Design and Implement of Customer Communication Behavior Analysis System

Qingzhang CHEN,
College of Computer, Zhejiang University of Technology, Hangzhou, China
qzchen@zjut.edu.cn

Yanqing OU, Hang SUN
College of Computer, Zhejiang University of Technology, Hangzhou, China
yqou1986@126.com sunh@zjut.edu.cn

***Abstract*—In this paper, design and implementation of a major telecommunications service providers for user behavior analysis of communication systems. Thesis the use of data mining of fuzzy cluster analysis and methods of thought, from the traffic analysis, custom classification, by the group, then, loaded with the extent of changes in trends in the call on the custom's phone records for analysis. Pass the test showed that the target customs of the system breakdown, calls the trend analysis results, enterprises can support the establishment of a new telecom operators and value-added services can be used to provide personalized services to users based on, as well as marketing planning and provide more scientific Decision Support.**

***Key words*—data mining; fuzzy clustering analysis; customer communication behavior analysis; customer clustering.**

I. INTRODUCTION

In recent years, global mobile communications have developed rapidly, with the progress of information technology and telecommunications industry, the monopoly is broken and an increasingly competitive telecommunications market, the development of the industry is facing new opportunities and challenges. May 24, 2008, the State of the Ministry of Industry and Information, National Development and Reform Commission, the Ministry of Finance jointly announced a three-sector "on the deepening of telecom reform Notices", which really started the restructuring of China's telecom industry off. In the "notice" that the development of third generation mobile communication (hereinafter referred to as 3G) as an opportunity to support the formation of the three has a national network of resources, relatively close to the strength and scale, business-wide capability and strong market competitiveness of the main[1] . The user is a corporate foundation for the survival and development, and maintain fully exploited to attract users and potential users of telecom enterprises to improve their core competitiveness.

In this paper, the objective is to design and realization of a behavior analysis of communication system users, which is based on large enterprise users of telecommunications data, through the call records of telecommunications users excavation analysis, to identify a user calls the law of conduct, the establishment of the user level and predict the future users the trend of call, allowing carriers a better understanding of enterprise user preferences, to provide targeted services. The effective functioning of this system can not only improve the level of service to users, to meet user needs, but more importantly is, you can find those companies with high profit contribution of business users, and mining laws in order to increase its volume of business in this category, management through a series of marketing strategies, enhance enterprise competitiveness, the consolidation of customer loyalty and avoid the loss of users and the development of the user's consumption potential, so as to enhance the business profitability.

II. RELATED TECHNOLOGY

A. Data Mining

1) The Definition of Data Mining

Data mining from a large number of incomplete, noisy and ambiguous, the practical application of random data, extraction of implicit, in which people do not know in advance, but is potentially useful information and knowledge of the process. And data mining is similar to a synonym for data integration, data analysis and decision support, etc[2].

2) Fuzzy Cluster Analysis

The concept of cluster analysis is from the multivariate statistical analysis, it is a mathematical method. In people's daily life, often to everything around him or a collection of values in accordance with their nature, purpose, divided into a number of attributes such as range or more groups, so that the classification process is called cluster analysis. And clustering methods are divided into the classical method of cluster analysis divided into hard and fuzzy cluster analysis classification of soft breakdown. In recent years, cluster analysis has been widely used in weather forecasting, geological exploration, pattern recognition and other fields[3].

B. Data Transformation Services(DTS)

In order to support decision-making, many organizations need to pool data for analysis. But usually the data in different formats are stored in different places.

doi:10.4304/jsw.6.8.1484-1491

Some may be a text file, although some have a table structure but does not belong to the same data source, which greatly hampered the focus on data. SQL Server provides a Data Transformation Services components, that is Data Transformation Services (DTS). DTS itself contains a number of tools and interfaces to provide all the support it currently OLE DB data sources between the import, export or transfer data, and this work has become simple and efficient. DTS can use all the OLE DB, ODBC data source driver or the text of this article, and between the SQL Server Import, export or transfer data[4].

