Data Clustering Mining Method of Social Network Talent Recruitment Stream Based on MST Algorithm

Hongjian Li(✉) and Nan Hu

China University of Labor Relations, Beijing 100048, China
abrajim@sohu.com

Abstract. In order to solve the problem that the data clustering mining method of social network talent recruitment stream is affected by the score of graph area and has a long time of index updating, a data clustering mining method of social network talent recruitment stream based on MST algorithm is designed. Based on the six-degree segmentation theory, the features of social network talent recruitment are extracted, the flow computation framework is established, the recruitment data processing process is optimized, and the similarity coefficient is used as similarity measure to construct the flow data clustering model and the mining pattern is designed by using the MST algorithm. Experimental results show that the maximum update time of the proposed method is 16.638 ms, which shows that the proposed method can shorten the update time of the index and is of high value.

Keywords: MST algorithm · Social network · Talent recruitment · Stream data · Clustering · Data processing

1 Introduction

In the early research of stream data clustering, the problem of stream data clustering was regarded as a special case of "large database" clustering problem. This kind of early algorithm belongs to single-layer algorithm structure in frame structure. There are many examples of streaming data, such as: in the business world, large warehousing supermarket transaction data. The supermarket chain's data center collects a large number of transactions from each store every day, each of which includes attributes such as customer purchases and consumption amounts, in chronological order. Single-layer algorithm structure tries to transform the dynamic characteristics of data flow into the traditional static mode, so that it can apply more mature traditional methods to solve the problem.

If each transaction completes and records can be collected immediately by the data center, the recorded data will flow to the data center continuously in the time dimension from the data abstraction point of view. At this stage, the focus of the algorithm research is to improve the performance of the traditional algorithm to adapt to the dynamic characteristics of the stream data.

W. Fu and L. Yun (Eds.): ADHIP 2022, LNICST 469, pp. 99–111, 2023.
https://doi.org/10.1007/978-3-031-28867-8_8

Small space, the algorithm is the representative of this kind of algorithm, it uses the improved K-center point algorithm to make it can be applied in the new problem domain. In the telecom industry, the mobile company will collect the user's call record every time. These records also include a number of attributes, such as calling number, called number, call time, the amount of money and so on. This paper analyzes the requirements of stream clustering model, and summarizes some clustering models that may be suitable for data stream. It is considered that there are three requirements to be satisfied in data stream clustering: (1) compressed expression, (2) rapid and growing processing of new data points, and (3) rapid and clear determination of outliers. In this paper, a two-layer algorithm framework for stream data clustering is proposed. The algorithm is divided into two parts: online layer and offline layer. Large numbers of call logs are arranged in chronological order, pooled in a mobile company's data center, and can be abstracted as a "stream." The online layer algorithm is responsible for fast and simple processing of the stream data and generating the profile data structure. In addition, a series of physiological signals, such as heartbeat, pulse and blood pressure, are transmitted to the analysis module in real time to predict the patient's health status at each time. Line layer algorithm makes use of these summary data information for more complex analysis, and generates more accurate clustering results. In addition to its theoretical value, the study of data stream clustering is of great practical significance. Many practical problems can be solved by clustering the data stream [1, 2]. In the industrial production, some large equipment safety testing instruments will collect the operation parameters of the equipment at this time at every moment, as a data signal for analysis and processing. And so on, these are examples of streaming data. For example, in network monitoring, the network state is analyzed by clustering the network flow, which provides a reliable basis for improving the network performance. It is not hard to see that these data patterns all have some common features. First of all, the volume of data is very large, the number of these data over time to rise sharply, such as Wal-Mart supermarket import and export transactions can reach hundreds of thousands of times. Dynamic analysis of user call records recorded on telecom switches using data flow clustering technology can help balance the load. Clustering analysis of the visit records of large websites can provide a basis for the decision-making of websites.

