ISSN 2706-6827 Vol. 4, Issue 2: 88-94, DOI: 10.25236/IJFS.2022.040215

Design and Implementation of Data Analysis System of Social Network

Jinglin Bai

Shandong University, Qingdao, China
bjl1306600@163.com

Abstract: *To solve the problem that data collection efficiency of social network analysis technology based on single-mode network structure is low and cannot meet the problem of data information processing of large-scale social network, the social network data analysis system based on the network node centrality theory is proposed. The naive Bayesian classification algorithm is used to improve the traditional data mining mode. The traditional social network data analysis technology is compared with the social network data analysis system proposed in this research from four aspects: multi-platform network data processing response time, data acquisition and processing accuracy, social network data feature condition introduction rate, and data feature condition introduction accuracy. The accuracy of data collection and processing in the traditional social network data analysis system is 89.26%, the response time of multi-platform network data processing is 7.32s, the introduction rate of data feature conditions is 82.65%, and the accuracy rate of data feature conditions is 78.88%. The data processing accuracy of the data analysis system proposed in this research based on the network node centrality theory is 98.99%, the response time of multi-platform network data processing is 1.35s, the data feature condition introduction rate is 97.91%, and the data feature condition introduction accuracy rate is 97.39%. The social network data analysis system proposed in this research is significantly better than the traditional social network data analysis system in all aspects, and its data processing performance is very good especially in multi-platform social network. The results show that the social network data analysis system based on network centrality theory proposed in this research is feasible and can meet the daily application and research requirements of social network data collection, storage, analysis and visual display, which has the advantages of fast data processing speed, high accuracy and wide range of data storage.*

Keywords: *Naive Bayesian classification algorithm; Network node centrality; Data analysis; Data storage; The social network*

1. Introduction

With the rapid development of computer technology, the popularity of QQ, WeChat, Weibo, Tougias, and Facebook and Twitter in China, the human-centered social network connecting information interaction between people has been bursting with endless passion and vitality. There is a growing number of services and applications based on social networks, making the relationship between modern social networks more and more complicated [1] [24]. Traditional social network analysis only needs to be based on the single-mode network structure, that is, all nodes are of the same type and the connections between nodes are of the same type. However, most social networks are multi-mode networks, that is, nodes and connections in the network belong to different types. One of the most important functions of social networks is to help users make new friends and expand their social circle [26]. For example, in the user clustering module, much interrelated information in social networks is often ignored based on single-mode networks. If the clustering is only based on the relationship degree of existing friends, a lot of useful information is lost, and the social circle formed is too small. User clustering based on multi-mode social networks can combine users' friends, interests and browsing information to diversify the clustering [2-4]. Therefore, how to extract the potential information for users with the complex multi-mode social network data has become the focus of researchers [25].

Data analysis refers to the process of collecting data, studying data and drawing conclusions with a certain purpose to find out useful information. According to the theory of statistical application, data analysis can be divided into descriptive statistical analysis, exploratory data analysis and confirmatory

ISSN 2706-6827 Vol. 4, Issue 2: 88-94, DOI: 10.25236/IJFS.2022.040215

data analysis [5]. The data analysis model based on the social networks has been the focus of many scholars. For example, Jorgensen et al. (2018) [6] demonstrated how to use Bayesian data expansion to adapt to missing cyclic data by using the social relationship model, including how to take partially observed covariables as auxiliary correlation factors or substantial predictors. Chang et al. (2018) [7] proposed a social network analysis method to perform big data analysis and demonstrated an internal development model of social network application program interface with six functions. It can be focused on the contacts who click on the likes or comments on the author's post, and the results can be extracted and visualized in data. Jarvie et al. (2018) [8] applied social network analysis to an example, conducted data modeling for 30 active major league baseball teams, and connected players to form 300 independent team networks. Generalized least-square regression was used to calculate and analyse social network indicators such as network density, network centralization and average weighting to predict winning and team ranking [9][10].

