ITR Lab Week 5: Social Networks as Graphs and Graph Mining

Current Usage Investigation

Social network analysis is a well-developed set of methods that has become an important tool for examining structures formed by various types of units within society. From a relationalist perspective, network analysis allows us to uncover social structures across multiple levels that emerge out of repeated relationships between these units. One of the most common units for social network analysis is definitely individual humans, with studies focusing on interactions ranging from romantic relationships to academic citations. However, in the context of cultural analysis based on large textual datasets, the concept of unit can be expanded from creators of texts to texts themselves, enabling a brand new set of inter-textual relationships. From another perspective, these inter-textual relations can also be aggregated to further generate relations between text producers. This approach provides richer insights into the pattern of both overarching and local structures of the corpus under study.

An interesting example for this is Arhab et al.'s exploration on online discussions about car parking behaviors, the major methodology of which is the application of community detection algorithms including Girvan-Newman to relevant social media datasets[1]. In their study, the units of analysis are the users of social media, and the relationships between them were constructed based on the Jaccard similarity of the posts they published on social media. Results showed that four major communities of users can be identified, each of which mainly focused on a certain theme about car parking, providing valuable evidence for car parking related urban planning and public policy development.

Community detection can also plays an important role as a part of the preprocessing pipeline. After community sub-graphs are extracted, further analysis can be conducted on those communities who are more closely linked in terms of internal interactions. For example, Ganguly et al. explored the diffusion of influence in various types of social networks on the basis of community detection[2]. After communities were identified within different social network graphs, two different models (IC-based and LT-based) were used to simulate the process of diffusion based on the seed nodes (which were selected based on centrality measures) of these communities, providing evidence for key factors influencing dissemination of information throughout the communities. Similarly, Park and Kwon developed a social media cyberattack detection model based on community detection and text analysis methods[3]. By combining community detection algorithm with inter-textual similarity measurements such as Word2Vec, communities' probability to relate to cyberattack could be measured, which was precise and efficient in identifying communities with higher risk of cyberattack.

Data Application

This week I applied community detection algorithm to a large dataset containing all posts on Chinese social media platform 'weibo' in January 2020 related to pandemic[4]. The data preprocessing pipeline was similar to those used in previous weeks and mainly included three parts: filtering posts that mention women, tokenization and stop words removal. After preprocessing, the inter-user network was constructed in reference to Arhab et al.'s project[1]. For any pair consisting of two users, the Jaccard similarity was calculated between the dictionaries constructed on users' posts. If the inter-user similarity exceeds a certain threshold (here we use $\gamma = 0.4$ which is recommended by Arhab et al.), an edge is added between these two user nodes. The largest connected sub graph can be visualized as follow:

