Construction of Social Network Big Data Storage Model Under Cloud Computing

Zihui Jin[1(✉)] and Ting Chen[2]

[1] Tianfu College of Swufe, Chengdu 621050, China
jinzihui@tfswufe.edu.cn
[2] Chengdu College of University of Electronic Science and Technology of China, Chengdu 611731, China

Abstract. In order to better store massive network information data effectively, a cloud computing based method for building a big data storage model for social networks is proposed, which combines cloud computing theory to identify user interest data of social network data, constructs a hierarchical user interest conceptual model, and improves the ability of identifying data interest based on Cloud Computing; According to the process of the impact of network information dissemination, the evaluation algorithm of social network data storage is optimized, and the classification management is carried out based on the calculation results, so as to achieve the goal of classification storage of massive data. Finally, the experiment proves that the model built has high practicability in practical applications.

Keywords: Cloud computing · Big data · Social networks · Storage model

1 Introduction

With the development of Internet technology and social media, the network has been deeply integrated into people's daily work and life. People are used to publishing and sharing information using the network platform, and are used to obtaining information and increasing knowledge through the network. The wide use of social network big data and computers facilitates people to obtain and store data, which makes the network data grow exponentially. The amount of data has to be measured in ZB. Many important information and knowledge are hidden behind the surge of data [1]. Social network big data under cloud computing theory is relative to information cognitive ability. Compared with massive data, what is really meaningful to mankind is the knowledge behind it. People hope to conduct a deeper analysis of social network big data in order to make better use of data and absorb knowledge. Therefore, the application of social network big data has become the focus of attention and a scientific problem that people urgently explore.

Reference [2] proposes multi relationship group impact modeling in online social networks. Starting from the types of social relationships among users, it classifies and mines the complex network topology relationships in online social networks, analyzes

W. Fu and L. Yun (Eds.): ADHIP 2022, LNICST 468, pp. 205–216, 2023.
https://doi.org/10.1007/978-3-031-28787-9_16

the online group environment of different dimensions that users may perceive, and proposes the definitions and mining methods of static group environment and dynamic group environment In different online social group environments, the group structure characteristics perceived by users in the environment are quantified from a macro perspective, and the influence mechanism among users is modeled and simulated from a micro perspective. Reference [3] proposes the research on competitive nonlinear dynamic information dissemination model of online social networks, analyzes the internal relationship between the competition mechanism between different types of information on the network, node state transformation and information dissemination evolution law by using Markov chain theory, constructs a probability model of node state transformation from the perspective of probability, and constructs a nonlinear dynamic information dissemination model of network system information diffusion from the perspective of statistics; The equilibrium point of the propagation dynamics differential equation of the proposed model is solved and its stability is analyzed; Through the simulation analysis of the relative change relationship between the parameters in the model, the process of information competition and dissemination is simulated. The era of social network big data has changed the way people collect, disseminate and analyze information. For managers and users, it is more necessary to improve the means and level of traditional data management and use. Clarifying data association, eliminating the deviation caused by data redundancy and data loss, and mining the information, knowledge or wisdom behind the data are the key problems to be solved in the analysis and utilization of cloud computing. For social network big data, the content that people hope to obtain through the network can be divided into two categories: the lowest and most original data set, which can be called "objective reality data" and the most valuable information with the highest matching degree of self cognition, which can be called "objective information". Objective reality data and objective information can be transformed into information, knowledge or wisdom through in-depth mining and utilization, and finally contribute to the improvement of human innovation ability. Objective reality data exist without the influence of any subjective cognition and judgment. The objective information is often the description and judgment of the original data, which has been processed and analyzed by people. To some extent, objective information reflects people's current cognitive ability and knowledge structure, which may be updated by users when it is reused. Based on this, the construction of social network big data storage model under cloud computing is proposed. Combined with cloud computing theory to identify user interest data in social network data, build a hierarchical user interest conceptual model, and improve data interest recognition ability based on cloud computing; The calculation results are classified and managed to achieve the goal of classifying and storing massive data. The research shows that the constructed model has good effect.