However, in the process of data processing, the above data stream clustering mining method is affected by the score of graph area, which has the defect of a long index update time. In order to simplify the recruitment data processing process and reduce the index update time, this paper proposes a social network talent recruitment flow data clustering mining method based on the MST algorithm. Innovatively based on the six degree segmentation theory, extract the characteristics of social network talent recruitment, establish a process calculation framework, and optimize the recruitment data processing process; The streaming data clustering model is constructed, and the MST algorithm is used to design the mining pattern, which improves the locality of calculation and realizes efficient data mining. Experimental results show that this method can reduce the update time.

2 Data Clustering Mining Method of Social Network Talent Recruitment Stream Based on MST Algorithm

2.1 Extracting the Recruitment Characteristics of Social Network Talents

Due to the characteristics of no time and space restrictions, enterprises and recruiters can browse recruitment information on electronic devices anytime, anywhere, communicate with each other, and strengthen the information exchange between the recruitment and the applicants. For employers and job seekers, online recruitment is conducive to both sides to use the least time in the broadest scope to find the right talent or position.

Social network recruitment is based on social network, is the application of social network in the field of human resources recruitment, refers to the enterprise recruitment activities based on social network platform. The breakthrough time limit and geographical scope of the Internet can quickly connect users around the world, so that job seekers can see the recruitment information of enterprises on the Internet platform, and can carry out rapid and two-way interaction between users in different regions and countries at any time. This feature of the Internet is more conducive to the recruitment units to find their potential suitable candidates. The enterprise establishes the enterprise talent database based on the social network platform, as well as the enterprise carries on the staff recruitment based on the social network platform release recruitment information. However, the traditional recruitment channel is restricted by time and region, and the network recruitment is not limited by time and space to make it possible to achieve a reasonable flow of talent. Meanwhile, social network recruitment refers to the process of employing social network service platforms such as Microblog and WeChat as recruitment channels.

Social networking is based on the "six-degree segmentation theory," which assumes that the world is unfamiliar to everyone and that it takes just six people to make a connection. The "six-degree separation" emphasizes the important role of "weak relationship" in interpersonal communication. The characteristics of the Internet, such as "no time and space limitation, interactive and real-time", are favored by enterprises. They will also use the official website of enterprises to show the corporate image to job seekers in an all-round and multi-level way, so as to provide job seekers with a true way to understand the enterprises and help job seekers choose suitable recruitment positions. Social networks move offline relationships into the network, widening the range of social interactions and creating a large network of relationships. But social network recruitment relies on the unique advantage of social network to play a strong role of weak relationship. Most enterprises can select suitable talents through online recruitment, the openness and rapid proliferation of online recruitment information for enterprises and candidates to provide more options.

The network regarding the present society, may say is the content rich, the function is formidable, makes the contribution for the enterprise employment advertise effective promotion. Social network recruitment is mainly divided into three parts, namely, personal management, recruitment information and networking. Social networking sites target job seekers through the management of resume information on personal platforms. Network information system, when dealing with a large number of resume information of job seekers, uses its own powerful function of data processing to sort out, screen

and analyze more quickly and accurately, thus establishing a powerful talent database for enterprises and providing a basis for the future talent vacancy. Based on the social network, the candidate can know the relevant information of the enterprise in real time, and match the information published by the enterprise with his ability and intention to find a satisfactory job. Candidates can also learn about the culture and values of the company through information posted on social networking sites. In the recruitment website, the enterprise recruitment information and talent supply and demand information is updated in time to facilitate enterprises and recruiters choose at any time. Enterprises can select job candidates according to search engines, and the function of automatic classification helps to find suitable job candidates quickly and effectively, and feed back to job candidates in time, thus improving the efficiency of recruitment. The recommendation function of the social network recruitment platform, based on the matching of the personal information of the candidates and the recruitment information of the enterprises, can recommend the highly matched candidates to the recruitment enterprises. The internet recruitment can improve the ability of information collection, analysis and processing, reduce the time and manpower cost of manual information collection, so as to reduce unnecessary capital waste and make the recruitment more effective.