C. ADO Data Access Technologies

ADO is a set of optimized exclusive access to the database object set is the application-level programming interface, which is based on OLE DB, OLE DB for a package. ADO provides the following ways: (1) to connect to data sources; (2) specify the command to access the data source can also be optimized with variable parameters or implementation; (3) the implementation of the command, for example, a script SELECT; (4) If This command data in table form of the return of Bank of China, these firms will be stored in a readily accessible for inspection, operation or modification of the cache; (5) the appropriate circumstances, can change the contents of the cache line write back to the database, update data sources; (6) The conventional method to provide error[5].

III. COMMUNICATION BEHAVIOR

A. Communications Data Analysis

According to preliminary research and analysis, traffic analysis, this paper will, by then the group, full of changes and trends in four areas calls for an analysis of user behavior to call, in addition to the need for user classification, each functional category can be sub - a sub-function.

User data is divided into static data and dynamic behavior of the data, the user's static data include: the user code, the user telephone number, user type, number and pay their way; the dynamic behavior of the user data is the user's phone records. Static data with the relative stability, and dynamic data can be continuously updated, which will in the database design requirements and the actual information the user's current line, analyzing the latest data reflect the behavior of users. Communication data sample is structured as follows and table Ⅰ:

TABLE Ⅰ Part of a User's Actual Communication Data

Field Name	Description
user code	1
user number	13588054757
numbers belong to	Hangzhou City
the call type	Caller
call date	2009-01-01
start time	16:24:49
calls	18s
weeks	6
the date type	Holidays

1) Static Data:

a) user code: identification of a user, a user may have several phone numbers

b) user number: user telephone number only

c) the user types: the definition of the user telephone number of the type of

d) numbers belong to: the definition of user phone numbers in their respective districts

e) Payment Method: the definition of the user a way to pay telephone charges

2) Dynamic Data:

a) caller id: the definition of a party to initiate the call telephone number, and called number, call date and starting time of a unique identifier with the call records

b) called party number: the definition of the called party telephone number

c) the call type: the definition of the type of call-owned

d) call date: the date of the definition of call

e) start time: the definition of a call the moment of its occurrence

f) calls: the definition of call duration

g) date: definition format yyyy-mm-dd

h) weeks: a few weeks, with 1 ~ 7, respectively, said Monday to Sunday

i) the date type: date of their types, and holiday work is divided into two types

B. Communications Parameters of the Definition of User Behavior

1) User Category

According to the pyramid theory, law (80% of the profit from 20% of users), classification of users, in line with the actual functioning of customer relationship management system, the first step.

If the company has to care about each of its users, it will not cost-effective, so in principle, will be based on user-to-business contribution to the separation of the user level, and then the same levels of users, users with basic information, and further fine - points with similar attributes to the user classification together, although not necessarily in the same class must have the same users, but must have some similar relationship, so that users can significantly reduce the cost of relationship management, marketing, if all users N, and 20% will be divided into M-ping categories, one of the categories, there are C% of people with a particular call, if that calls for acts of marketing, takes each element of cost D, the total marketing cost:

$$N \times 20\% \times \frac{1}{M} \times C\% \times D = \frac{N \cdot C \cdot D}{500M} \tag{1}$$

a) aimlessly original cost of marketing to N * D, marketing costs for the $\frac{500M}{C}$ fold gain ratio;

b) compared with the pyramid law to spend 20% * N * D, marketing costs for the $\frac{100M}{C}$ fold gain ratio;

Separate categories of users, you can of a particular user, at a given time, to make specific marketing, to achieve cost savings in marketing.

2) Traffic Analysis

Analysis of the main flow chart shows the use of user calls, the main purpose is to understand the user calling behavior, which can occur through the analysis of the causes of behavior, to understand trends in user behavior to predict future behavior. User behavior accurately spaced from the date, date type, time, call type, in terms of different users to conduct flow analysis.

3) Group Call

To identify specific users, information on their call details, call volume statistics, call the object's position and so on, to analyze the user's calling behavior. Calls received from the user group behavior: a certain number of seconds the call date, the date a certain number of calls to a particular date within the object of a call such as call position.

4) Full Extent

Full extent of: record call volume ratio loaded. If hours 60 minutes, if a user calls in this hour 30 minutes, when the full rate of above 50%. To a week for all full-time rate calculated. Loaded with a single user by observing the situation, the provision of services for users.