2 Social Network Data Cloud Computing Storage Model

2.1 Social Network Data User Interest Collection Model

In order to realize the construction of lower level secondary user interest model in social network big data environment, the following research ideas are proposed, as shown in Fig. 1:

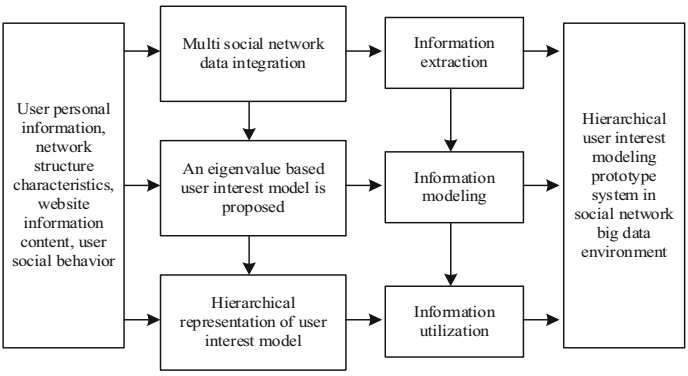

Fig. 1. Hierarchical user interest conceptual model

It can be seen from Fig. 1 that using the hierarchical user interest conceptual model can easily mine user interests from historical records, and conveniently modify and refine the user interest model. The model uses a series of keywords to describe users' interests. In the vector space model, each keyword corresponds to a weight. This method is simple to use and easy to update the user model. At the same time, it requires that these keywords are orthogonal, and does not describe the real relationship between keywords.

Social network big data application is a process of data insight, including "descriptive analysis", "predictive analysis" and "normative analysis" of cloud computing. It aims to analyze valuable "information", "knowledge" or "wisdom" from massive data, finally achieve the goal of improving the ability to support decision-making, and optimize the data interest recognition ability system, as shown in Fig. 2:

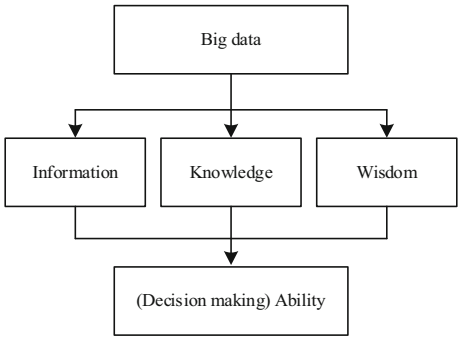

Fig. 2. Improvement of data interest recognition ability based on Cloud Computing

Therefore, for developers and researchers, the goals of social network big data application can be summarized as follows: achieve "insight" based on cloud computing analysis and build a new individual or organizational knowledge structure system; Create or improve the decision-making ability of individuals and organizations [4]. Combined with the goal of social network big data, this paper attempts to decompose the application

process of social network big data in stages and construct the application process model, so as to explore the internal evolution logic and external influencing factors of cloud computing applications. We can try to use the corresponding theories of information management and knowledge management to describe its process, as shown in Fig. 3:

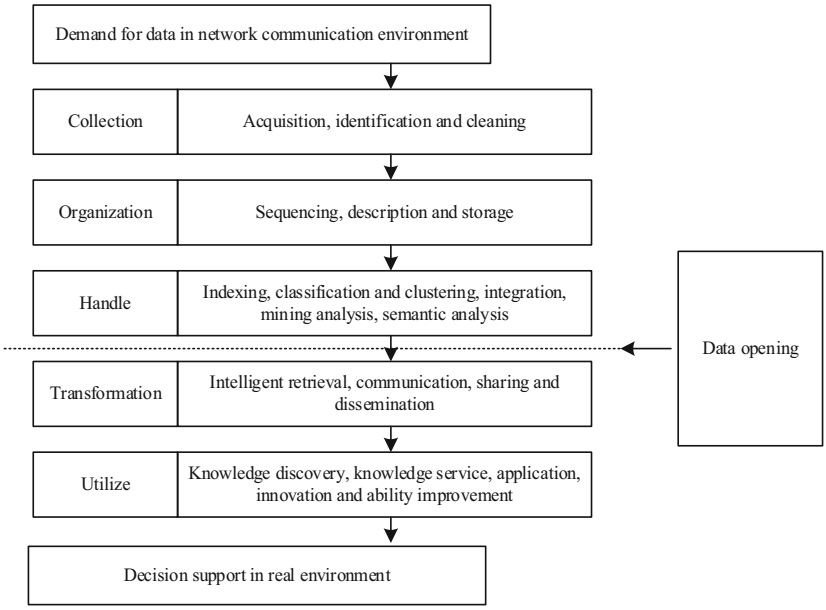

Fig. 3. Process model of network big data acquisition application

In social media, the network language presented in the Pseudo Environment reproduces the cognition and behavior of network users, and realizes the transformation of the subjective reality data presented by users' subjective cognition and behavior into network language symbols [5]. Other users who read these useful data or information will then transform these online language symbols into their own subjective cognition and behavior through reading and thinking, and then produce new continuous offline or online actions, or simply affect cognition, which will have a certain impact on the objective reality. The process of the influence of network communication is shown in Fig. 4.

In the interpretation and interpretation of these data or information, distortion will inevitably occur. For the publisher, when interpreting data or information, the network "non mirror" mimicry propagation leads to the secondary distortion of information. For the receiver, in the process of network data or information interpretation, due to the individual differences of users, there are differences in subjective cognition and behavior, resulting in deviation and secondary distortion. Once propagated again, it will lead to further deviation of distorted information.

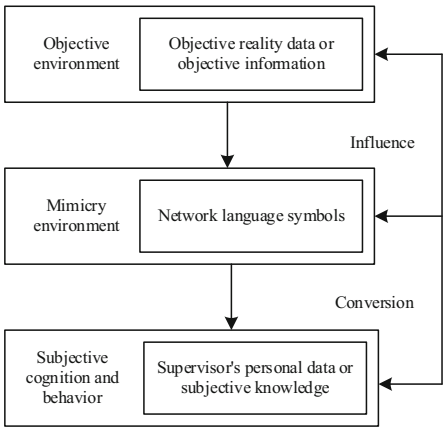

Fig. 4. The process of the influence of network information dissemination

2.2 Social Network Data Storage Evaluation Algorithm

Due to the wide use of the network, the amount of network data is increasing rapidly, but it does not bring the precipitation and accumulation of valuable knowledge, showing a picture of lack of knowledge [6]. Therefore, it is necessary to explore the process and implementation mechanism of knowledge accumulation in the network environment. People carry out information inquiry in cyberspace, which is an information activity consciously engaged in in in order to change the existing knowledge structure. For how to change the existing knowledge structure, this paper puts forward the consistency characteristics of information users' knowledge structure and information needs, and points out that the satisfaction of users' information needs takes self knowledge structure as the starting point. The absorption of information will interact with the recipient's original knowledge structure and produce the recipient's new knowledge structure. The form of Brooks equation is shown in formula (1):

$$K(S) + D_I - \Delta \mathrm{I} \times K(S + D_S) = P \tag{1}$$

In formula (1), $K(S)$ is the user's original knowledge structure; ΔI refers to the amount of information absorbed, that is, the information that can be understood and integrated into their own knowledge structure, and $K(S + D_S)$ refers to the knowledge structure formed after absorbing new information. The flow and transformation of knowledge need to be realized among individuals with similar knowledge structure but some differences [7]. Network data or information is transmitted through the Internet platform. When users with data or information needs obtain, clean up, identify, mine and analyze valuable objective data or information with high matching degree from the data set based on the cloud computing processing platform, they can interact with their original knowledge structure, establish a new knowledge structure and form knowledge innovation. Users can also learn and match the old objective information in the data set with the help of artificial intelligence technology, so as to establish a new knowledge structure and form knowledge discovery. Therefore, this study attempts to expand Brooks equation to

obtain as shown in formula (2):

$$E = P - K_a(S_a) + K_b(S_b) \tag{2}$$

When users acquire and absorb knowledge for multiple data sets or groups of objective information in data sets, multiple knowledge structure units K_a are transformed into multiple new knowledge structure points K_b by absorbing several matching information or knowledge S_a, so as to establish a new general knowledge structure system S_b. The establishment of this new knowledge structure realizes the explicit transformation of tacit knowledge and the implicit transformation of explicit knowledge [8]. According to the characteristic division results of massive data, the massive data is controlled through protocol transformation to realize the real-time collection, analysis and combination of massive data in the process of large-scale transmission. The model construction is shown in Fig. 5.

