Topical Community Detection: an Embedding User and Content Similarity Method

Thi Bich Ngoc Hoang[1,2]

[1] Institut de Recherche en Informatique de Toulouse, UMR5505 CNRS,
Université de Toulouse, France
[2] University of Economics, the University of Danang, Vietnam

Abstract. Community detection aims at partitioning a network into subgroups of densely connected nodes. While many approaches focus on community detection based on users' relationships, the latter may be not effectively enough since some communities may be topic dependent. In this paper, we propose a method that detects communities by considering users and their topics. More specifically, our approach combines cues extracted from the users' exchanges and the ones extracted from their posts. The data collection and the evaluation measures we intend to apply our method on are also presented in this paper. Yet, evaluation is not included in this paper.

Keywords: Social Media Analysis, Community Detection, Embedding User, Content Similarity, Twitter

1 Introduction

Community detection aims at partitioning a network into subgroups of densely connected users [4] with similar interest, background or purpose. Community detection is an important topic to discover the complex structure of social networks, and is applied in several fields such as biology, sociology and computer science.

It has been widely applied in several domains such as influence analysis [10], bibliometry [26], network security [5], and criminology [8].

It is also applied in social network analysis. A social network can be represented as a network composed of nodes and links, where the nodes denote the users, and the edges denote the relationships between these users.

In that context, community detection aims at grouping similar users into clusters, where users within a group tend to be more similar as compared to nodes outside the group.

Since community is originally determined based on linkage structure, previous community detection methods tend to purely consider the network's topology [1,9,21]. However, this information is not satisfactory in accurately defining the community membership because the topology is often sparse and noisy [25]. Other studies considered only the content to identify groups of users [22,24] but the results are not highly convincing since the inappropriate content attributes could miss-lead the process of community detection or some communities may be topic dependent. A few studies investigated combining users' link and users' content [15,20]. The authors consider each user as a node and the user's post content as the attributes of the nodes under the form of keywords. The authors then use a single assignment clustering method to detect communities. The applicability of these methods is limited: as each node can belong to a single community only, these methods cannot detect overlapping communities.

In this paper, we take on a method to identify communities in social networks that considers both users' interaction and their message content while using light computing. The users' interaction is defined based on the retweet action while the users' messages are considered in term of the semantic similarity. We integrate these two factors to detect communities. Our current work is to evaluate this method, thus results could not be included in this paper.

The rest of the paper is organized as follows: Section 2 presents the related work. Section 3 describes three proposed approaches to identify communities in social networks. Section 4 presents the experiment scenario, the data, and evaluation measures that could be used to evaluate this scenario. Finally, Section 5 concludes this paper.

2 Related work

Most of recent work in social community detection focused on users' interaction [1,9,21] while very few of the studies consider the users' message content [22,24] or combine these two factors to identify groups in social networks [15,19,20]. We only report studies that consider these both factors since they are comparatively few and new, and are the more related to our work.

Ruan et *al.* [19] extract communities by combing link strength and content similarity in graph structure. Link strength is measured based on whether the link is likely to reside within a community with high probability while content similarity is estimated through cosine similarity. They first create content edges among nodes, then sample the union of link edges and content edges with bias, retaining only edges that are relevant in local neighbor hood. Finally, they partition the simplified graph into clusters. Also integrating topology and content, Qui et *al.* [15] introduced a method of identifying social communities under the framework of non-negative matrix factorization. Their method uses adjacency to represent the network connectivity, then associates the corresponding semantic description of each community by adding an attribute to each node. This attribute corresponds to a keyword extracted from the content . The authors assume that the description for the same community should be semantically similar while the description among different communities should be different.

zhao *et al.* [27] proposed an approach to identify the topical opinion leader in social community question answering by combining the topic sensitive influence and the topical knowledge expertise. To measure the true topical influence of users, the authors incorporated the network structure, the topic interest similarity and the topical knowledge to measure the true topical influence. In their method, they infer each user's topic interest and knowledge authority from past posts. They confirmed the existence of homophily which implies that a user follows another having the similar topic of interest.