5) Call Frends

Comparison of several months of on-call and call the number of seconds, from the trend can clearly see the volume of calls there has been a downward trend, the more obvious decline in the extent of its loss of the greater probability, it is estimated that next month's call volume increases and decreases timely to make marketing and customer service strategy, which will retain the loss of business.

C. Algorithm Design

1) User level clustering algorithm

a) According to the classification attributes the users choose, extract the level user's original data from the database.

b) Membership functions initialize data. Fuzzy clustering algorithm requests the initial data managed must be numerical data, unified less than or equal to 1. Only after the completion of the numerical treatment of various types of raw data, the next step of fuzzy clustering can be carried on. Membership function initialize the data is aimed at translating all the different types of attribute data into numerical data. User data will be divided into two types by the system, namely, numerical and generic type.

Numerical attribute values include the number of seconds and call the number to call under a variety of situations. If every value is considered as a separate entity, may result in over-classification refinement and complication of fuzzy clustering algorithm, so interval division of numerical attributes is needed. In the division of numerical attribute values, the same attribute values are in the same interval, the same interval of the attribute value has the same membership function, the others should be processed based on the distribution of distance, less than the value range is divided into a range. l is the the number of attribute-based classification , C_l is the l interval, $N(C_l)$ is the type of C_l contains the number of attribute values, $C_l^{(i)}$ is the first class section l of attribute value i , Attribute value is the membership function for :

$$\mu_R(C_l^{(i)}) = N(C_l)/n, l = 1,2,3...; i = 1,2,3 \text{ , } n \text{ is the number of data elements} \quad (2)$$

Attribute value is a type of attribute values of token classification, such as the payment method, the number belongs to, these attributes are get a certain value from a limited range. The same attribute values will be classified as a class, and its membership function value mainly considers the proportion of all types of property value in the total number of classification, the membership function as follows:

$$\mu_R(C_l^{(i)}) = N(C_l)/n, l = 1,2,3...; i = 1,2,3 \text{ , } n \text{ is the number of data elements} \quad (3)$$

Initialization is complete, integrated classification of a property value of all less than or equal to 1 to initialize the data sheets.

c) Using fuzzy matrix clustering method, suppose the domain is U, in this system is initialized data tables, the number of elements is $|U|$, that is, the number of users participate in fuzzy clustering., and then establish the fuzzy similarity relation R among the initialized data element U. In the use of the absolute value decrease method, we can calculate the R matrix elements:

$$r_{ij} = 1 - \sum_{k=1}^{m} |x_{ik} - x_{jk}| \quad (4)$$

d) The data of R matrix record the degree of similarity among users, regard the R matrix as the adjacency matrix of weighted graph G = (V, E), using Prim algorithm for the Maximum Weight Spanning Tree. Prim algorithm: first step, the initialization: U={ u_0 }, TE={} ; second step, seek the maximum side (u_0, v_0) , $TE + \{(u_0, v_0)\} \Rightarrow TE, \{v_0\} + U \Rightarrow U$ in all sides $u \in U, v \in V - U$; third step: if U=V, the algorithm finished, otherwise repeat the second step.

e) Carried out on the maximum spanning tree pruning classification, according to the needs of practical problems, set an appropriate $\lambda \in [0,1]$, T (e) is the weight of the edge e, if $T(e) < \lambda$, edge e will be deleted, the connectivity branch we get is the classification based on λ .

2) Algorithm of Determining The Call Trends

Part of trend analysis calculates the probability of increase or decrease of the call number in next a month based on the increase or decrease in the number of call and the number of seconds in the past months. Here consider the number of seconds and the call number as

two main factors influence the call flow next month, and thus compose the domain: $U = \{CommunicationSeconds(u_1), CommunicationTimes(u_2)\}$, On a single factor, the domain is: $V = \{increase(v_1), in\,var\,iable(v_2), decrease(v_3)\}$ .If the increase probability of the number of seconds achieve 30%, changeless probability achieve 40%, decrease probability is 30%, and the evaluation of the number of seconds is (0.3,0.4,0.3). The number of seconds and the evaluation of call number compose the matrix R, and then give the two factors the weights a_1 and a_2 , require weights meet the normalized requirement: $a_1 + a_2 = 1$ .These two weights compose a vector A=(a_1 , a_2) on the domain U. Thus, we can get the assessment of all months:

$$B = A \circ R \tag{5}$$

First of all, achieve the number of seconds and call number of all months.Then, calculate the single factor evaluation. Generally time interval is smaller, call trend will be more similar, the influence of the cases of the change in different months is different for the total call trend. Beside the last comparison, the other various processing base on the results of the comparison. If the calls increase, $v_1 = 0.5^n + v_1$; calls changeless, $v_2 = 0.5^n + v_2$; calls decrease, $v_3 = 0.5^n + v_3$. n is the compare order, the initial value of v_1 , v_2 and v_3 is 0. at last, according to the comparison result, if increase $v_1 = 0.5^{n-1} + v_1$,changeless $v_2 = 0.5^{n-1} + v_2$, decrease $v_3 = 0.5^{n-1} + v_3$.

At last, according to $B = A \circ R$, we can get the assessment of all the month selected of increase or decrease in traffic situation. Here the influence weights that the two factors of the number of seconds and the call number on the total call flow is same, A= (0.5，0.5) . After calculated, $B = (b_1, b_2, b_3)$, b_1 is the increase probability of the call volume, b_2 is the changeless probability of the call volume, b_3 is the decrease probability of call volume.

3) The Algorithm of Full Loaded Degree Analysis

Full load degree is the ratio of record talking. In this paper, statistical analysis methods and ideas are used to analyses users full load degree, there are two main steps:

a) For all users full load degree of analysis. Firstly, Statistics talking seconds between each user's choice of beginning and ending time according to the selection conditions. Secondly, each user's talking seconds divided by total talking seconds to get user's full load degree. Finally, rank the user's fill load degree according from large to small.

b) Select top users in the rank to analysis the full loaded degree of single user. Regard one week as a unit, stat the user's talking number of seconds in seven days a week. Use the talking seconds of different time obtained in the statistics to divide by the total number of seconds of the corresponding dates of the time, we can get the full rate of the time in seven days a week, and then indicate the user's full case based on the rate we chose

IV. System Design

A. System Structure

Completion of system calls based on user-recorded data collection analysis, in order to judge the acts of users using the telephone, can generally be divided into two functional modules: the first functional module is a database management module, the second is a call recording function analysis module. System framework as shown in Fig.1.

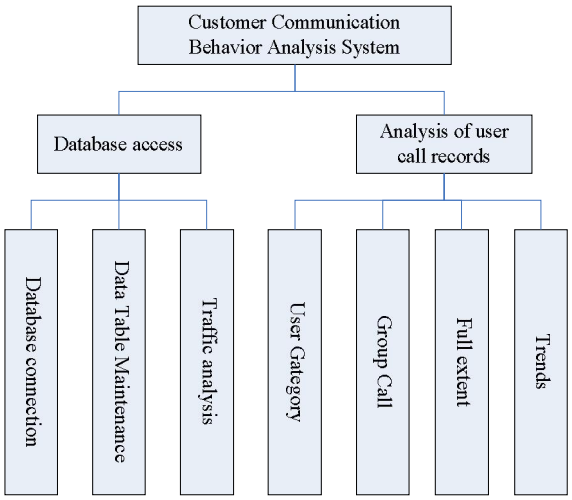

Figure 1. System framework map

B. System Analysis Module

1) Database Management Module

The module is the main background to complete the data in database management functions, including database connectivity and data tables to maintain two parts. Management functions in the database for static data and dynamic data management is very important. The management of static data that users enter basic information, and delete data, the date of the type of changes. Dynamic data management in connection with the call records to the user input and delete function. As the database system to set up the referential integrity constraints, data input, modification, relevance has been made to delete certain degree of protection.