To sum up, traditional data analysis based on single-mode network structure leads to the low efficiency of data collection and mining, data overflow and other phenomena. Therefore, a social network data analysis system is proposed based on the theory of network node centrality, and naive Bayesian classification algorithm is adopted to improve the traditional data mining mode and solve the fundamental problem [27]. The traditional social network data analysis technology is compared with the proposed social network data analysis system in terms of response time, data mining accuracy, data feature condition introduction rate and data feature condition introduction accuracy of multi-platform network data mining, which proves that the proposed social network data analysis system based on network centrality theory is feasible.

2. Methodology

2.1 Basic theory of data analysis system design of social network

Centrality is the degree to which a point in a social network or a person is at the center of the network. This degree expressed numerically, is called centrality. In the dissemination of design network information, the importance of each node is different, some important nodes have greater influence, while others are insignificant. Therefore, according to the different methods of measuring center degrees, it can be divided into degree centrality, closeness centrality, betweenness centrality [11] [12].

Degree centrality represents the overall centrality of the network graph and reflects the degree of concentration of the social network. The degree center degree of the core point in the star network diagram is n-1, the rest points are 1, and the center potential is 1. The calculation equation is as follows:

$$CAD_i = d(i) = \sum_i x_{ij} \tag{1}$$

In Eq. 1, i is for row number, j for column number.

The degree of intermediary centrality represents the degree of "intermediary" at that point, that is, the degree of the medium. There are relative and absolute intermediary centrality. It is calculated as the ratio of the geodesics at any other point and the number of geodesics passing through that point. The calculation formula is as follows:

$$C_{ABi} = \sum_{j<K} \mathrm{b}_{jk}(i) = \sum_{j<K} g_{jk}(i) / g_{jk} \tag{2}$$

In Eq. 2, j and k represent any two nodes, G_{jk} represents the number of shortest paths connecting jk, and $G_{jk\,(i)}$ represents the number of i in the shortest path.

Proximity to center represents a measure that is not controlled by others. In layman's terms, it is how close one point is to all the other points. There is relative and absolute proximity. The calculation method is the sum of the geodesic distances between this point and other points. The calculation equation is:

$$C_{APi}^{-1} = \sum_j d_{ij} \tag{3}$$

In Eq. 3, d_{ij} is the geodesic distance.

ISSN 2706-6827 Vol. 4, Issue 2: 88-94, DOI: 10.25236/IJFS.2022.040215

2.2 Naive Bayesian classification algorithm

The thought foundation of naive Bayesian algorithm is: for the item to be classified, the probability of each category to be classified under the condition of the occurrence of this item is solved. Whichever is the largest, the category to be classified is considered to belong to. Suppose there are M categories D (D_1, D_2 … D_m) and F (F_1, F_2 … F_m), then the classified text is categorized into the category most closely related to it, that is, the probability of each text going is categorized into the closest category [13] [27]. Assuming that the influence of each attribute value on a given category is independent of each other, the equation of naive Bayes' theorem is:

$$P(D|F) = \frac{P(F|D)P(F)}{P(D)} \tag{4}$$

2.3 Framework design of social network data analysis system and functions of each module

The social network forms a virtual social network by simulating real society. For the needs of later research, the user data obtained should be random, instead of collecting special user data. Data analysis system mainly deals with the following data: user information, user relationship information (including user fans and user attention), user interest information and so on. This system is mainly aimed at the social network data collection, mining and user dynamic changes. The system is mainly divided into five parts: data collection, data storage, data mining, control module and end-user. Each module can independently accomplish specific functions in the system, as shown in Figure 1 below.