To measure the topical knowledge expertise, the authors employ the topic-relevant metrics that accounts for knowledge capacity, satisfaction and contribution.

Surian et *al.* [20] identified communities on Twitter network by either considering the tweet content or considering the follow relationship among users. They used Latent Dirichlet Allocation (LDA) method to infer the topics from tweets and used Louvain method to detect communities based on following interaction. They then measured the alignment between topics and communities for the users who were part of the largest connected component. They concluded that there are clear differences in the distribution of topics across communities defined by the follower network. The authors considered only the largest community detected by the two methods; thus their conclusion may not be relevant for the other communities.

Different from the above studies, we suggest to identify communities in social network by considering both the embedding user (retweet relationship among users) and the semantic similarity of the users' message content.

3 Proposed method

In this paper, we address the problem of community detection in social networks by considering both the *embedding users relationship* and *embedding users' message content*.

In this approach, each user is considered as a node. We first detect users who resent a post from other users and consider these interactions as relationships between users (edges between nodes). In social networks, most of users re-post messages from their friends (which appear on their timeline) when they agree with the message content or find the messages interesting. These re-posting messages may be extensions of or same as the original ones. The re-posts not only show indirect user relationships but also correspond to high semantic similarity in their content. We thus hypothesize that using the re-post interactions can help detecting communities with similar interest or background. In the next step, we use a community detection method to detect communities. Several community detection methods can be used in this purpose such as InfoMap [18], Label Propagation [16], Leading Eigenvector [12], Louvain [2], Spinglass [17], or Walktrap [14].

Whatever the social media considered, each user, either the user who writes an original message or the one who resent it, is a node. The approach is then implemented in the two following steps:

- We identify edges based on the resent relationship. If the user A resent a user B's post then we add an edge between the user A and the user B.
- We use traditional community detection methods to detect communities for above identified nodes and edges.

The result will be the number of communities and the members of each community. The approach will be evaluated using the data set and metrics that are described in the next section.

4 Evaluation framework

To evaluate the proposed approach, we developed an evaluation scenario that we will run on a tweet data set that we built. We will make this data set available to the research community on demand. Moreover, we will use usual metrics for community detection evaluation as presented in this section.

4.1 Data set

The data set we will use includes 20,000 retweets extracted from the 1 percent of tweets collected during the second week of January 2017 by IRIT, France within a spam detection project [23]. Each tweet in this data set is composed of several pieces of information regarding a twitter's post such as the author of the tweet, the content of the tweets and other objects. These $20,000$ retweets in our collection were created by $30,271$ users.

4.2 Evaluation measures

In community detection, algorithms are compared either on their efficiency (time taken to partition the network) or effectiveness (how relevant the extracted communities are) or both. With regard to effectiveness, various measures are used [11,3,6]; among this we will apply two well-known and widely used measures which are Modularity and Normalized Mutual Information as well as the newly defined f-divergence-based metric [7].

Modularity [13] is used to measure the difference of fraction of the edges that fall within communities and expected number of edges in a random graph:

$$Modularity = \frac{1}{2M}\sum_{xy}(A_{xy} - \frac{d_x d_y}{2M})\delta(c_x, c_y)$$

where x and y are nodes, M is the number of edges in the network, $\mathrm{d_x}$ and $\mathrm{d_y}$ are the degrees of x and y respectively; $\delta(\mathrm{c_x},\mathrm{c_y})$ equal to 1 when x and y belong to the same community and 0 in the other case.

Normalized Mutual Information (NMI) [3] is the measure used to evaluate the similarity between two partitions X,Y. The measure is:

$$NMI(X,Y) = \frac{-2\sum_{i=1}^{c_X}\sum_{j=1}^{c_Y} N_{ij} log(\frac{N_{ij}N}{N_{i.}N_{.j}})}{\sum_{i=1}^{c_X} N_{i.} log(\frac{N_{i.}}{N}) + \sum_{j=1}^{c_Y} N_{.j} log(\frac{N_{.j}}{N})}$$

where $\mathrm{N_{ij}}$ is the number of nodes in the community i (in X) that appear in the partition j (in Y); $\mathrm{c_x}$ and $\mathrm{c_y}$ are the number of communities in X and the number of communities in Y respectively; $\mathrm{N_{i.}}$ is the sum over row i of matrix $\mathrm{N_{ij}}$; $\mathrm{N_{.j}}$ is the sum over column j.

