Overlapping Community Detection in Directed Heterogeneous Social Network

Changhe Qiu[1,2,3], Wei Chen[1,2(✉)], Tengjiao Wang[1,2], and Kai Lei[3]

[1] Key Laboratory of High Confidence Software Technologies, Peking University, Ministry of Education, Beijing, China
pekingchenwei@pku.edu.cn
[2] School of Electronics Engineering and Computer Science, Peking University, Beijing 100871, China
[3] Shenzhen Key Lab for Cloud Computing Technology and Applications (SPCCTA), School of Electronics and Computer Engineering, Peking University, Beijing, China

Abstract. In social networks, users and artifacts (documents, discussions or videos) can be modelled as directed bi-type heterogeneous networks. Most existing works for community detection is either with undirected links or in homogeneous networks. In this paper, we propose an efficient algorithm OcdRank (Overlapping Community Detection and Ranking), which combines overlapping community detection and community-member ranking together in directed heterogeneous social network. The algorithm has low time complexity and supports incremental update. Experiments show that our method can detect better community structures as compared to other existing community detection methods.

Keywords: Community detection · Directed heterogeneous social network · Ranking

1 Introduction

Community detection[4,5] and ranking[2] in network are two dominating methods in social network analysis. However, both have their own defects[1].

We define X to be the artifact set, x to be one artifact, and Y to be the user set, y to be one user, and G_i to be the i_{th} community. Each artifact belongs to one community (because for example, in microblog, one blog tends to focus only one topic). Each user could belong to multiple communities (user may have many interests).

RankClus[1] combines clustering and ranking together in undirected heterogeneous network. However, its time complexity is high. It distributes users into all communities, which is unreasonable according to the truth.

In our method, we put forward transfer model to pass intermediary variable to connect two seemingly unrelated variables: artifacts and communities.

J. Li and Y. Sun (Eds.): WAIM 2015, LNCS 9098, pp. 490–493, 2015.
DOI: 10.1007/978-3-319-21042-1_47

2 OcdRank

The algorithm is iterative. All the artifacts are divided into k parts (in the first iteration, the distribution is random). Community-based ranking then carry out in every community. With the ranking score of nodes, we use link transfer model to evaluate the correlation coefficient between artifacts and communities. Artifacts and their associated users step into the most similar communities. Every loop the communities will be adjusted and the ranking score of nodes within the community will be changed. When iteration times reach a specified number or clusters changes only by a very small ratio, the iteration will terminate. Then, we get the meaningful communities.

Community-based ranking gives members in the same community a discriminating criterion. If the utilized ranking algorithm is PageRank [2], the score of users in our method can be formulated as

$$r(y'|G_i) = \frac{1-q}{N} + \mathrm{q}(\alpha \sum_{x \in Pub(y')} \frac{r(x|G_i)}{Out(x|G_i)} + (1-\alpha) \sum_{y \in Fol(y')} \frac{r(y|G_i)}{Out(y|G_i)}) \quad (1)$$

Where q is damping factor to make all the nodes in network can be accessed. And N is the number of nodes in community G_i. $Out(x|G_i)$ is the out degree of x in condition of community G_i and $Out(y|G_i)$ is the out degree of y accordingly. $Pub(y')$ is the artifact set y' published, $Fol(y')$ is the user set following y'. $\alpha \in (0,1)$ determines how much weight to put on each factor based on one's belief.

After getting the ranking score of every object in community G_i, we use transfer model to evaluate the correlation coefficient of artifacts and communities.

$$r(x|G_i) = \sum_{y \in (Dir(x) \cap Y')} \mathrm{r}(y|G_i) \quad (2)$$

Where $Dir(x)$ is the users directed to artifact x, Y_i is the user set, $\mathrm{r}(y|G_i)$ is the community-based ranking score of user y. Since every artifact only has relation with one community, but the associated users of it could be in different communities.