2) Call Record Analysis Module

The module is the core of the whole system, the background database and database management module for its services are. On demand system in five aspects of the user's phone records for analysis:

a) Traffic analysis: traffic analysis is divided into time slots, while the flow between the user segment, compared with the user calls the type of flow compared in three parts.

b) User category: user needs to meet the requirements in the subdivision, can be divided into two parts: first, the

user classification; Second, fuzzy clustering level users. First, the pyramid law classification of users, and then the users on the same level of further sub-clustering.

c) Call group: to meet the system requirements call for a specific user details, traffic statistics and analysis of call requests, such as object position.

d) Full extent of: all the users into a single user full analysis and full analysis of two periods, respectively, to meet all the fully loaded rate of analysis and individual analysis of the user's full rate.

e) Call trends: the basic unit for months under the specific circumstances of the call option and the number of seconds of data statistics, and intuitive manner in order to show that at the same time next month estimated the changes in call volume. The system used to meet the needs of some of the requirements of the call trend.

C. Database Design

The system uses Microsoft SQL Server 2000 database as a solution, through the collection, collation of data, the establishment of the database. User data are static data and dynamic behavior of the data, static data are relatively stable, and dynamic data can be continuously updated, which requires a database and users to the actual information on the current line, analyze the latest data reflect the behavior of users. In addition, in order to date for different types of users to an analysis of phone records, an increase of the maintenance schedule to date information.

The overall database design to meet the pattern of relationships in accordance with the principles of 3NF, taking into analysis of call records can be of. Due to the large number of user data and rational design of the database can improve the efficiency of data analysis, but also ensures the correctness of data, effectiveness and compatibility. User static data and dynamic data were stored in a different table, set the corresponding primary key to ensure the integrity of entities. Main table and the external table can be the primary key and foreign keys to connect and form a reference to the relationship between the light and to reduce data redundancy, at the same time conducive to the consistency of data.

In meeting the requirements of 3NF on the basis of the database design includes three data tables, with the expansion of the function of the system, you can add more data sheets. At present the database to preserve the user's static data, dynamic call recording and the date of the type of information. According to the various functional modules of the system analysis, design the following data structures and data tables:

1) Basic information the user: the user code, user numbers, user types, their numbers and pay. As shown in table Ⅱ.

2) Recorded information call: Caller ID, called number, call type, call date, start time, call duration. As shown in table Ⅲ.

3) Date of information: the date, week, date, type. As shown in Table Ⅳ.

TABLE Ⅱ The User Basic Information Table

Field Name	Data Type	Description
User code	varchar(10)	Identification of a user, a user may have several phone numbers
User number	varchar(20)	The only phone number the user (primary key)
User Type	varchar(40)	Telephone number the user types
Number of local-owned	varchar(10)	User phone number in their respective districts
Payment Method	varchar(10)	Users a way to pay telephone charges

TABLEⅢ Call History Table

Field Name	Data Type	Description
Caller ID	varchar(20)	Call initiated by a party to the telephone number (primary key), and called number, call date and starting time of a unique identifier with the call records
Called number	varchar(20)	Called telephone number (primary key)
Call type	varchar(40)	the type of call
Call Date	varchar(10)	Call date (the primary key, foreign key)，form:yyyy-mm-dd
Start Time	char(8)	Call the moment of its occurrence (primary key)，form:hh:mm:ss
Call the length of time	int(4)	Call duration, unit for the second

TABLE Ⅳ Date Information Sheet

Field Name	Data Type	Description
Date	varchar(10)	Date (primary key)，form:yyyy-mm-dd
Weeks	char(1)	Day of the week, respectively, with 1 to 7, said Monday to Sunday
Date Type	char(6)	The date of their types, holiday and work is divided into two types

D. Packaging Design Category

To function as a function of various types of packaging to provide programming interface to enable the realization of a simple function. The main use of the system to achieve graphics drawing package type operation functions; CADOConn using ADO object class implements the basic database operations functions: to deal with such major database connection and recordset operation, Packaging Connection object and Recordset objects commonly used functions, such as open, close the connection, do not go to the record of the implementation of the SQL commands, access to records such as field values.

E. System Function Module Example

System used in Microsoft Windows XP operating system development platform environment, using VC + + 6.0 as a system development process, using Microsoft SQL Server 2000 as a database development tool.

1) Users sample classification module

The user interface level of classification, as shown in Fig.2.