Social networks are also able to recommend well-matched companies to job candidates, thereby enabling efficient communication and matching of information among candidates, social networks and recruitment companies, laying a solid foundation for person- and organization-matching, as shown in Fig. 1:

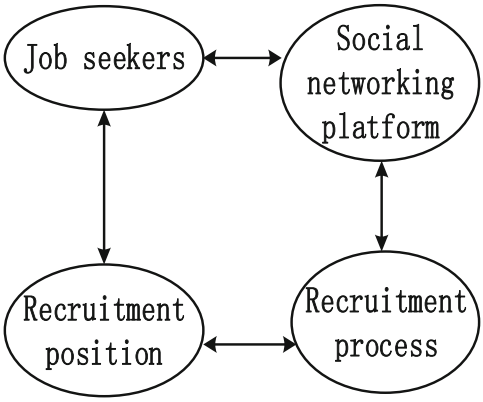

Fig. 1. Recruitment match diagram

As can be seen from Fig. 1, social network recruitment uses social network platform to realize the understanding, communication and interaction of the information of both sides, thus accomplishing the recruitment goal. Online recruitment is based on the initiative of both employers and job seekers to communicate on the Internet, which can not only improve the timeliness of online recruitment, but also provide good opportunities for job seekers to understand the enterprises and positions. Therefore, this kind of online recruitment is being accepted and used by more and more enterprises and job seekers.

Specifically, enterprises use social platforms to implement recruitment, big data technology as a support to analyze whether the characteristics and capabilities of candidates match the organization, thus enhancing the effectiveness of recruitment. Different from the traditional way of recruitment, it does not demand the absolute consistency of time and space, which facilitates the choice of time between the two sides, and saves a lot of time.

2.2 Optimize Recruitment Data Processing Process

The prominent feature of social network is that the network coverage is divergent, and the network technology makes the network between people become larger and larger, so as to enlarge the user's network. High throughput, low latency, good scalability and stable running system is an ideal system for stream data mining, which can not be realized without reasonable design and planning of port architecture, programming interface and high available technology.

Social network recruitment based on social network platform is to make use of the wide network of social network platform, which can make the information spread sufficiently and make the best use of social resources, and provide continuous and sustainable impetus for the career development of job seekers and the long-term development of enterprises. The so-called system architecture is the combination of various subsystems in the computing platform. The mining of stream data is inseparable from the specific streaming computing architecture. There are one master node and several slave nodes in the master-slave computing architecture. The master node is responsible for task allocation, resource scheduling, load balancing and fault tolerance. For enterprises, in the network of social networks with the social platform to find the right person can achieve twice the result with half the effort. The scheduling of the whole port depends entirely on the master node, and the slave node receives the tasks assigned by the master node and completes them respectively. In the symmetrical architecture, there is no master node for scheduling, and each node has the same function and good scalability. Resource scheduling and load balancing need to be coordinated by related distributed protocols. For job seekers, can make up for their own narrow circle of the shortcomings of the full received a large number of external recruitment information, in order to find suitable positions to pave the way. Job seekers in the production of resumes and interviews, the presentation is not necessarily the real daily state, there will be deliberate performance of the phenomenon.

Generally speaking, in the streaming data environment, the programming interface that uses directed task topology to represent the relationship between computing tasks is designed to facilitate programmers to implement the functions in the task topology. But by browsing the homepage space of the job seeker on the social networking platform, through a large number of data analysis to obtain the inner information of the job seeker's friends and personality traits. Unlike batch computing, where data is stored on a persistent device in advance, large data streaming does not allow data to be similarly persisted. The main processing flow of the recruitment data is shown in Fig. 2:

As can be seen from Fig. 2, the main processing flow of recruitment data includes three steps: data selection, data preprocessing and data transformation. Using MST algorithm, we can analyze the interpersonal communication of job seekers and understand