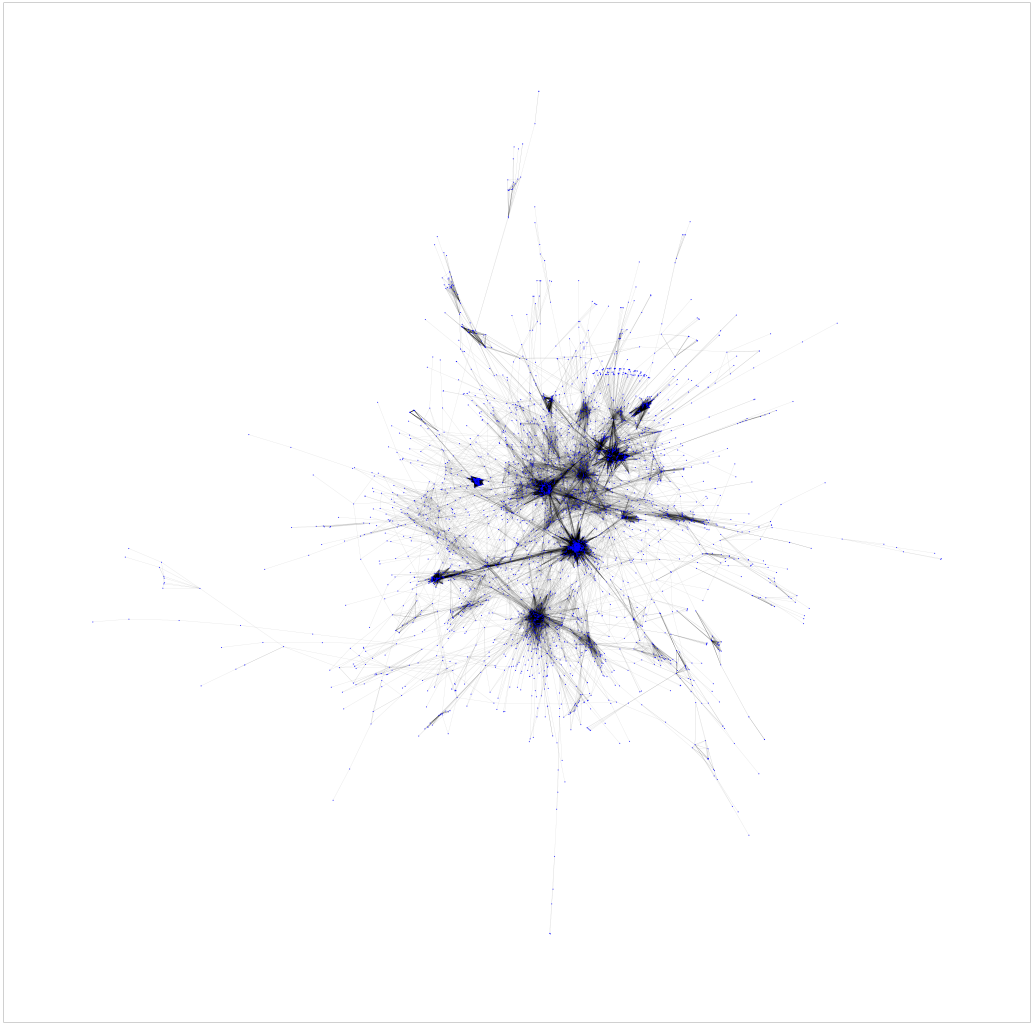

Figure 1: Largest Connected Network

To explore the inner structure of this largest sub graph, I applied Girvan-Newman algorithm to the graph to find out main communities inside. After that, I inspected posts published by several users that were classified into different communities, which provided insights into what topics different users were focusing on considering women at the starting period of the pandemic.

Apart from community detection, another major bottleneck considering time consumption in this data analysis pipeline is the construction of the user similarity graph: as the amount of users expand, the comparison need to be made increases exponentially. Although locality-sensitive hashing can help increase the efficacy of calculation, I will consider exploring different ways to represent posts and users and calculate inter-user similarity to make it more effective. Besides, although the dataset didn't include information related to forwarding relationship between posts, it's actually available for collection. The comparison between the forwarding network and the post-similarity network may be interesting.

Reflection

The results showed that the community detection algorithm was effective in identifying user communities focusing on different topics. For example, although many posts published by various users mentioned women, only a small community focused their discussion on women's role in the ongoing pandemic: an user in this community expressed her dissatisfaction with the discourse used by China's official media concerning the fight against the epidemic: 'Honestly, could these journalists stop putting men first in their headlines? Things like 'Husband Sends Wife Off to Battle' or 'Husband Bids Tearful Farewell to Wife' make it sound like they're the ones making a grand sacrifice, as if they're the real heroes...' (话说这些写新闻的，能不能不要把男人放前面？什么"老公送妻子出征"，什么"丈夫挥泪别过妻子"，搞得很悲壮的是他们，好像他们是男英雄似的...). Another interesting discovery is that geographical location seemed to be playing an important role in the formation of communities: apart from communities focusing on some specific topics such as women's role in the pandemic and Chinese new year, it seemed that some other communities were identified for their emphasis on local news about the pandemic. For example, a small community is consisted of users paying continuous attention to the pandemic in Tianjin, a municipal city in northern China. This may be caused by the usage of local dialects and expressions, or a particular cohesion within some of the local communities, as not all regions have formed their corresponding online communities. This result was rather different with those given by traditional topic models, as it might take some minor expressive traits that may not be relevant to any specific topic into account, thus placing more emphasis on the locality of communities. Since geolocation information is available for a small portion of the posts in the dataset, I may consider incorporating these geographic information variables, thus extending my analysis of networks on social media to three dimensions: discussion networks, forwarding networks and geographic networks. The similarities as well as differences between these three dimensions may provide me with insights that are different from the unidimensional analysis I was currently conducting.

References

[1] Nabil Arhab, Mourad Oussalah, and Md Saroar Jahan. Social media analysis of car parking behavior using similarity based clustering. *Journal of Big Data*, 9(1):74, December 2022.

[2] M. Ganguly, P. Dey, and S. Roy. Influence maximization in community-structured social networks: a centrality-based approach. *The Journal of Supercomputing*, 80:19898–19941, January 2024.

[3] Jeong-Ha Park and Hyuk-Yoon Kwon. Cyberattack detection model using community detection and text analysis on social media. *ICT Express*, 8(4):499–506, December 2022.

[4] Runbin Xie, Samuel Kai Wah Chu, Dickson Kak Wah Chiu, and Yangshu Wang. Exploring Public Response to COVID-19 on Weibo with LDA Topic Modeling and Sentiment Analysis. *Data and Information Management*, 5(1):86–99, January 2021.