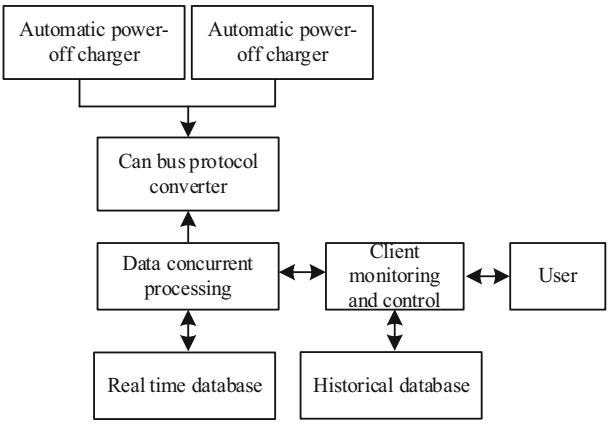

Fig. 5. Construction of data classification management model

Take the server as the core, maintain the connection between the CAN protocol converter and the client, and provide support for data transmission control. Embedded database is to convert all the non current data in the server into the historical database, and read the real-time data of battery and motor through calling and communication interface. There are usually two kinds of data in the historical database: control command data and monitoring command data.

2.3 Implementation of Social Network Data Storage

The social network data storage model is a one-stop platform with spark as the core, including data acquisition, data processing, data mining and data visualization functions. It has good openness, scalability and versatility [9]. It mainly includes four modules: data capture, data preprocessing and storage, data mining and analysis and data visualization. As shown in Table 1, the three types of files collected by wif fence equipment and

their corresponding data scales are counted. These three types of files are hotspot real-time information files, terminal real-time information files and virtual identity real-time information files.

Table 1. Statistics of data scale in big data analysis storage model

File name	Growth rate (estimated value)	Estimated total (3 months)
Hotspot real-time information file	168 million pieces/day; 21.3 GB/day	15.5 billion, about 1909 GB
Terminal real-time information file	136 million pieces/day; 17.35 GB/day	12.5 billion, about 1555 GB
Virtual identity real-time information file	139.9 billion pieces/day; 1.63 GB/day	1.4 billion, about 155 GB

In terms of database, it is mainly divided into relational database and non relational database. Among them, relational database includes traditional OSQL database and new Newsql database; Non relational database mainly refers to NOSQ database, which is divided into key value storage database, column storage database, document database and graphic database [10]. Due to the capture limitation of social network API, this paper collects user data by means of distributed capture and analysis of social network web pages when implementing this module. In the capture process, it is not only necessary to simulate the login operation of social network browser, but also need to schedule and manage the tasks of distributed crawler 28. The distributed social network crawler is based on Actor model and implemented by Scaa and Akka. The data storage management process is shown in Fig. 6:

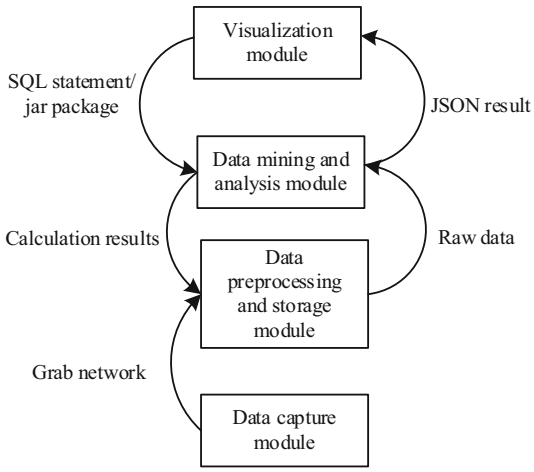

Fig. 6. Data storage management flow chart

The ultimate purpose of establishing user interest model is to realize the accurate push of information service. According to the established multi-level user interest model, the preference description of users on social networking sites can be made, and the user's personal preference information business card can be made, so as to realize the marking of information delivery objects, better meet the personalized needs of social networking users and improve the success rate of push. The hierarchical classification storage model of social networking data is shown in Fig. 7.