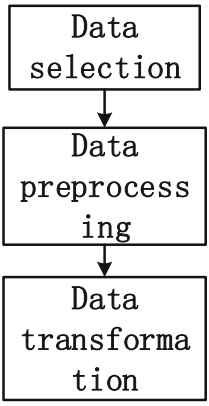

Fig. 2. Main processing flow of recruitment data

the quality and word-of-mouth image of job seekers indirectly through the circle of friends. Social networking platforms provide job seekers with a freer and fuller space to express their opinions and choices than a stylized resume. Therefore, it is not easy to replay data after node failure in the big data streaming environment, and the high availability technology of batch computing can not be fully applied to the streaming environment. That is to say, social network recruitment can make the human resource managers know more about the personality, interest and experience of the candidates, and help them to match the positions offered by the enterprises, which improves the success rate of recruitment, at the same time, reduces the turnover rate to a greater extent. In data mining, the main task of classification mining is to discover value more accurately through learning. An algorithm of classification mining can be divided into two main parts: building model and using model to predict. The classification algorithm for stream data can build decision tree and mine incremental data, and the incremental data mining is also dynamic.

2.3 Build Streaming Data Clustering Model

Using the existing initial data to build a clustering model, the data to be classified will continue to arrive, the classification model can be used to classify these objects. Clustering is the process of dividing data objects into different subsets. In the application of data mining and data analysis, it is often necessary to quantitatively express the differences between data points or between individuals, so as to evaluate the similarity of data points or individuals and their categories [3–5]. There will also be an endless stream of data to update the decision tree. The data to be predicted in the future will be predicted by the latest decision tree.

Using the existing initial data to build a clustering model, the data to be classified will continue to arrive, the classification model can be used to classify these objects. Clustering is the process of dividing data objects into different subsets. In the application of data mining and data analysis, it is often necessary to quantitatively express the differences between data points or between individuals, so as to evaluate the similarity

of data points or individuals and their categories [3–5]. There will also be an endless stream of data to update the decision tree. The data to be predicted in the future will be predicted by the latest decision tree.

$$L(p,q) = \sum_{\beta=1} |p - q_\beta|^{\alpha-1} \tag{1}$$

In formula (1), p represents the data set, q the data space, β the data distance difference, and α the data dimension.

Therefore, the calculation of similarity is an important index in the research field of data mining. Since the properties of clustering model are divided according to similarity, it is very important to calculate the accuracy of similarity. This type of data, produced with the growing reach of the Internet and the widespread use of automatic digital devices, goes beyond the limits of existing information systems that can be persisted, processed accurately, and accessed repeatedly. Set the fixed clustering parameters in advance, and then get the final result according to the parameters. In addition, we can set the threshold to distinguish the strength of similarity through the threshold, and then carry out the subsequent clustering mining process [6, 7].

In clustering model, the selection of similarity measure is often represented by distance measure, which measures the difference and similarity between data points. Data matrix can be used to represent the characteristics of the corresponding data, through the matrix list to declare the corresponding attribute structure, each row in the matrix represents a data. The structural characteristics of t data can be represented by the following matrices T:

$$T = \begin{bmatrix} t_{11} & \cdots & t_{1x} \\ \vdots & \ddots & \vdots \\ t_{y1} & \cdots & t_{xy} \end{bmatrix} \tag{2}$$

In formula (2), t represents the number of clustering objects, x represents the attributes of data, and y represents the structural characteristics of data. Stream data has many characteristics: the data arrives in real time, the data flow speed is unpredictable, and it requires fast and instant response. The data changes over time, on a vast and unbounded scale. Among many clustering models, Euclidean distance measure is usually used to measure the attribution degree of the target point and the cluster. However, large datasets are often accompanied by higher dimensions.

On the basis of formula (2), if distance is used as the similarity, then the value on the diagonal line is 0, if similarity coefficient is used as the similarity measure, then the value on the diagonal line is 1, and the formula of the transformation matrix T' of the matrices T is as follows:

$$T' = \begin{bmatrix} r_{11} & & \\ \vdots & \ddots & \\ r_{x1} & \cdots & r_{xx} \end{bmatrix} \tag{3}$$

In formula (3), r represents the horizontal columns of the matrix that represent the corresponding data points. Data processing is basically single-pass scanning algorithm,

once the data is difficult to be taken out again. The degree of data structuring is low, which needs multilevel and multidimensional processing. As the dimension of the set of numbers to be clustered increases, the distance between two points will become smaller and smaller, and the farthest neighbor and the nearest neighbor will be almost the same. Therefore, Euclidean distances are no longer suitable for measuring large, high-dimensional data sets.