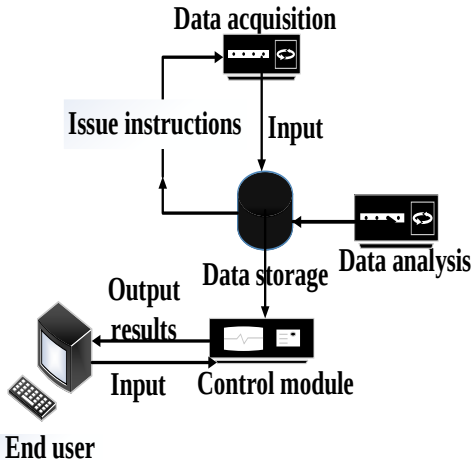

Figure 1: Schematic diagram of system architecture

The data acquisition module is mainly responsible for collecting data from social networks. Based on breadth-first search strategy, open API interface collects user relationship data and user interest data, including user's fan list and follow list. In the process of collecting user relations, the collected user fans or concerned users are added to the collection queue after Bayesian classification filtering for iterative collection to ensure the integrity of data collection [14]. User interest data is mainly collected by filtering the contents clicked and browsed by users, as shown in Figure 2 below.

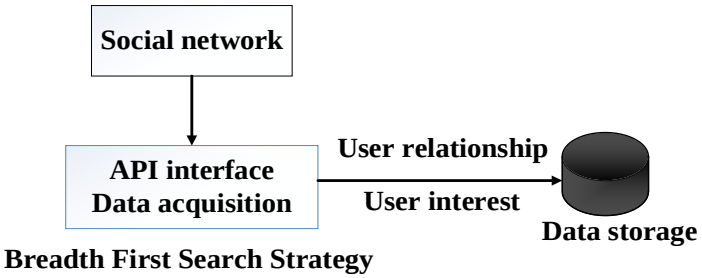

Figure 2: The schematic diagram of the data acquisition module

Data storage is to sort out the collected data, mainly to provide good access data for subsequent data

ISSN 2706-6827 Vol. 4, Issue 2: 88-94, DOI: 10.25236/IJFS.2022.040215

analysis. Establishing an appropriate data storage module can simplify and improve the subsequent analysis speed [15].

The data mining module is used to provide a series of data mining operation interface, data storage system as the output and input of the potential information mining. Because of the complicated logic of data mining, the stability and maintainability of the system can be greatly improved.

The control module first obtains the user's input in the end-user part, then gives instructions to other modules to complete the operation required by the user, and feedbacks the final result to the end-user. As the dispatching center of the system, it can connect each module to realize complete system functions.

The end-user module provides an interface for data information interaction, which is used to input the user's requirements and give the end-user direct feedback of the system running results.

2.4 The workflow of social network data analysis system

The data collection process of social network needs to be run in the network environment, because the required amount of data is relatively large, so the program runs for a long time. In the process of data collection, the operation may be interrupted due to the local network conditions and the insufficient capacity of the resource server [16]. If the program is terminated unexpectedly for some reason and the progress of collection is not recorded in time, the collected data need to be judged and processed again, resulting in very low efficiency of the system. As shown in Figure 3, the system controls the whole running program through the message queue. The data to be processed are added to the queue, and the processed data are deleted from the queue. The thread controls the number of calls to the API interface to prevent unexpected interrupts. The program delay method is adopted to make the program run for a period of time or call the API interface for a certain number of times [17]. The running is stopped for a while to get around the frequency limit of API interface calls. The program is then allowed to sleep on itself for a few minutes until the frequency limit for the API interface to continue requesting calls is met.

