Multimodal Social Network Dataset Based on Goldbach Conjecture Proved Event in Zhihu

LIU Tingzhen[1,†]，GENG Shijie[1]，HUANG Zhiquan[2]，WU Senxin[3]，WANG Zixi[4]

1. School of information Science and engineering, Shenyang University of Technology, Shenyang 100871;
2. School of computer science and technology, South China University of Technology, Guangzhou 510006;
3. School of Resources and Environment, South China Agricultural University, Guangzhou 510642;
4. School of , Hangzhou Dianzi University, Hangzhou 310018;

†Corresponding author, E-mail: firstsg@outlook.com

Abstract

At the end of 2018, a high school student asked a question in Zhihu community, claiming that he had proved Goldbach's conjecture. The problem caused an explosive reaction and a large number of users participated in the discussion. And has caused the widespread influence. On January 1, 2019, the questioner issued his "proof". His proof was soon proved wrong. The heated discussion caused by the incident contains a lot of information of social science analysis value. Therefore, we follow up the event in the first time and build a time series dataset for the event. Taking the questioner's "proof" as the dividing line, all the answers, comments, sub comments and user information of writing these texts before and after two days were recorded. This series of temporal information can reflect the dynamic features of the interaction between user opinions, and the impact of exogenous shocks (proof release) on community opinions. The dataset can be used not only for the demonstration of various social network analysis algorithms, but also for a series of natural language processing tasks such as fine-grained sentiment analysis for long texts, as well as multimodal tasks combining natural language processing and social network analysis. This paper introduces the characteristics and structure of the dataset, shows the visualization effect of social network, and uses the dataset train the benchmark model of emotion analysis.

Keywords: Social network analysis; Natural language processing; Dataset; Multimode; Opinion Dynamics

Introduction

Nowadays, online social networking sites have become an important platform for people to express their views and emotions. Designing effective algorithms for analyzing data on social networks is of great value, because social networking sites record a large number of users' social interaction, dialogue, communication and cooperation patterns. The data in social networking sites is divided into structured data and unstructured data. Structured data reflects the interaction between people in social networks. Through structured data, researchers can analyze how views spread and vary in the network, that is, social network analysis. In this field, researchers propose various communication dynamics models to describe social phenomena. However, due to the lack of suitable real data, most of these studies are carried out on the simulated network. Unstructured data generally refers to the natural language corpus in social networks. Rich corpus is very important for the research of natural language processing. However, most of the Chinese corpus data sets from the Internet are short texts with short length and simple connotation. The data of rich and long online Chinese corpus is relatively scarce. However, long text data with rich connotation is of great significance to solve natural language understanding problems such as fine-grained emotion analysis.

In view of the lack of social network structured data and Chinese long text corpus reflecting real social events, this paper constructs a data set for Chinese social network emergencies, which contains rich structured and unstructured data at the same time. The data set can not only be used for the empirical experiment of social network viewpoint dynamics model, but also for the training of Chinese language model. It can also be used for multimodal tasks combining graph structure data with natural language corpus. In addition, because there is an exogenous impact in the emergency, our data set takes the impact event as the central point and organizes the data of the first and second days. Therefore, the data set can be used to analyze the spatio-temporal dynamics of views, and how exogenous shocks affect the evolution of views.

In this paper, the characteristics, structure and possible application fields of the data set will be introduced. And visual display.

Related Work & Research Motivation

Social Network Analysis

Social network analysis seeks the relationship between the micro behavior of social individuals and the macro state of social system by studying the network composed of social actors as nodes. It often studies the evolution of systems with the help of complex network theory. Viewpoint dynamics is a key direction of social network analysis based on complex network theory. For hot events, everyone has his own point of view. At the same time, the communication between people will change their views. Modeling the evolution of views in the crowd in order to understand its dynamic characteristics will help such as group decision-making, false news identification, extreme thought account identification, precision marketing and the construction of interpersonal networks that can reach consensus. Researchers adopt a series of models represented by DeGroot [1,2], ising [3,4], bounded confidence model [5,10], voter model [6] and their mean field approximation [7,8,9] to study how viewpoint dynamics causes the phase transition of social system [10,11]. Such work generally follows the paradigm of complex network research, studies the behavior, stability, convergence and equilibrium point of the network under certain dynamic characteristics, and is verified by means of randomly generated network and numerical simulation [12]. The reason why we do not conduct empirical research on the real public opinion phenomenon is that most of the existing social network data sets can not reflect how the views in the social system change over time. There is also a lack of sufficient information for the algorithm to infer the complex view interaction. A few work [13,14] used the data sets built by researchers based on the data of online social networking sites, but most of these data sets are only built for this study and have limited significance for other tasks.