Based on the above principle, assuming that the h central point is known, the $h + 1$ central point can be selected by the following principle $G_{(h+1)}$. Can be expressed as:

$$G_{(h+1)} = \frac{1}{\max\left[g(d,h)^2\right]} \tag{4}$$

In formula (4), g represents the shortest distance between the candidate center point and the former h center point, and d represents the maximum value of the shortest distance. The results show that the similarity of objects in clusters and the difference of objects in clusters are evaluated according to the attribute values of the objects, usually involving distance vectors.

Clustering may find previously unknown groups in the data object. On the other hand, the length of the data stream is theoretically considered to be infinitely long, and its range is also considered to be infinite. For example, if a router only routes 10,000 different IP address pairs, then despite the large number of IP packets, there is no challenge for querying IP address pairs whose traffic is greater than a certain threshold. By comparison, Manhattan distance is a better representation of the actual similarity between data than Euclidean distance in high dimensional space. But when the IP address pair space is far beyond the real storage, the small number of IP packets also makes many queries difficult.

On the same dataset, different clustering methods may produce different clustering. Clustering analysis is not divided through the people, but through the algorithm automatically. Finally, the infinity of stream data makes it impossible for stream data mining to retain the original data, but can only maintain a series of profiles in memory, and generate the final results based on the profiles.

2.4 MST Algorithm Design Mining Patterns

The MST is constructed in each grid, and then the MST of each grid is connected to obtain the MST algorithm of the original dataset. Then the clustering is completed. The process of data mining is a process of discovering various models, summaries and derived values from a given set of data. The main stages involved include: data preparation, data mining, result representation and interpretation. The mining method based on MST is a significant research field in the mining method based on graph theory. MST is an important research direction of connectedness in graph theory. It has many excellent structural properties in the representation of set data, and has been widely used in many fields. Data selection mainly refers to extracting the relevant data from the existing database or data warehouse to form the target data. Data preprocessing is to merge the data, to solve the semantic fuzziness, data processing omission and cleaning dirty data, etc.

In the MST -based clustering algorithm, firstly, the relationship among all the samples in the data set should be represented as the graph structure of MST. Among them, the node of graph structure is all the sample points in the data set, and the edge of graph structure is the similarity measure between the two connected sample points. The aim of da transformation is to eliminate the dimension of data, that is to find out the really useful features from the initial features, narrow the processing range and improve the quality of data mining. Then, according to the given weight closed value or the number of clusters, delete the edges with the largest weight in turn, and obtain the connected subgraphs of the graph structure, each of which represents a cluster.

Data preparation is an important step of data mining. Whether the data preparation is good or not will affect the efficiency and accuracy of data mining and the effectiveness of the final pattern. This stage is the actual work of mining, first of all to determine the task of mining or what is the goal, such as data summary, classification, clustering, association rules or sequential pattern discovery. Among them, the nodes belonging to the same subgraph have the greatest similarity, and the similarity with other subgraph nodes is greater than the nodes within the subgraph. In general, for a given dataset, the MST generation algorithm is not unique for each execution because there may be two or more edges of the same length. According to the task of mining, the choice of algorithm is the most important step, and the choice of algorithm directly affects the quality of mining.

The extracted information is analyzed according to the end user's decision goal, and the most valuable information is distinguished and submitted to the decision maker by the decision support tool. The social network talent recruitment stream data clustering mining process is shown in Fig. 3:

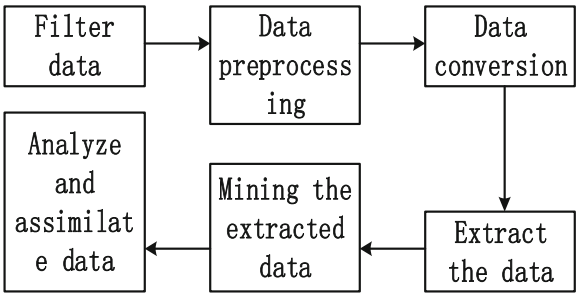

Fig. 3. Social Network Talent Recruitment Flow Data Clustering Mining Flowchart

As can be seen from Fig. 3, the process of data clustering and mining in social network recruitment flow mainly includes: data selection mainly refers to extracting relevant data from existing databases or data warehouses to form target data. Data preprocessing is to merge data, to solve semantic fuzziness, to deal with the missing data and dirty data in the data.

