Hindawi
Journal of Robotics
Volume 2022, Article ID 8286007, 6 pages
https://doi.org/10.1155/2022/8286007

Research Article

Implementation of a Community Data Processing System Based on Data Mining

Li Li

School of Philosophy and Public Administration, Henan University, Kaifeng, Henan, China

Correspondence should be addressed to Li Li; 30040056@vip.henu.edu.cn

Received 31 October 2021; Revised 20 December 2021; Accepted 23 December 2021; Published 15 February 2022

Academic Editor: Shan Zhong

With the development of social information technology, intelligent community as a new way of life is changing people's lives step by step. The community as a unit of residence in China has developed quite maturely, but the community services are not perfect, and the management is relatively mechanized. Therefore, improving work efficiency and enriching and perfecting community service is an increasingly important issue. Data mining is to extract valuable information by analyzing the internal connections, rules, and patterns of these data so as to provide more favorable decision support for community managers and provide users with more humane and modern community intelligence services. This research focuses on the implementation of community data processing systems based on data mining. Firstly, data preprocessing analysis is carried out, the realization of data storage and cache is studied, the process and characteristics of cluster analysis are studied in detail, and the simulation results of a community data processing system based on data mining are summarized. This study uses data mining technology to dig out the daily consumption data of users in the community mall, cluster the data, and analyze the consumption situation and consumption types of different types of users. In this study, data mining technology is used to mine the fault repair data of the community, and classification prediction technology is used to classify and predict different types of faults so as to help managers troubleshoot problems existing in the community environment.

1. Introduction

Intelligent community, in theory, is the intelligent control of community data, the combination of hardware equipment and network information technology to connect the whole community network, and intelligent control through mobile terminal software [1, 2]. With the rapid development of information technology and the rapid development of intelligent devices and information devices, the intelligent community is no longer an empty talk. People's pursuit of intelligent and humanized life in the new society has also become the driving force for the market expansion and development of intelligent communities [3]. Data mining, as its name implies, is actually an analysis process of mining or extracting valuable data or knowledge with unknown laws from intricate and irregular data [4]. With the rapid growth of community databases, a large amount of potentially valuable data is scattered in various database tables. If data mining cannot be used to efficiently integrate and analyze these data, it will not only waste data information but also cannot help community managers make important decisions [5, 6]. With data mining as an effective means, smart home systems with push function as the way, and with the purpose of providing high-quality and humanized community services, to improve the lifestyle of community residents to the greatest extent, provide decision-making assistance for community managers, provide users with a comfortable, convenient, intelligent, and colorful life, and embrace the arrival of the information age. Based on the existing intelligent community system, this paper makes an in-depth analysis of data mining architecture and constructs a data mining architecture suitable for the community. The clustering analysis method and classification prediction method of data mining technology are applied to the consumption data and historical repair data of community users. According to the data integration and analysis results,

the decision-making basis for community service managers is provided. In addition, in the process of promoting the modernization of social governance, there are information barriers between government functional departments and resources that cannot be reasonably integrated, which leads to many community management services not being effectively and quickly supplied. The construction of a community management service platform can change this situation to some extent and promote the modernization of social governance. Through the constructed community management service platform, to obtain a large number of community and community residents data through data mining, to help the government make decisions on this basis and improve the intelligent level and professional level of social governance.

2. Data Preprocessing

For numerical data, data mining analysis methods have inherent advantages, and they have high computational efficiency and fast execution speed. Therefore, this study tries its best to convert data into numerical data for operations, which are mainly completed in data preprocessing [7]. Data preprocessing can be divided into two categories: the first is to shield and omit some irrelevant data in the clustering algorithm; the second is to process some nonnumerical data and convert them into numerical data.

In the first type of data preprocessing, since the basic information used to identify users, such as user number and name, is meaningless to analyze the identification information of such character types in the clustering algorithm, the records are identified only when they are stored and screened during algorithm analysis. For example, the data selected by a certain cluster include user number, age, communication consumption, entertainment consumption, and domestic consumption, which is a group of five dimensions of data space. After the omission, the data set becomes the four-dimensional data of age, communication consumption, entertainment consumption, and domestic consumption, and the data of all dimensions are numerical, which meets the requirements of cluster analysis.