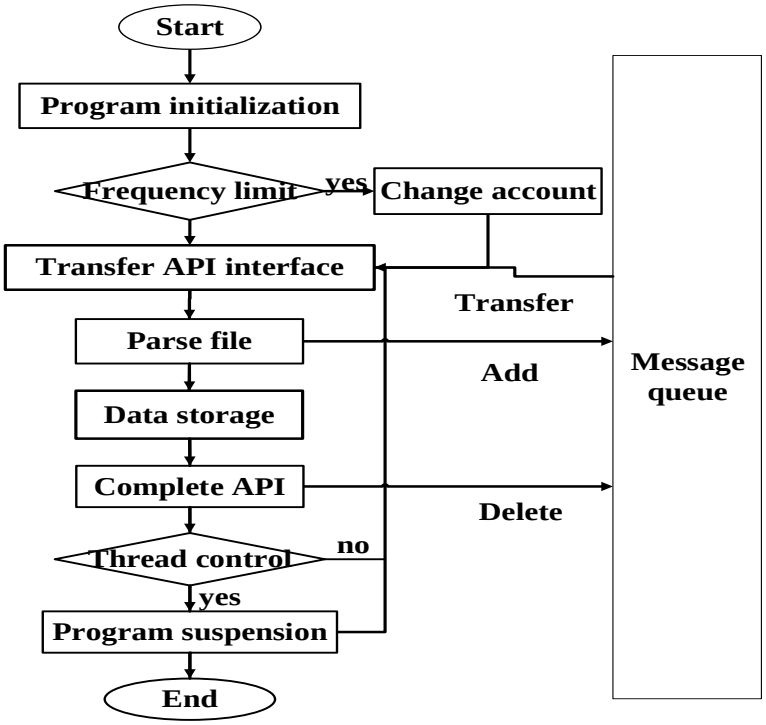

Figure 3: The workflow of the system

3. Results and discussion

The proposed social network data analysis system based on the theory of network node centrality is

ISSN 2706-6827 Vol. 4, Issue 2: 88-94, DOI: 10.25236/IJFS.2022.040215

simulated and tested. The random test of social network data under multiple platforms is carried out through a simulation experiment. The traditional social network data analysis technology is compared with the social network data analysis system proposed in this research from four aspects: multi-platform network data processing response time, data collection and processing accuracy, social network data feature condition introduction rate, and data feature condition introduction accuracy, which proves that the proposed social network data analysis system based on network centrality theory is feasible [29].

The specific development environment is as follows: the operating system is the Windows10 platform; the development platform is MyEclipse10, Java programming editor; the database is Mongo DB, which is used to store user relationship data and user interest data.

Through the simulation experiment, the randomly selected test of social network data under multiple platforms is conducted, and the results are shown in Table 1 below. The accuracy of data collection and processing in the traditional social network data analysis system is 89.26%, the response time of multi-platform network data processing is 7.32s, the introduction rate of data feature conditions is 82.65%, and the accuracy rate of data feature conditions is 78.88%. The data processing accuracy of the data analysis system proposed in this research based on the network node centrality theory is 98.99%, the response time of multi-platform network data processing is 1.35s, the data feature condition introduction rate is 97.91%, and the data feature condition introduction accuracy rate is 97.39%. Therefore, the social network data analysis system based on the node centrality theory and combined with naive Bayesian classification algorithm is significantly better than the traditional social network data analysis system in all aspects and it is very good especially in multi-platform social network data processing performance, which solves many problems from the root of the traditional social network data mining technology [28].

Table 1: Comparison of data processing performance between two network data analysis systems

Test content	Traditional social network data analysis system	The data analysis system proposed in this research
Data acquisition and processing accuracy (%)	89.26	98.99
Network data processing response time (s)	7.32	1.35
Data feature condition introduction rate (%)	82.65	97.91
Data feature conditions introduce accuracy (%)	78.88	97.39

4. Conclusion

Social networking sites are open computer platforms where users can simply register to publish information. Therefore, social networks regulate users too casually, resulting in too messy user data information on relevant platforms, and a lot of junk information also appears on the network platform [20]. Therefore, data collection should be filtered and classified. Data classification is an important content in data analysis. Text classification can be regarded as a guided learning process, which classifies target text according to the classification model and text features, which are obtained by training the training data set. As a typical machine learning method, naive Bayes has the advantages of simple implementation, high classification accuracy and fast classification speed, which is a relatively simple and superior text classification model, which is suitable for spam filtering and filtering [18] [19].

This study mainly introduces the design and implementation of an automated system for data collection and analysis based on network node centrality theory, including the breadth-first search strategy. The data acquisition module is connected to an open API interface, which provides a series of data mining operation interface data mining modules for the subsequent data analysis and data storage module for data access [21]. As the dispatching center of the system, it can connect each module to realize the control module of complete system functions. Through the demand analysis of the data analysis system, the overall framework is designed, and some targeted collection, storage and mining methods are put forward for the data information of a large number of user relationships and user interests generated by the social network [23]. Traditional social network data mining technology is based on the single-mode network, with slow data processing speed, low accuracy and poor

ISSN 2706-6827 Vol. 4, Issue 2: 88-94, DOI: 10.25236/IJFS.2022.040215

responsiveness under the background of the multi-platform social networks. To solve this problem, the naive Bayesian classification algorithm is used to optimize the logic and improve the computing ability of the system [22].