However, the non- uniqueness of MST constructs on the same dataset does not affect the clustering result based on the minimum generated cluster, because the generated cluster is independent of the specific MST structure. The clustering algorithm based on

MST has its own unique advantages. The task of this step is not only to present the results (e.g., using information visualization), but also to filter the information. If the decision makers can not be satisfied, we need to repeat the above data mining process. MST is an important data structure in graph theory, and it has good geometric properties.

The essence of MST is the minimal cost spanning tree, i.e. the weight of all the edges and the minimal connected subgraph in the graph structure. Mining the web data after data preprocessing, using various mining technologies to mine the rules and patterns hidden behind the data, and according to the specific application, filter out the rules or patterns that are not used in the pattern mining stage, and transform the useful rules and patterns into knowledge and apply them to the specific field. So that the structural characteristics of the data set is more prominent, segmentation MST clustering results produced more scientific, rational and easy to operate. The performance of the algorithm is not limited by the shape of clusters, and irregular classes can be identified.

3 Experimental Study

3.1 Experimental Preparation

In order to test the effectiveness of the proposed method, we design a comparative experiment in the environment of Eclipse.

The experimental environment is a cluster environment composed of 16 computing nodes. Among them, the resource configuration of each machine is as follows: CPU: Intel (R) Core (TM) i 3-2100 @ 3.l0 GHz (dual core), memory: Apacer 4G DDR3 $\times$ 2, hard disk: Hitachi SOOG/7200 RPM operating system: Red Hat Linux 6.1 JDK version: jdk -1.6.0-30. The experiment used two real datasets, the Live Journal social network (SOC) (default data set) and the Texas Road Network (TX). Based on the MATLAB R2015a program, using the MST generation algorithm, we input two real datasets, randomly generate the weights of each edge, and the distribution of each edge weight satisfies the uniform distribution mining pattern. Based on the weight results, the cluster mining results are computed.

In addition, all three comparison methods use 75% of the data in the dataset to initialize, randomly manipulate the whole dataset and simulate the generation of stream data. Based on the initialized microclusters, online clustering mining is carried out for the data generated by simulation (sending 1500 data per second).

3.2 Experimental Results

For the same dataset, we use the clustering mining method based on genetic algorithm and the clustering mining method based on deep learning to compare the two methods. As the number of graph partitions is larger, the index updating process is more complex and the updating time is longer. Therefore, taking the number of graph partitions as the test index, the maximum update time of the three methods is tested when the number of graph partitions is different. The experimental results are shown in Tables 1, 2 and 3:

According to Table 1, for area score index of 4 maps, the maximum update time of the social network talent recruitment flow data clustering mining method in this paper is

Table 1. Maximum update time for area score index of 4 maps (ms)

Number of experiments	Data clustering mining method of social network talent recruitment stream based on genetic algorithm	Data clustering mining method of social network talent recruitment stream based on deep learning	Data clustering mining method of social network talent recruitment stream
1	16.514	15.699	12.141
2	15.334	16.347	11.547
3	16.254	16.925	10.698
4	17.121	16.548	12.146
5	16.992	15.377	11.214
6	15.847	16.201	12.096
7	16.224	15.649	13.147
8	17.131	17.003	12.588
9	16.255	16.528	13.164
10	15.288	16.945	12.410

Table 2. Maximum update time for area score index of 8 maps (ms)

Number of experiments	Data Clustering Mining Method of Social Network Talent Recruitment Stream Based on Genetic Algorithm	Data Clustering Mining Method of Social Network Talent Recruitment Stream Based on Deep Learning	Data Clustering Mining Method of Social Network Talent Recruitment Stream
1	19.488	21.004	14.512
2	18.554	19.822	16.528
3	23.516	21.225	15.494
4	21.334	22.313	15.461
5	23.818	24.164	14.332
6	22.156	23.255	13.520
7	24.549	24.818	14.511
8	22.616	25.191	15.260
9	19.477	24.462	14.220
10	18.463	23.646	13.744