In the second type of data preprocessing, it is mainly aimed at the digital standardization of nonnumerical characteristic data, such as residence location, gender, and so on, and converts them into feature vectors, which are used to identify the corresponding points of the user's feature space, so as to realize the clustering algorithm analysis of the feature space. For example, 0 indicates female and 1 indicates male. Different positive integers represent different communities. The address number is composed of the building number, unit number, and room number. The address number is the floating point number added on the basis of the corresponding integer of the cell, which can not only ensure the numerical value of the data but also carries out correlation analysis for the user's living location.

3. Data Storage and Caching

Considering that the data can be reused and to ensure the consistency of various data, ARFF format files are used to cache the collected data sets to improve system performance [8]. Data set features are mainly divided into the following two types:

(1) Static characteristic data: such as name, gender, age, residential address, and other attributes that remain unchanged for a long time

(2) Real-time statistics: such as user consumption data, reports for repair, water and electricity consumption, and so on, need to query the database in real time to obtain statistical data

Data storage solution is shown in Figure 1:

According to the characteristics of these two kinds of data, this paper adopts the storage and query scheme, as shown in Figure 1, to optimize the data. Since static feature data have not changed for a long time, statistics can be queried and updated from the community management system database at intervals (monthly or even quarterly). While real-time data can be optimized based on cache file modification time, for example, the cache file stores all the consumption data of a user before December 31, 2019. For the subsequent data mining, it only needs to query the user consumption data from January 1, 2020 to the query time in the community management system database, and then add them to the cache data.

4. Clustering Analysis

In the front page, users input the initial K value of cluster analysis, the start and end time of cluster data, and select the data source to be clustered. The data of community mall can be carried out separately, or the community mall can be combined with the basic information of community users for cluster analysis. The value of k is verified in Java Script to determine whether it is empty, whether it is a numeric parameter, whether it is greater than 0, and so on. After the verification is successful, the data are passed to the action for processing [9]. If there is no corresponding cached data, the corresponding data in the data source will be queried according to the time, and the data will be input into the clustering algorithm for analysis after data pretreatment and caching. The analysis results of the algorithm should also be processed and evaluated to obtain the final results of the cluster analysis.

4.1. Analysis of Clustering Algorithm. The principle of the clustering algorithm is as follows: firstly, scan data one by one, and each data feature is classified into the same class or generated into a new class according to the distance from the scanned data; then the distance between the various basis classes is combined until a certain requirement is reached and stopped. In this paper, the classical K-means algorithm is adopted for clustering, and the cluster analysis process is shown in Figure 2:

(i) During initialization, n total data objects and K objects are given as the initial nodes of clustering.

(ii) Scan other node objects in turn, and calculate the similarity distance between these objects and the initial central node. The calculation formula uses the

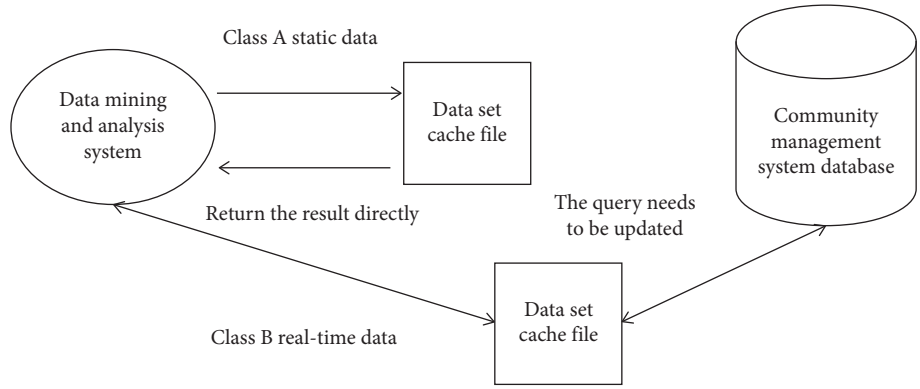

Figure 1: Data storage solution.