Through the simulation experiment, it is found that the automatic system of data collection and analysis based on the network node centrality theory has the characteristics of fast data processing speed, high accuracy and wide range of data storage, which can meet the daily application and research requirements of social network data collection, storage, analysis, and visual display, and provides a new research idea for the research field of social network data analysis. In the following work, researchers should further strengthen data processing ability and meet the demand of data mining for social network platforms through good data collection.

References

[1] Zengler K, Zaramela L S. The social network of microorganisms-how auxotrophies shape complex communities. Nature Reviews Microbiology, 2018, 16(6), pp. 383-390.
[2] Kim J, Hastak M. Social network analysis: Characteristics of online social networks after a disaster. International Journal of Information Management, 2018, 38(1), pp. 86-96.
[3] Sijtsema J J, Lindenberg S M. Peer influence in the development of adolescent antisocial behavior: Advances from dynamic social network studies. Developmental Review, 2018, 50, pp. 140-154.
[4] Hagen L, Keller T, Neely S, et al. Crisis communications in the age of social media: A network analysis of Zika-related tweets. Social Science Computer Review, 2018, 36(5), pp. 523-541.
[5] Shen L, Wang S, Dai W, et al. Detecting the Interdisciplinary Nature and Topic Hotspots of Robotics in Surgery: Social Network Analysis and Bibliometric Study. Journal of medical Internet research, 2019, 21(3), pp. e12625.
[6] Jorgensen T D, Forney K J, Hall J A, et al. Using modern methods for missing data analysis with the social relations model: A bridge to social network analysis. Social networks, 2018, 54, pp. 26-40.
[7] Chang V. A proposed social network analysis platform for big data analytics. Technological Forecasting and Social Change, 2018, 130, pp. 57-68.
[8] Jarvie D. Do Long-time Team-mates Lead to Better Team Performance? A Social Network Analysis of Data from Major League Baseball. Sports Medicine, 2018, 48(11), pp. 2659-2669.
[9] Sun D, Peverill M R, Swanson C S, et al. Structural covariance network centrality in maltreated youth with posttraumatic stress disorder. Journal of psychiatric research, 2018, 98, pp. 70-77.
[10] Akbari, E., Naderi, A., Simons, R.-J., & Pilot, A. (2016). Student engagement and foreign language learning through online social networks. Asian-Pacifc Journal of Second and Foreign Language Education, 1(1), 4. https://doi.org/10.1186/s40862-016-0006-7
[11] Asterhan, C. S. C., & Bouton, E. (2017). Teenage peer-to-peer knowledge sharing through social network sites in secondary schools. Computers & Education, 110, 16–34. https://doi.org/10.1016/j.compedu.2017.03.007
[12] Benson, V., & Filippaios, F. (2015). Collaborative competencies in professional social networking: Are students short changed by curriculum in business education? Computers in Human Behavior, 51, 1331–1339. https://doi.org/10.1016/j.chb.2014.11.031.
[13] Benson, V., Saridakis, G., & Tennakoon, H. (2015). Purpose of social networking use and victimisation: Are there any diferences between university students and those not in HE? Computers in Human Behavior, 51, 867–872. https://doi.org/10.1016/j.chb.2014.11.034
[14] Borrero, D. J., Yousafzai, Y. S., Javed, U., & Page, L. K. (2014). Perceived value of social networking sites (SNS) in students' expressive participation in social movements. Journal of Research in Interactive Marketing, 8(1), 56–78. https://doi.org/10.1108/JRIM-03-2013-0015.
[15] Borrero, J. D., Yousafzai, S. Y., Javed, U., & Page, K. L. (2014). Expressive participation in Internet social movements: Testing the moderating efect of technology readiness and sex on student SNS use. Computers in Human Behavior, 30, 39–49. https://doi.org/10.1016/j.chb.2013.07.032.
[16] Boyd, D. M., & Ellison, N. B. (2008). Social network sites: Defnition, history, and scholarship. Journal of Computer-Mediated Communication, 13(1), 210–230. https://doi.org/10.1111/j.1083-6101.2007.00393.x.
[17] Cheung, C. M. K., Chiu, P. Y., & Lee, M. K. O. (2011). Online social networks: Why do students use facebook? Computers in Human Behavior, 27(4), 1337–1343.
[18] Chung, J. E. (2014). Social networking in online support groups for health: How online social networking benefts patients. Journal of Health Communication, 19(6), 639–659. https://doi.org/10.1080/10810730.2012.757396.
[19] Doleck, T., Bazelais, P., & Lemay, D. J. (2017). Examining the antecedents of social networking