13.164 ms, the maximum update time of the genetic algorithm based social network talent recruitment flow data clustering mining method is 17.131 ms, and the maximum update

Table 3. Maximum update time for 16 chart area score indexes (ms)

Number of experiments	Data clustering mining method of social network talent recruitment stream based on genetic algorithm	Data clustering mining method of social network talent recruitment stream based on deep learning	Data clustering mining method of social network talent recruitment stream
1	36.154	35.164	22.314
2	32.166	34.917	23.487
3	31.255	33.626	22.619
4	32.848	32.588	24.466
5	33.649	34.649	23.718
6	32.717	34.718	24.513
7	32.478	33.644	22.462
8	33.915	32.502	23.647
9	32.477	33.784	22.164
10	33.025	34.519	21.005

time of the deep learning based social network talent recruitment flow data clustering mining method is 17.003 ms; According to Table 2, for area score index of 8 maps, the maximum update time of the social network talent recruitment flow data clustering mining method in this paper is 16.528 ms, the maximum update time of the genetic algorithm based social network talent recruitment flow data clustering mining method is 24.549 ms, and the maximum update time of the deep learning based social network talent recruitment flow data clustering mining method is 25.191 ms; According to Table 3, for 16 chart area score indexes, the maximum update time of the social network talent recruitment flow data clustering method in this paper is 24.513 ms, the maximum update time of the genetic algorithm based social network talent recruitment flow data clustering mining method is 36.154 ms, and the maximum update time of the deep learning based social network talent recruitment flow data clustering mining method is 35.164 ms. This is because this method is based on the six degree segmentation theory, extracts the characteristics of social network talent recruitment, establishes a process calculation framework, optimizes the recruitment data processing process, and improves the mining efficiency; At the same time, the similarity coefficient is used as the similarity measure to construct the stream data clustering model, and the MST algorithm is used to design the mining pattern, which improves the locality of calculation and eliminates unnecessary time overhead.

4 Conclusion

By using the distributed index technology, the proposed method can make sure that two vertices belong to the same connected component, and only need several communication operations, not need to detect. At the same time, this paper analyzes the data mining technology, the related concepts and characteristics of stream data, and the key technologies of stream data mining. Aiming at the cluster mining in data mining, this paper analyzes the related concepts and challenges of cluster and stream clustering. In addition, the use of zoning acceleration technique effectively improves the locality of the computation and eliminates the cost of unnecessary communication. Future research will focus on improving the accuracy of mining methods on the basis of ensuring low mining time.

References

1. Cheng, H., Liao, Z., Wang, S.: Simulation of mixed attribute data clustering mining based on feature selection. Comput. Simul. **37**(7), 399–403 (2020)
2. Zheng, L., Zhang, H.: Big data clustering mining technology based on swarm intelligence algorithm in cloud environment. Mod. Electron. Tech. **43**(15), 115–118 (2020)
3. Gu, D.: Large data clustering mining based on P-WAP in Hadoop cloud platform. J. Changchun Normal Univ. (Nat. Sci.) **39**(5), 29–35 (2020)
4. Hua, T., Yi, H.: Big data mining based on around-centroid clustering algorithm. Appl. Res. Comput. **37**(12), 3586–3589 (2020)
5. Zang, Y., Xie, L., Zhang, Y., et al.: Data mining algorithm based on power marketing clustering analysis. Inf. Technol. **44**(4), 56–59, 64 (2020)
6. Li, X., Wu, X., Tong, B.: The research on data mining based on dynamic fuzzy clustering—taking the comprehensive strength analysis of Anhui city as an example. J. Guiyang Coll. (Nat. Sci.) **15**(1), 52–57 (2020)
7. Jin, H.: Research on artificial bee colony clustering data mining algorithm for accurate prediction. Digit. Technol. Appl. **38**(10), 95–97 (2020)