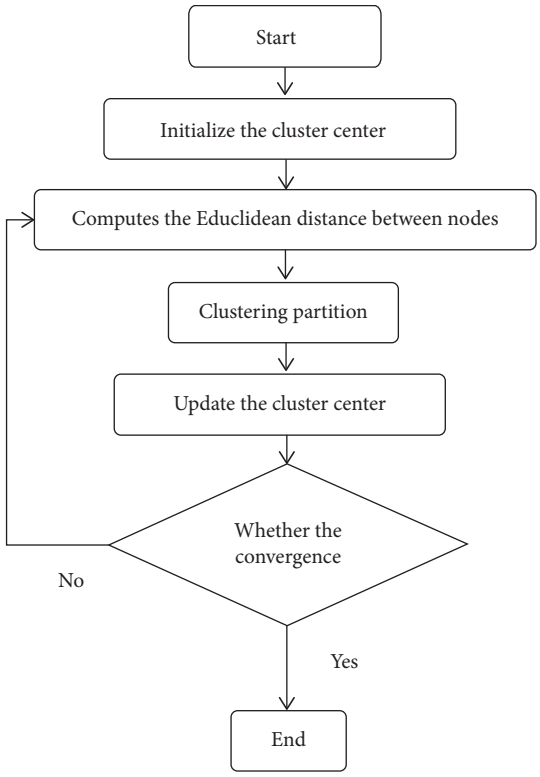

Figure 2: Cluster analysis process.

classical Euclidean distance formula to measure the index of similarity as follows:

$$d(x, y)\sqrt{\sum_{i=1}^{P}\left|x_i - y_i\right|^2}. \quad (1)$$

Among them, the similarity index meets the following mathematical function requirements: (a) $d(x, y)$ represents the distance between two objects and is a nonnegative value; (b) $d(x, x) = 0$, the distance between an object and itself is always equal to 0; (c) $d(x, y) = d(y, x)$, the distance function has symmetry; and (d) $d(x, y)$ cannot be greater than the sum of $d(x, z)$ and $d(z, y)$, because the direct distance from object x to object y cannot be greater than the distance from any third object z in the path (triangle inequality)

(iii) These objects are respectively assigned to the most similar (shortest Euclidean distance) cluster

(iv) Recalculate and update the cluster central node

(v) This process is repeated until the standard measure function (mean square deviation) converges

Consumption table of cluster analysis is shown in Table 1.

The consumption data after preprocessing and storage (as shown in Table 1) can be analyzed according to the abovementioned clustering algorithm. All the user's feature attributes constitute each corresponding point of the user's feature space, and some methods can be adopted to reduce the accuracy and improve the clustering efficiency through distance division of the feature space. Finally, find the feature space of similar user groups and form a list of users with similar features and user group classification. The closer the point is to the user, the higher the similarity to the user. In the face of the huge amount of user data in the community, especially when there are many user feature dimensions, 10,000 users will generate 100 million similarity distances, and the calculations are extremely complicated. At the same time, considering the high latitude data vector, the k-means algorithm is not highly efficient. Therefore, the system will reduce the accuracy of some very similar user attributes in the process of clustering so as to reduce the number of clustering indexes and improve efficiency. For example, users whose age ranges from 24 to 26 years old are clustered in accordance with the age of 25. In addition, users living on the same floor are classified as one, that is, the last two room numbers in the address number are omitted. For example, the user with address number 1.0231101 and the user with address number 1.0231102 are identified as 1.02311.

The k-means algorithm requires the initial value of the number of clusters to be set in advance. The more clusters there are, the more classification results there will be and the more detailed classification conditions there will be. However, the more, the better. In this system, community managers can input the number of clusters in advance on the page, and the background server will use the obtained

Table 1: Consumption table of cluster analysis.

Age	Shopping consumption	Entertainment consumption	Tourism consumption	Household consumption	Education consumption
25	800	1800	1200	400	1000
35	1000	3100	2400	800	200
40	1000	1600	2000	1000	3000
60	800	100	600	2300	0

number of clusters for clustering analysis by the *K*-means algorithm. By default, the number of clusters is 3.

4.2. The Key Code for Calculating the Euclidean Distance

```
Private float distance (float [] element, float [] center){
Float distance = 0.0f;
Float x = element [0] – center [0];
Float y = element [1] – center [1];
Float z = x*x + y*y;
Distance = (float) Math.sqrt (z);
Return distance;
}
```

The key code for calculating the error sum of squares criterion function:

```
Private void count Rule (){
Float jcf = 0;
For (int i = 0; i < cluster.size (); i++){
For (int j = 0; j < cluster.get (i).size (); j++){
jcf+ = error Square (cluster.get (i).get (j), center.get (i));
}}
jc.add (jcf);}
```