ISSN 2706-6827 Vol. 4, Issue 2: 88-94, DOI: 10.25236/IJFS.2022.040215

sites use among CEGEP students. Education and Information Technologies, 22(5), 2103–2123. https://doi.org/10.1007/s10639-016-9535-4.

[20] Eid, M. I. M., & Al-Jabri, I. M. (2016). Social networking, knowledge sharing, and student learning: The case of university students. Computers and Education, 99, 14–27. https://doi.org/10.1016/j.compedu.2016.04.007.

[21] Elphinston, R. A., & Noller, P. (2011). Time to Face It! Facebook intrusion and the implications for romantic jealousy and relationship satisfaction.

[22] Nwagwu, W. E. (2017). Social networking, identity and sexual behaviour of undergraduate students in Nigerian universities. The Electronic Library, 35(3), 534–558. https://doi.org/10.1108/EL-01-2015-0014.

[21] Pantic, I. (2014). Online social networking and mental health. Cyberpsychology, Behavior, and Social Networking, 17(10), 652–657. https://doi.org/10.1089/cyber.2014.0070.

[22] Parboteeah, D. V., Valacich, J. S., & Wells, J. D. (2009). The infuence of website characteristics on a consumer's urge to buy impulsively. Information Systems Research, 20(1), 60–78. https://doi.org/10.1287/isre.1070.0157.

[23] Park, N., Song, H., & Lee, K. M. (2014). Social networking sites and other media use, acculturation stress, and psychological well-being among East Asian college students in the United States. Computers in Human Behavior, 36, 138–146. https://doi.org/10.1016/j.chb.2014.03.037.

[24] Tang, J. H., Chen, M. C., Yang, C. Y., Chung, T. Y., & Lee, Y. A. (2016). Personality traits, interpersonal relationships, online social support, and Facebook addiction. Telematics and Informatics, 33(1), 102–108. https://doi.org/10.1016/j.tele.2015.06.003.

[25] Tashakkori, A., & Teddlie, C. (2003). Issues and dilemmas in teaching research methods courses in social and behavioural sciences: US perspective. International Journal of Social Research Methodology: Theory and Practice, 6(1), 61–77. https://doi.org/10.1080/13645570305055.

[26] Teo, T., Doleck, T., & Bazelais, P. (2017). The role of attachment in Facebook usage: a study of Canadian college students. Interactive Learning Environments, 4820(April), 1–17. https://doi.org/10.1080/10494820.2017.1315602.

[27] Tokunaga, R. S. (2010). Following you home from school: A critical review and synthesis of research on cyberbullying victimization. Computers in Human Behavior. https://doi.org/10.1016/j.chb.2009.11.014.

[28] Tower, M., Latimer, S., & Hewitt, J. (2014). Social networking as a learning tool: Nursing students' perception of efcacy. Nurse Education Today, 34(6), 1012–1017. https://doi.org/10.1016/j.nedt.2013.11.006.

[29] Van Hoof, J. J., Bekkers, J., & Van Vuuren, M. (2014). Son, you're smoking on Facebook! College students' disclosures on social networking sites as indicators of real-life risk behaviors. Computers in Human Behavior, 34, 249–257. https://doi.org/10.1016/j.chb.2014.02.008