality into c_i. A larger dataset with information of more users would also provide a more reliable result.

The time and space complexity of the random walk algorithm is $O(n^2)$, because the algorithm involves the multiplication of an $n \times n$ matrix and $1 \times n$ vector. However, the matrix is sparse, thus it is possible to improve the computation speed using a more effective biased local random walk algorithm [10].

4.5. Construction of similarity-based reference system

In above paragraphs we considered constructing this recommendation system only by the social relations between uploaders because we assume that they are more connected through these relations. However, we know from data that many super star uploaders rarely follow other people and in our data the relation is divided and not enough to show one ups attributes. And we need to ask that do these uploaders really need to follow other people? They just need to update their own contents. The social relation is not significant in these uploaders, and actually this also the real character of Bilibili. One more reliable method is to find the similarity among uploaders according to their submitted videos. Different uploaders have different styles and trends.so we crawled more data to demonstrate and calculate the similarity. We got total submitted video and their tags for each uploader. for instance, “Lexburner” has 1 in science catalogue, 144 in animation catalogue, 4 in film catalogue, 5 in national creation catalogue, 17 in daily life catalogue. And then we make use of this to find similarity of ups. Let be the vector of videos counting in each region for uploader , be the character vector for uploader :

Then, we get another transition matrix based on contents similarity. And then we could weight this matrix with the matrix we get in the biased local random walk, but it seems the necessity is weak—-we just don’t need the relation between the uploaders and this is more realistic. And we test this method with the same settings, the precision is somehow improved to 34%. This is remarkable. Video tag is just one extra measurement and we can crawl more other useful data if possible. Combining them together is also one application for some other websites but not suitable for Bilibili.

5. Conclusion

In this paper, we studied the feature of social network on Bilibli and constructed two reference systems using both the similarity and centrality of nodes. This effective reference system could help the website to improve the user experience and attract more users. We suggest that reference system should not only base on the individual but the social relations of it and the features of the whole network as

well. Improvements could be made on the computational speed and parameter selection of this reference system. Performance of this system on a larger scale is also worth studying.

References

[1] Borgatti S P, Mehra A, Brass D J, et al, “Network analysis in the social sciences[J],” Science, 2009, 323(5916): 892-895.

[2] Covington P, Adams J, Sargin E, “Deep neural networks for youtube recommendations[C],” Proceedings of the 10th ACM conference on recommender systems. ACM, 2016: 191-198.

[3] Grover A, Leskovec J. node2vec, “Scalable feature learning for networks[C],” Proceedings of the 22nd ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2016: 855-864.

[4] Susarla A, Oh J H, Tan Y. “Social networks and the diffusion of user-generated content: Evidence from YouTube[J],” Information Systems Research, 2012, 23(1): 23-41.

[5] Bakshy E, Rosenn I, Marlow C, et al. “The role of social networks in information diffusion[C],” Proceedings of the 21st international conference on World Wide Web. ACM, 2012: 519-528.

[6] Wang, Tianyi , et al. “Understanding Graph Sampling Algorithms for Social Network Analysis,” 31st IEEE International Conference on Distributed Computing Systems Workshops (ICDCS 2011 Workshops), 20-24 June 2011, Minneapolis, Minnesota, USA IEEE, 2011.

[7] Gephi: https://gephi.org

[8] NetworkX: https://nexworkx.github.io

[9] Feng, Shan, ”Research on reference system in social network [D]”, 2015.

[10] Liu, Weiping, and L. Lu. ”Link Prediction Based on Local Random Walk.” EPL (Europhysics Letters) 89.5(2010):58007.