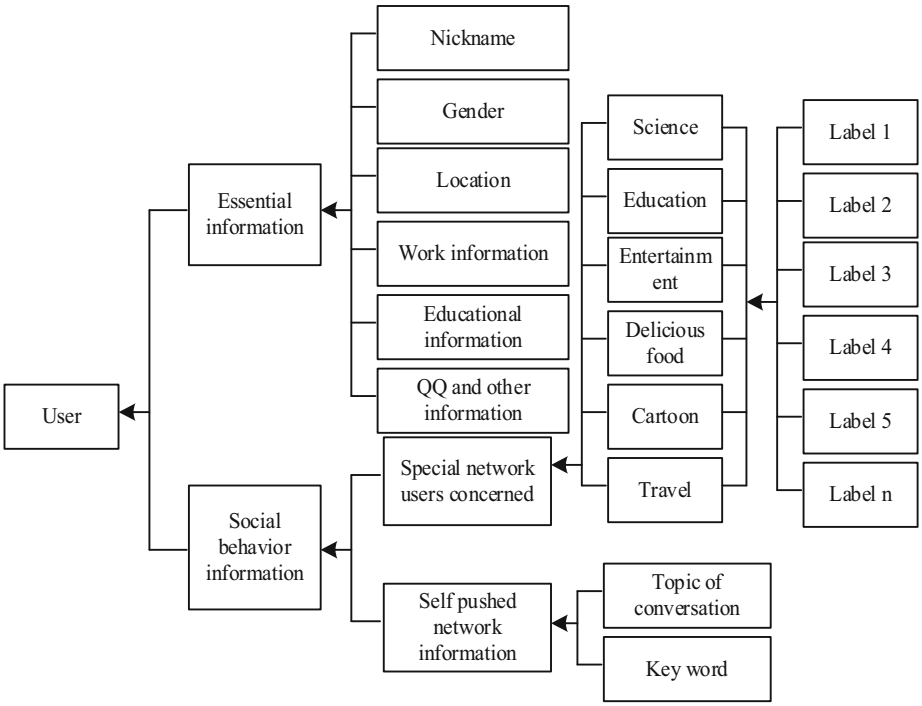

Fig. 7. Hierarchical classified storage model of social network data

In order to speed up the data loading speed in the web page, we also choose to use Reds for data caching, because the web page displays the relevant data and algorithm results of the social network big data analysis platform. These data will not expire within a certain period of time. If there are multiple same data access requests in a short period of time, sending the request to Hive every time is a waste of computing resources. Therefore, the data visualization module in this paper sets the validity period for all visual data when designing. Every time the web page requests background data, it will first go to Reds to find out whether the data is within the validity period. If it is valid, it will be returned directly from the cache: if it fails, route the request to hive, wait for the result to return, and cache the result into Reds again, so as to achieve the goal of classified storage of massive data.

3 Analysis of Experimental Results

Due to the huge amount of data to be processed in the experiment, the social network user forwarding behavior prediction algorithm needs to rely on the fast computing capacity of the social network big data analysis platform when it is implemented. Therefore, the method used in the algorithm to optimize the prediction model under the multi task learning framework has been written and compiled into the algorithm library of the data mining and analysis sub module.

During the experiment, eight sub nodes in the platform provided computing services. Each machine provided 4-core CPU and 10 GB memory. HDFS with a total capacity of 2.5 tb provided file storage services. Taking 20148.31 as the time dividing point, the experimental data set is divided into experimental training set and experimental test set. Randomly select a social network user W and acquire the interest of the social network user according to the above method. Extract the following user list of the social network user W, and obtain the tags of the special social network user w pays attention to (according to the tag tree method), so as to obtain the initial interest tag set of W; Calculate the weight of all interest tags according to the number of sub levels of interest tags. Calculate the value of each interest tag. In addition, the number of forwarded social networks is about twice that of original social networks. Therefore, in order to ensure data balance, we need to sample the original social networks and forwarded social networks, with a sampling ratio of 1:2. Table 2 shows the data sets actually used in the experiment.