IRIT, UMR CNRS 5505, Université de Toulouse, France

Accordingly, if the communities in X match with the communities in Y then NMI index is equal to 1; if the communities in X are totally different from the communities in Y then the NMI index is 0; otherwise the amount will be in the range from 0 to 1.

f-divergence based metric [7]

$$\begin{aligned} MD_{\chi^2}(X,Y) &= 1 - \frac{\chi^2(P_{XY}, P_X P_Y)}{K_{max}} = 1 - \frac{\sum_{x,y} \frac{(p(x,y)-p(x)p(y))^2}{p(x)p(y)}}{K_{max}-1} = \\ &= 1 - \frac{\sum_{x,y} \frac{p^2(x,y)}{p(x)p(y)} - 1}{K_{max}-1}. \end{aligned} \tag{1}$$

where K_{max} denote the maximum number of communities in X and Y respectively, i.e. $K_{max} = \max\{K_X, K_Y\}$, where K_X and K_Y are the number of communities in partitions X and Y respectively.

4.3 Baselines.

The results of our method will be compared to the state of art [20] described in the related work section 2. For this, we will first re-implement the state of the art methods and applied them to our data set to obtain fair baselines.

5 Conclusion

In this paper, we propose a new approach to detect communities in social network considering both embedding users and messages content. We described the approach as well as the experiment scenario to evaluate the method on a real tweet collection. We have also described the evaluation metrics we will use to evaluate our approach and compare it to related work.

We are implementing the evaluation scenario and expect the result yields in the very near future. We are also designing variants of the proposed approach.

Acknowledgement. This work has been performed in the context of the PREVISION project, which has received funding from the European Union's Horizon 2020 research and innovation programme (H2020-SU-SEC-2018) under grant agreement No 833115. The paper reflects only the authors' view and the Commission is not responsible for any use that may be made of the information it contains (`https://cordis.europa.eu/project/id/833115`).

Ethical issue. While detecting communities from social media raises ethical issues, they are beyond the scope of this paper.

References

1. Ball, B., Karrer, B., Newman, M.E.: Efficient and principled method for detecting communities in networks. Physical Review E **84**(3), 036103 (2011)