The correlation coefficient of artifacts and communities can be viewed as a decomposed vector. Every community could be formed as the normalized sum of all artifacts in it. Let $v(x) = (r(x|G_1'), r(x|G_2'), \ldots, r(x|G_k'))$ be the vector of artifact x, the vector of a community G_i' $(i = 1, 2, \ldots, k)$ is

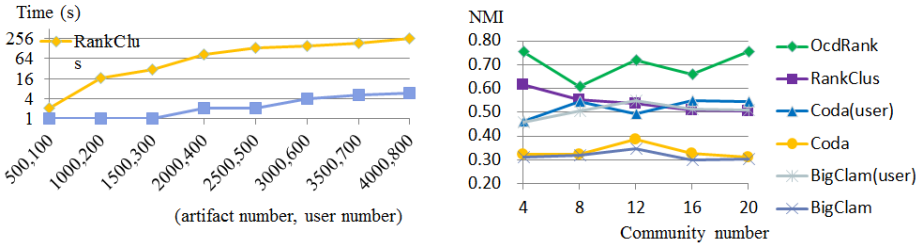
$$v(G_i') = \sum_{x \in X'} v(x) r(x|G_i') \quad (3)$$

$v(G_i')$ is the vector of G_i', $r(x|G_i')$ is the community-based ranking score of x. Each artifact and the associated users can be assigned to the nearest community according the cosine similarity scores.

When new artifact needs to be assigned to one community, there will be three considerations: 1) The artifact point to another one. We can directly assign it to the community where the pointed artifact is. 2) The associated users are partly in the user set Y. The correlation coefficient of artifact and communities can be computed directly by equation (2). The iteration goes on as our algorithm described. 3) There's no associated user in the user set Y. A new community should be created.

3 Experiments

We use Twitter (http://arnetminer.org/heterinf) and Weibo datasets. The baselines are BigClam[4], Coda[5] and RankClus[1] respectively. BigClam runs on undirected networks while Coda directed. Both are overlapping community detection methods in homogeneous network. RankClus[1] works in undirected heterogeneous network.



(a) Running time of OcdRank and RankClus (b) performance of different algorithms

Fig. 1. Comparison between OcdRank and other algorithms

Normalized Mutual Information (NMI) is an information-theoretic measure of similarity between two partitioning of a set of elements[3].

Fig. 1 (a) is the comparison of running time between RankClus and OcdRank.

We use Weibo dataset with different ground truth of different number of communities. For BigClam and Coda, we treat the heterogeneous networks as homogeneous. We use network of user links to compare with the ground truth of user network (BigClam(user) and Coda(user)). The parameter α of OcdRank is 0.5. We run each algorithm 10 times and get mean value. The result is shown in Fig. 1(b).

In Weibo, the rank result in each community is interesting. The users with higher ranking scores are usually the official accounts in same special field.

4 Discussion

OcdRank works on directed bi-type heterogeneous social networks. The ranking of users based on community can be used for experts finding. The ranking of artifacts can be used for semantic community detection. For very large datasets, running on distributed system is one of our future works. Another future work is semantic community detection, which would utilize the semantic information of artifacts.

Acknowledgments. This research is supported by the Natural Science Foundation of China (Grant No. 61300003), Research Foundation of China Information Technology Security Evaluation Center (No. CNITSEC-KY-2013-018) and Research Foundation Program of Ministry of Education & China Mobile (MCM20130361).

References

1. Sun, Y., Han, J., Zhao, P., Yin, Z., Cheng, H., Wu, T.: Rankclus: integrating clustering with ranking for heterogeneous information network analysis. In: Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology, pp. 565–576. ACM (2009)
2. Brin, S., Page, L.: The anatomy of a large-scale hypertextual Web search engine. Computer Networks and ISDN Systems **30**(1), 107–117 (1998)
3. Lancichinetti, A., Fortunato, S., Kertész, J.: Detecting the overlapping and hierarchical community structure in complex networks. New Journal of Physics **11**(3), 033015 (2009)
4. Yang, J., Leskovec, J.: Overlapping community detection at scale: a nonnegative matrix factorization approach. In: Proceedings of the Sixth ACM International Conference on Web Search and Data Mining, pp. 587–596. ACM (2013)
5. Yang, J., McAuley, J., Leskovec, J.: Detecting cohesive and 2-mode communities indirected and undirected networks. In: Proceedings of the 7th ACM International Conference on Web Search and Data Mining, pp. 323–332. ACM (2014)