Therefore, we construct a real emergency data set. The event of "high school students proving Goldbach conjecture" was recorded. All answers, comments, sub-comments and user information of writing these texts in the first two days and the second two days were divided by the "proof" issued by the questioner. Based on these data, social networks can be constructed. The timing information can reflect the dynamic characteristics of the interaction between user views and the impact of exogenous impact ("proof" release) on community views.

Multimodal Analysis of Social Networks

In the current social network analysis tasks based on the data of online social networking sites, the follow or interactive behavior between users is generally used as the basis for establishing the connection between user nodes. However, simple follow and interactive behavior can hardly reflect the strength of the connection between users. For example, on many websites, users will follow a large number of users they see. At this time, the ability of follow behavior to reflect the association between users is very weak. There are many measures for interactive behavior. Some studies [15] use the number of interactions between two users (such as comments and forwarding) as the measure of their connection, but this method does not consider the intensity of a single interaction. For example, some users like to frequently forward the lottery microblog, but this does not mean that he has strong social ties with the blogger who initiated the lottery. In contrast, his substantive exchanges of views with some users better reflect social connections (although they may be less frequent). In view of this situation, it is necessary to analyze the semantics of user forwarding and comment text. To train this model to connect text with social network, we need to use multimodal data sets with both natural language text and social network structure.

At present, most of the commonly used social network data sets do not contain natural language corpus. For example, blogcatalog data set describes the connection between bloggers and their social relations (such as friends), and contains the blogger's interest tags. Zacharykarate Club data set is a real social network constructed by researchers through observing an American University karate club, including 34 nodes and 78 sides. Nodes represent members in the club, while edges represent the friendship between members.

A few social network data sets contain natural language corpus, such as wikifa data set, which contains 11381 users, forming 189004 different voters / voting pairs, with a total of 198275 votes. Each vote has a short reason written by the voter.

Our data set contains all the answers, comments and sub-comments under the question "if high school can prove Goldbach's conjecture, will it be escorted by the Department of mathematics of Tsinghua University and Peking University?" that is, it contains not only the comment text, but also the mutual reply relationship between these texts. Based on these reply relationships, social networks can be constructed. We also recorded the user information who wrote these answers and comments. The data set has three dimensions: social network connection, text corpus and user information, so it can be used in the multimodal task of the integration of natural language text and social network.

“Zhihu Goldbach Conjecture Proof Event” Dataset

Describe

On December 22, 2018, the user "proved" asked in Zhihu community: "if high school can prove

Goldbach's conjecture, will it be escorted by the Department of mathematics of Tsinghua University and Peking University?", claiming that he has proved Goldbach's conjecture and will announce the proof process at 0:00 on January 1, 2019. The question triggered an explosive response. By the time the "proof" was released, the number of fans of the questioner had increased from zero to more than 36000. Under this problem, ordinary student users, big V in academic circles, and users of different ages and identities have expressed various views on this event. This event is similar to other social network emergencies. This kind of emergency information will generally realize multiple coupling interactions such as superposition and combination without expectation, and have an impact on the steady state of the original community. For example, people "fled the city" caused by the outbreak of infectious diseases, and "beating, smashing and looting" caused by the spread of rumors. Therefore, the analysis of the viewpoint network of this event is helpful to grasp the internal evolution mechanism of public opinion of other emergencies.