2. Blondel, V., Guillaume, J., Lambiotte, R.: Fast unfolding of communities in large networks. Journal of Statistical Mechanics: Theory and Experiment **2008** (October 2008). https://doi.org/10.1088/1742-5468/2008/10/P10008
3. Danon, L., Diaz-Guilera, A., Duch, J., Arenas, A.: Comparing community structure identification. Journal of Statistical Mechanics: Theory and Experiment **2005**(09), P09008 (2005)
4. Fortunato, S.: Community detection in graphs. Physics Reports **486**, 75–174 (2010)
5. Gao, C., Ma, Z., Zhang, A.Y., Zhou, H.H., et al.: Community detection in degree-corrected block models. The Annals of Statistics **46**(5), 2153–2185 (2018)
6. Haroutunian, M., Mkhitaryan, K., Mothe, J.: f-Divergence Measures for Evaluation in Community Detection (regular paper). In: Workshop on Collaborative Technologies and Data Science in Smart City Applications (CODASSCA 2018). pp. 137–145. AUA NEWSROOM (American University of Armenia, affiliated with the University of California)), http://newsroom.aua.am/ (2018)
7. Haroutunian, M., Mkhitaryan, K., Mothe, J.: A New Information-Theoretical Distance Measure for Evaluating Community Detection Algorithms. Journal of Universal Computer Science **25**(8), 887–903 (2019), `http://www.jucs.org/jucs_25_8/a_new_information_theoretical/jucs_25_08_0887_0903_haroutunian.pdf`
8. Karataş, A., Şahin, S.: Application areas of community detection: A review. In: 2018 International Congress on Big Data, Deep Learning and Fighting Cyber Terrorism (IBIGDELFT). pp. 65–70. IEEE (2018)
9. Karrer, B., Newman, M.E.: Stochastic blockmodels and community structure in networks. Physical review E **83**(1), 016107 (2011)
10. Li, X., Cheng, X., Su, S., Sun, C.: Community-based seeds selection algorithm for location aware influence maximization. Neurocomputing **275**, 1601–1613 (2018)
11. Mothe, J., Mkhitaryan, K., Haroutunian, M.: Community detection: Comparison of state of the art algorithms. In: 2017 Computer Science and Information Technologies (CSIT). pp. 125–129. IEEE (2017)
12. Newman, M.: Finding community structure in networks using the eigenvectors of matrices. Phys. Rev. E **74**(036104) (September 2006). https://doi.org/10.1103/PhysRevE.74.036104
13. Newman, M.E., Girvan, M.: Finding and evaluating community structure in networks. Physical review E **69**(2), 026113 (2004)
14. Pons, P., Latapy, M.: Computing communities in large networks using random walks. Lecture Notes in Computer Science, Springer **3733** (2005). https://doi.org/10.1007/11569596$_3$1
15. Qin, M., Jin, D., Lei, K., Gabrys, B., Musial-Gabrys, K.: Adaptive community detection incorporating topology and content in social networks. Knowledge-Based Systems **161**, 342–356 (2018)
16. Raghavan, U., Albert, R., Kumara, S.: Near linear time algorithm to detect community structures in large-scale networks. Physical Review E **76**(036106) (September 2007). https://doi.org/10.1103/PhysRevE.76.036106
17. Reichardt, J., Bornholdt, S.: Statistical mechanics of community detection. Phys. Rev. E **74**(016110) (March 2006). https://doi.org/10.1103/PhysRevE.74.016110
18. Rosvall, M., Bergstrom, C.: Maps of random walks on complex networks reveal community structure. PNAS **105**(4), 1118–1123 (July 2007). https://doi.org/10.1073/pnas.0706851105
19. Ruan, Y., Fuhry, D., Parthasarathy, S.: Efficient community detection in large networks using content and links. In: Proceedings of the 22nd international conference on World Wide Web. pp. 1089–1098 (2013)
20. Surian, D., Nguyen, D.Q., Kennedy, G., Johnson, M., Coiera, E., Dunn, A.G.: Characterizing twitter discussions about hpv vaccines using topic modeling and community detection. Journal of medical Internet research **18**(8), e232 (2016)
21. Wang, X., Cui, P., Wang, J., Pei, J., Zhu, W., Yang, S.: Community preserving network embedding. In: Thirty-first AAAI conference on artificial intelligence (2017)

22. Wang, X., Jin, D., Cao, X., Yang, L., Zhang, W.: Semantic community identification in large attribute networks. In: Thirtieth AAAI Conference on Artificial Intelligence (2016)
23. Washha, M., Qaroush, A., Sedes, F.: Leveraging time for spammers detection on twitter. In: Proceedings of the 8th International Conference on Management of Digital EcoSystems. pp. 109–116. ACM (2016)
24. Yang, J., McAuley, J., Leskovec, J.: Community detection in networks with node attributes. In: 2013 IEEE 13th International Conference on Data Mining. pp. 1151–1156. IEEE (2013)
25. Yang, T., Jin, R., Chi, Y., Zhu, S.: Combining link and content for community detection: a discriminative approach. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 927–936 (2009)
26. Zhang, G., Jin, D., Gao, J., Jiao, P., Fogelman-Soulié, F., Huang, X.: Finding communities with hierarchical semantics by distinguishing general and specialized topics. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence. pp. 3648–3654. AAAI Press (2018)
27. Zhao, T., Huang, H., Fu, X.: Identifying topical opinion leaders in social community question answering. In: International Conference on Database Systems for Advanced Applications. pp. 372–387. Springer (2018)