Update the key code of the central node of the clustering algorithm:

```
Private void set New Center (){
For (int i = 0; i < k; i++){
Int n = cluster.get (i).size ();
If (n ! = 0){
Float [] new Center = {0, 0};
For (int j = 0; j < n; j++){
New Center [0]+ = cluster.get (i).get (j) [0];
New Center [1]+ = cluster.get (i).get (j) [1];
}
New Center [0] = new Center [0]/n;
New Center [1] = new Center [1]/n;
Center.set (i, new Center);}}}
```

4.3. Cluster Analysis. The system uses multiple methods to display clustering results at multiple levels and strives to provide community managers with clear and concise data results and convenient operation interfaces. The system sets up three forms to display the cluster analysis results: a pie chart, a cluster center distribution table, and a user feature list.

(i) Pie chart: intuitive, can clearly and concisely show the proportion of each category

(ii) Cluster center distribution table: it enables community managers to understand the distribution of each cluster center and understand the characteristics of cluster user groups

(iii) User feature table: it can fully display the user group corresponding to each clustering result and its detailed information

In addition, each category should have a push service based on the information push system. The purpose of cluster analysis in this study is to cluster irregular user consumption data and user basic information data. And different types of information are pushed to different consumer groups to achieve humanization and intelligent management of the community.

Example analysis of clustering results is shown in Table 2.

According to Table 2, for young users with relatively high entertainment and shopping consumption, the community manager can push promotional activities such as mall discounts for these users based on the actual situation of surrounding commerce, advertising, and investment promotion in the community. For middle-aged and elderly users who consume less entertainment and more household services, community managers can push advertising information such as medical treatment, social security, and home promotion to them.

5. Analog Implementation

In this test, the consumption data of community users in a quarter from October to December 2019 provided by a company were used for testing. The test results are shown in Table 3.

According to Table 3, the test results show that among them, the share of users in the first category is up to 45%. This category of users has the highest entertainment consumption and belongs to the entertainment consumption type users, whose average entertainment consumption is about 2500 yuan. Therefore, for these kinds of users, they can push peripheral business, advertising, and entertainment information. The second category accounts for 23% of users, who spend the most on study and tourism, with an average of 1700 yuan on education and 1300 yuan on tourism. These users belong to educational tourism users, so they can push educational information around the community, such as English learning, primary and secondary school training, music, dance, and other educational information, as well as surrounding tourism

Table 2: Example analysis of clustering results.

Cluster type	Generation	Consumption type	Push ads or information
First	Youth	High entertainment consumption, low household consumption	Entertainment information
Second	Middle aged	Entertainment consumption medium, household consumption medium	Entertainment and household information
Third	Elderly	Low entertainment consumption, high household consumption	Household information

Table 3: The test results.

Clustering categories	Proportion (%)	Age	Entertainment	Tourism	Shopping	Education	Living	Propensity to consume
First	45		2500	1100	200	500	400	Entertainment consumption type
Second	23		400	1300	600	1700	800	Educational tourism type
Third	32		200	900	400	200	2100	Home saving

information for them. The third category accounts for 32% of the total number of users. This category of users has a high level of home consumption and belongs to the frugal type of home users, with an average home consumption of about 2100 yuan. Therefore, they can push medical, social security, home, and other service information around the community to them.

For the classification test of obstacle report data, in this test, the obstacle repair data of a community in 2020 are used as the classification training data. Use the obstacle repair data of a community in 2020 as the real test data, and use the decision tree obtained from the training data to predict and analyze the real data. The test results show that elevator failure, line failure, and sewer failure are key problems. Community managers need to focus on maintenance and maintenance to avoid failure. Lighting faults, access control faults, monitoring faults, and other faults are not critical problems. The system makes a return visit decision and judgment for specific obstacles in the community. The classification decision judgment results of common disorders are shown in Table 4.

In the data reported for repair, only 1.51% of those reported for repair need to be paid attention to and visited again. Obviously, only a small part of the repair records have the value of a return visit. The list contains all the records that need to be returned. The community administrator can view the detailed repair reports and directly push messages to the user.

According to the test results of cluster analysis and classification prediction analysis of mining and analysis systems, the application of big data analysis technology is conducive to aiming at these characteristics of user consumption data. After analysis, it was decided to use the clustering analysis method in data mining to analyze user consumption data. Big data analysis technology is used to classify the data of obstacle repair reports according to certain rules and find out their classification rules, so as to predict the distribution of different types of obstacles in the community and the proportion of obstacle repair reports according to their classification rules. In this way, community managers can not only actively eliminate obstacles but also master the distribution and regional coverage of obstacles in the whole community.