Our dataset has several characteristics:

1. The viewpoint dynamic analysis of an emergency requires comprehensive data in all time aspects of the event. If there is no timely data capture in each time section of the event, it is difficult to carry out corresponding analysis. As the event progresses, some users will delete or modify their speeches, making the analysis based on speech time no longer accurate. In order to obtain the first-hand data, we follow up the event at the first time, grab and process the data every day, so that the data set reflects the complete process of viewpoint evolution.
2. The data set contains an exogenous shock, that is, on January 1, 2019, the questioner released his "proof" of Goldbach's conjecture. At the first time of publishing the proof, some users interpreted the proof and confirmed it as an error. This impact has triggered further changes in community views, even if more users have a negative tendency towards the event. We have captured the data of equal time before and after the "proof" release, and can study how exogenous shocks affect the evolution of community views based on this data.
3. Taking the exogenous shock event as the axis of symmetry, we collected the data of answers and comments in the past few days. Every day's data can construct a social network, and the time series data reflect how the network state changes with time. Combined with rich user attribute information, we can study the viewpoint dynamics characteristics emerging in the interaction between space and time attributes under emergencies, and how these characteristics are affected by user node attributes.
4. Our dataset contains all answers, response comments and sub-comments under the question. Since all users are discussing the same topic, there is an exchange of views between the user who answers and the user who comments under the answer, and between the user who comments and the user who makes sub-comments under the comment, that is, there is a connection on the social network. In this way, the social network describing the relationship between users can be obtained, so as to establish a relationship between the answer, comment text and the social network structure. In addition, we also recorded all the followers of the questioner. These followers are users interested in "proof events", so they have a connection with the questioner. By checking the statements of followers under the question, we can infer the attitude (positive or negative) and intensity (degree of concern) of this social connection. Based on these data, the cross modal task of inferring social network connection parameters from natural language texts can be completed.

Dataset structure

Since the questioner released his "proof" on January 1, 2019, the duration of the data before and after needs to be symmetrical. In addition, considering the update speed of answers and comments, we crawl the data at 22:00 every day on December 30, 2018, December 31, 2018 and January 1, 2019. Since the question has been deleted after January 1, we crawled the data under the question "what are the errors in the proof of Goldbach conjecture by high school students?" on January 2, 2019. The data volume of each type in four days is shown in the table below:

Table 1 Data volume of different types in four days

Date	Number of answers	Number of comments	Number of sub-comments
2019.12.30	1879	5029	6806
2019.12.31	2203	6229	8014
2019.1.1	2532	6936	8499
2019.1.2	489	1515	1437

The data set is organized with the date as the root directory. Each date directory contains three subdirectories: answer, GetData, follower and. The answers, comments (sub-comments) and the followers of the questioners of the day are recorded respectively. Since the questioner voluntarily cancelled his account on January 1, 2019, there was no follower information on January 1, 2019 and January 2, 2019. The specific attributes of each type of data are shown in the table below:

Table 2 Specific attributes of each type of data

Answer	Comments	sub-comment	Questioner's follower
Answerer nickname	Number of likes	Number of likes	Nickname
Answerer avatar	Creation time (UNIX timestamp)	Creation time (UNIX timestamp)	Profile
Answerer profile	Whether it been deleted	Whether it been deleted	Number of answers
Answerer gender	Content	Content	Number of followers
Whether it is an excellent answerer (if yes, including the field classification of excellent answerers)	Commentator nickname	Commentator nickname	
Whether it is folded (if yes, including reasons for folding)	Commentator avatar	Commentator avatar	
Whether it is anonymous	Commentator profile	Commentator profile	
Creation time (UNIX	Commentator	Commentator	

timestamp)	gender	gender
Update time (UNIX timestamp)	Whether it is answerer's comment	Whether it is answerer's comment
Content		

Social network visualization

We take users as nodes and interactive (comment) relations as edges to build social networks. Visualize the data of January 2, 2019, and the network structure is shown in the figure below:

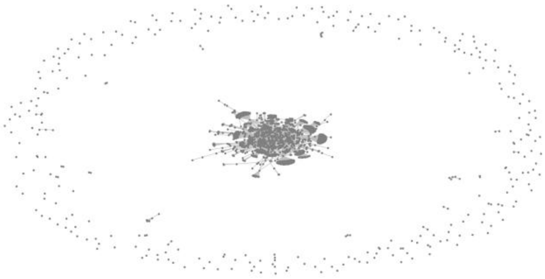

Fig.1 January 2, 2019 social network structur

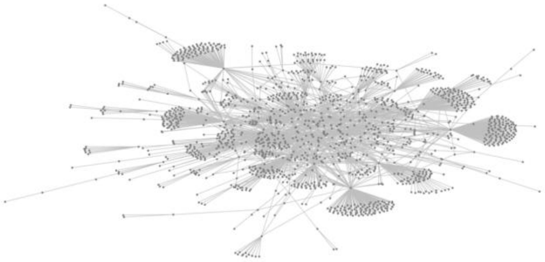

Fig.2 January 2, 2019 social network structure (amplification Center)