Table 2. Data set properties

Number of network users	Number of network forwarding	Number of network originality	Number of concerns
92,068	716,185	9,197,365	1,273,852

Table 3. Characteristics of data sets used

Content	Number of network forwarding	Number of network originality
Experimental training set	589,252	1,169,856
Experimental test set	128,365	250,365

In order to verify the performance improvement of the proposed prediction model, logical regression (LR), support vector machine (SVM) and passive aggressive algorithm (PA) are selected as control algorithms to train and verify on the data set respectively (Table 3).

Table 2 shows the overall experimental results with accuracy and recall as indicators, in which the data storage management behavior prediction algorithm represents the algorithm proposed in this paper (Table 4).

Table 4. Experimental results

Index	PA	Logistic regression	Support vector machine	Forwarding behavior prediction algorithm
Accuracy	0.755	0.865	0.878	0.895
recall	0.389	0.468	0.508	0.522

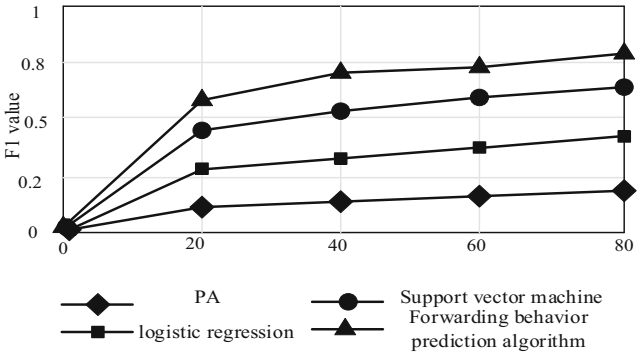

Fig. 8. F1 value of experimental results

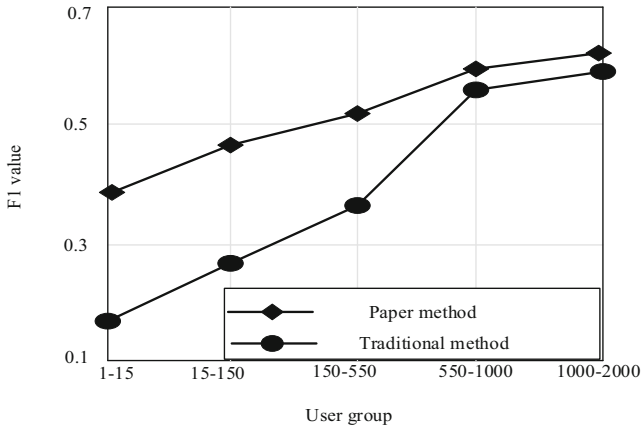

Fig. 9. Comparison of user set data storage

According to the experimental results in Fig. 8 and Fig. 9, on the one hand, because the PA algorithm is too simple, its accuracy and recall are relatively low; On the other hand, compared with the logistic regression algorithm as the benchmark algorithm, the algorithm proposed in this paper can significantly improve the accuracy and recall. In addition, compared with the previous two linear algorithms, although the method in this paper has a better classification effect, the overall classification effect is not as good as

the algorithm proposed in this paper because the comparison algorithm is all based on the model trained by global data, and the algorithm proposed in this paper introduces local parameters to improve the classification effect. Through the comparison of F1 values in Fig. 8, it can be seen that the algorithm proposed in this paper can obtain better F1 values. In order to determine the impact of local parameters on the model, the benchmark algorithm logistic regression is used as the control algorithm in the experiment. In order to verify the classification of the model proposed in this paper under different forwarding historical data, users are divided into five groups according to the number of social networks: 10, 10–100, 100–500, 500–1000 and 1000–2000. One user is extracted from each group for 20 times. At the same time, the method in this paper with good results is used as a representative to compare with the algorithm proposed in this paper. The results are shown in Fig. 10:

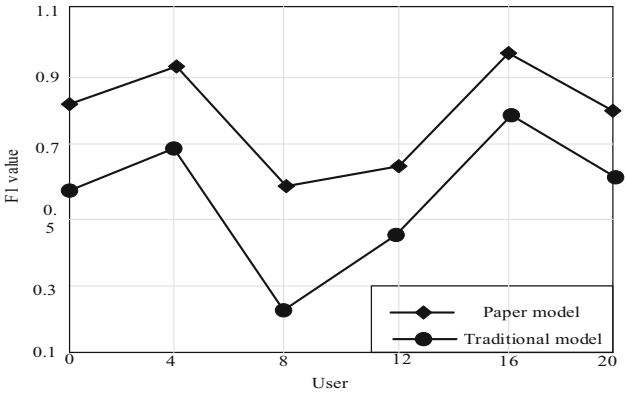

Fig. 10. Comparison of random user set data storage

As can be seen from Fig. 10, the predicted F1 value of the model proposed in this paper is much smaller than that of the logistic regression model, and it can also get better results on several users with poor performance of the logistic regression model, which shows that the model can meet the requirements of individuals.

To sum up, the model in this paper can significantly improve the accuracy and recall of the two performance indicators, and the fluctuation of the predicted F1 value for each user is much smaller than the logistic regression model.

4 Conclusion

The research of this paper is to build a social network crawler as a data source, use cloud computing to store the captured data in a distributed manner, use the fast computing ability of cloud computing to process the stored data, and finally use js visualization tools to display the data and related algorithm results on the web page, so as to finally realize a social network big data analysis platform with cloud computing as the core that can complete this whole process. At the same time, a prediction algorithm for the forwarding

behavior of social network users is proposed for the captured social network data. In order to improve the performance of the algorithm, a multi task learning framework is also introduced. Through the obtained prediction model of forwarding behavior of social network users, the new social networks received by users can be predicted, and these social networks can be reordered so that users can easily view the social networks they are interested in. In addition, the algorithm runs on the social network big data analysis platform based on cloud computing, which verifies the effectiveness of the platform and the scalability of the algorithm module.

References

1. Devi, K., Muthusenthil, B.: Intrusion detection framework for securing privacy attack in cloud computing environment using DCCGAN-RFOA. Trans. Emerging Telecommun. Technol. **33**(9), 4561 (2022)
2. Meng, Q., Liu, B., Zhang, H..-Y.., et al.: Multi-relational group influence modeling and analysis in online social networks. Chin. J. Comput. **44**(06), 1064–1079 (2021)
3. Liu, X., He, D.: Research on of competitive nonlinear dynamic information diffusion modeling in online social network. Chin. J. Comput. **43**(10), 1842–1861 (2020)
4. Lian, J., Fang, S., Zhou, Y.: Model predictive control of the fuel cell cathode system based on state quantity estimation. Comput. Simul. **37**(07), 119–122 (2020)
5. Xu, Z., Zhu, D., Chen, J., et al.: Splitting and placement of data-intensive applications with machine learning for power system in cloud computing. Digit. Commun. Netw. **8**(4), 476–484 (2022)
6. Thabit, F., Alhomdy, S., Jagtap, S.: A new data security algorithm for the cloud computing based on genetics techniques and logical-mathematical functions. Int. J. Intell. Netw. **2**(2), 18–33 (2021)
7. Li, Y., Zhou, F., Xu, Z.: Privacy-preserving k-nearest-neighbor search over mobile social network. Chin. J. Comput. **44**(07), 1481–1500 (2021)
8. Ramdani, F., Wirasatriya, A., Jalil, A.R.: Monitoring the sea surface temperature and total suspended matter based on cloud-computing platform of google earth engine and open-source software. IOP Conf. Ser. Earth Environ. Sci. **750**(1), 12–31 (2021)
9. Luo, H., Yan, G., Zhang, M., et al.: Research on node importance fused multi-information for multi-relational social networks. J. Comput. Res. Dev. **57**(05), 954–970 (2020)
10. Xu, M., Zhang, Z., Xu, X.: Research on spreading mechanism of false information in social networks by motif degree. J. Comput. Res. Dev. **58**(07), 1425–1435 (2021)