Table 4: The classification decision judgment results of common disorders.

Fault category	Times	Fraction of coverage (%)	Focus or not
Sewer faults	15	7.18	Yes
Line faults	23	2.36	Yes
Elevator faults	34	5.61	Yes
Monitoring faults	18	2.33	No
Access control faults	17	2.48	No
Lighting faults	21	3.19	No
Other faults	9	1.41	No

6. Conclusion

The community data processing system based on data mining includes three major systems: community management, data mining and analysis, and information push. Among them, community management helps community managers to complete daily management works, such as user data maintenance, fault reporting, and water and electricity payment. At the same time, it is also the main source of community user data and provides platform support for data mining and analysis systems and information push systems. Data mining and analysis, on the one hand, it mines the daily consumption data of users in the community mall. By clustering the data, it can analyze the consumption situations and consumption types of different types of users and provide support for the strategic decision-making of the community by combining the basic information of users grasped by the community, so as to provide more targeted and humanized services for users. On the other hand, it mines the fault repair data of the community and classifies and predicts different types of faults by using the classification and prediction technology, so as to help managers troubleshoot problems existing in the community environment and reduce the incidence of faults. Information push service is an important link to combine community service with a smart home system. It uses a XMPP protocol to push community information to users' mobile phone terminals by using a smart home system, so that users can get all kinds of community service information in real time, narrowing the distance between community managers and

users. Therefore, on the basis of data mining technology, we can make full use of the innate advantages of the community to obtain user information, analyze user data, and use the analysis results to provide help for the daily management and strategic decision-making of the community, so as to improve the quality of service and the sense of belonging of the community.

At the same time, due to the lack of personal research ability and the limitation of data collection, there are still many deficiencies in this study, which can be further improved in future research. As for data mining technology, this paper has not studied deeply enough. More effective algorithms can be used to analyze community data. In the management data and user consumption data of the community, there are still a lot of wasted useful information for the system to mine and analyze, such as the analysis of the user's payment data and the investigation of some abnormal users. There is still room for improvement and development. In the process of follow-up work and study, we will continue to study the technology involved in this topic and constantly improve and enrich the research results of this topic.

Data Availability

The data used to support the findings of this study are included within the article.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] K. He and Z. Yang, "Smart community management system construction," *Bulletin of Surveying and Mapping*, vol. S2, pp. 248–250, 2016.

[2] N. Singh, V. Singh, and T. E. Carlson, "Pim-GraphSCC: PIM-based graph processing using graph's community structures," *IEEE Computer Architecture Letters*, vol. 19, no. 2, pp. 1-2, 2020.

[3] T. M. Roddenberry, M. T. Schaub, H. T. Wai, and S. Segarra, "Exact blind community detection from signals on multiple graphs," *IEEE Transactions on Signal Processing*, vol. 68, no. 99, pp. 1-2, 2020.

[4] A. Mahmood, K. Shi, S. Khatoon, and M. Xiao, "Data mining techniques for wireless sensor networks: a survey," *International Journal of Distributed Sensor Networks*, vol. 2013, pp. 185–193, 2013.

[5] H. Li, Y. Liu, R. Zhang, and G. Liu, "Design of community ECG monitoring system based on Wireless sensor Network," *Journal of Tianjin Polytechnic University*, vol. 34, no. 01, pp. 64–67, 2015.

[6] S. F. Tonellato, "Bayesian nonparametric clustering as a community detection problem," *Computational Statistics & Data Analysis*, vol. 152, 2020.

[7] J. Wang, X. Wang, and C. Zheng, "Design and implementation of intelligent community 3D display system," *Computer Technology and Development*, vol. 028, no. 009, pp. 156–161, 2018.

[8] X. Ma, J. Geng, and C. Fan, "A network community detection algorithm based on multi-view data fusion," *Computer Applications and Software*, vol. 35, no. 07, pp. 310–314, 2018.

[9] E. Luo, G. Wang, and C. Li, "Research on clustering algorithm for multi-dimensional data deduplication in big data environment," *Mini-Micro Systems*, vol. 37, no. 03, pp. 40–44, 2016.