It can be seen that the network, like most social networks, has scale-free characteristics. The user's nickname is displayed in the figure, and the local visualization results of the network are shown in the figure below:

Fig.3 2019.1.2 local structure of social network (including user nickname)

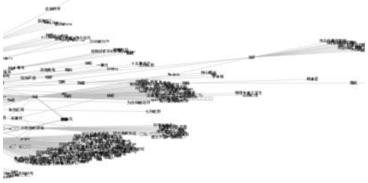

Fig.4 2019.1.2 local structure of social network (including user nickname)

Fig.5 2019.1.2 local structure of social network (including user nickname)

Access Dataset

Dataset: https://github.com/Goldbach-Research-Group/data

All relevant codes (e.g. operation data): https://github.com/Goldbach-Research-Group

Reference

[1] Zhou Q, Wu Z, Altalhi A H, et al. A two-step communication opinion dynamics model with self-persistence and influence index for social networks based on the DeGroot model[J]. Information Sciences, 2020, 519: 363-381.

[2] Liang H,Yuan F,Zhou Z ,et al. Opinion separation in leader-follower coopetitive social networks[J]. Neurocomputing, 2021.

[3] Vazquez F , Krapivsky P L , Redner S . Constrained opinion dynamics: freezing and slow

evolution[J]. Journal of Physics A Mathematical & General, 2003, 36(3):L61-L68.

[4] Ben-Naim E , Krapivsky P L , Vazquez F , et al. Unity and discord in opinion dynamics[J]. Physica A Statistical Mechanics & Its Applications, 2003, 330(1-2):99-106.

[5] Hickok A , Kureh Y , Brooks H Z , et al. A Bounded-Confidence Model of Opinion Dynamics on Hypergraphs[J]. 2021.

[6] Yang H X , Wu Z X , Zhou C , et al. Effects of social diversity on the emergence of global consensus in opinion dynamics[J]. Physical Review E, 2009, 80(4 Pt 2):046108.

[7] Zhu Y , J Jiang, Li W . The critical behavior of Hegselmann-Krause opinion model with smart agents[J]. 2021.

[8] Fennell S C , K Burke, Quayle M , et al. Generalized mean-field approximation for the Deffuant opinion dynamics model on networks[J]. PHYSICAL REVIEW E, 2021, 103(1).

[9] Kolarijani M , Proskurnikov A V , Esfahani P M . Macroscopic Noisy Bounded Confidence Models With Distributed Radical Opinions[J]. IEEE Transactions on Automatic Control, 2021, 66(3):1174-1189.

[10] Schawe H , L Hernández. Collective effects of the cost of opinion change[J]. Scientific Reports, 2020, 10(1):13825.

[11] Jiang L L , Hua D Y , Zhu J F , et al. Opinion dynamics on directed small-world networks[J]. European Physical Journal B, 2008, 65(2):251-255.

[12] Ravazzi C , Hojjatinia S , Lagoa C M , et al. Ergodic opinion dynamics over networks: learning influences from partial observations[J]. IEEE Transactions on Automatic Control, 2021, PP(99):1-1.

[13] Liu Q , Xiao R . An Opinion Dynamics Approach to Public Opinion Reversion with the Guidance of Opinion Leaders[J]. Complex Systems and Complexity Science, 2019.

[14] Wu H , Hu Z , Jia J , et al. Mining Unfollow Behavior in Large-Scale Online Social Networks via Spatial-Temporal Interaction[J]. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(1):254-261.

[15] Biessmann F , Papaioannou J M , Harth A , et al. Quantifying spatiotemporal dynamics of twitter replies to news feeds[C]// IEEE International Workshop on Machine Learning for Signal Processing. IEEE, 2012.