Processes as shown in Fig.3. Which input data are: dates, date, type, user type, the way calls, call type and the user level. The output is: to meet the input conditions and the total number of user calls the number of seconds.

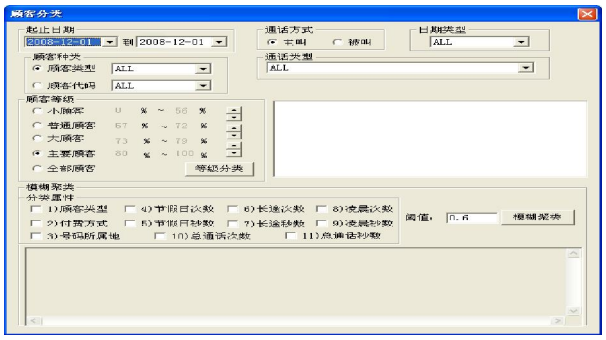

Figure 2. Features a user interface classification

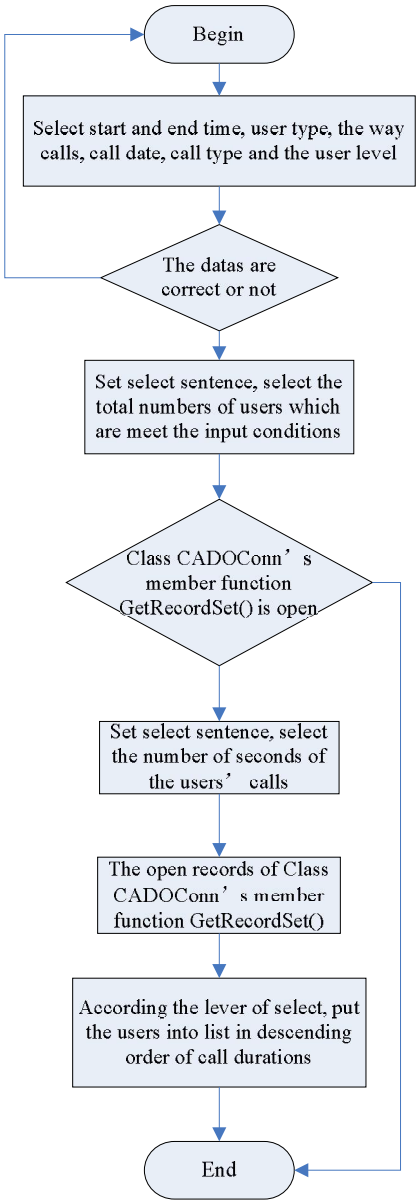

Figure 3. Flow chart of the user classification

2) System maintenance module

a) Database Connectivity

Database connection system and the background databases, including databases and disconnect the database connection in two parts.

(a) to connect the database to store a user call records and basic information on the database and client application to connect. Input: including the server name, user name, password and database name. Output: the success of connection, database connection errors. as shown in Fig.4, in Fig.5.

(b) disconnect the database functions: disconnect the current client application and database connectivity. Input: determine disconnect or cancel the operation. Output: the database has been disconnected tips.

b) Data Table Maintenance

Data sheets to achieve the maintenance of data input, edit, delete function, the specific call records for the user basic information and user input and deleted, the date of the type of changes.

(a) Call History / user data loading. Functions: to add to the database need to analyze the call records or information to add users. Input: data conversion and data packet path the path of the text. Output: data loading data loading success or error. as shown Fig.6 ,in Fig.7.

Figure 4. Connection database interface

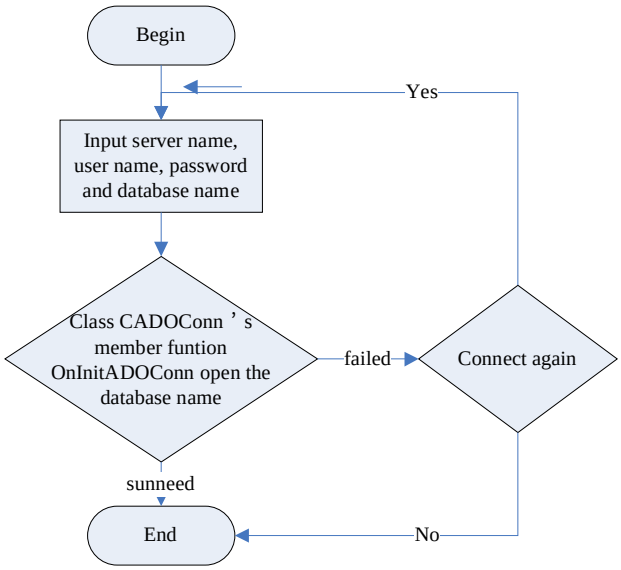

Figure 5. Connection database operations flow chart

Figure 6. Data loading interface

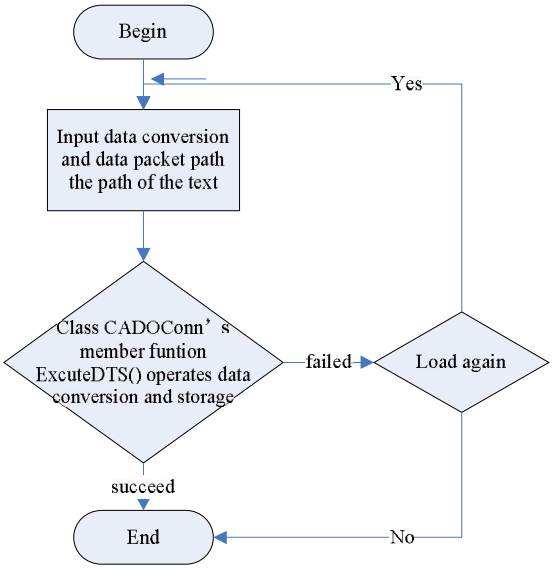

Figure 7. Flow chart of the data loading operation

(b) the user to delete call records. Function: To remove a user call records, to facilitate the updating of phone records that need to be analyzed. Input: Select the dates need to be deleted. Output: the deletion of the success or error.

(c) the basic information the user to delete. Functions: analysis of phone records do not need to delete users. Input: Select the user number, to identify the deletion. Output: success or failure of the deletion.

(d) the date of the type of modification. Function: modified type, the date can be divided into two types of holiday or work day, the date of type-based analysis of phone records to prepare. Input: Select the date and date types. Output: the date of the type of amendments to the success or error.

V. Test and Analysis of Data

A. System Test Ready

Experimental testing of the system, all data are taken from real mobile phone users of China Mobile's personal voice recording. The form of text information as the original record. Phone users in the act of analysis, the database must be connected, and call records and user information into the database of basic information. Fig.8 is a user calls the form of raw data records.

Figure 8. Records the user calls the form of raw data

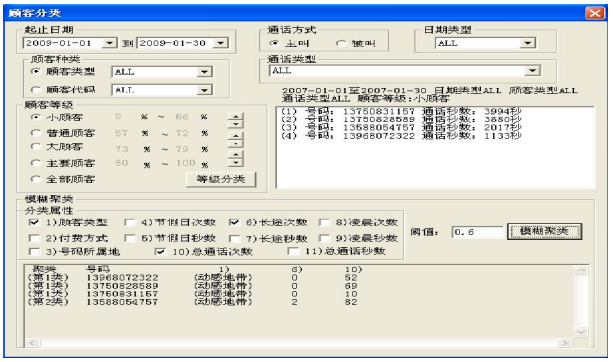

Figure 9. Classification results of user testing plans

B. Examples of System Function ModuleTest

Telecommunications customers in the user classification system as an example, functional testing, Fig.9 Classification results for the user test plans.

Know to pass the test, select the total number of seconds to pass the top 44% of users as a small user of the database in January 2009 in line with the conditions of 4 users classification, and then select "User Type", "long-distance number," "total calls "as a Category 4 property further fuzzy clustering of small users, the two prices at this time it will tie in with marketing programs, programs for those who will focus on classification, so that more targeted marketing, thereby reducing marketing costs.

VI. Conclusion

The telecommunications industry and economic development is a closely related industry, it is also an increasingly competitive industry. Telecommunications companies in the complex and ever-changing market environment for better survival and development, on the one hand, we must continue to improve technology, innovative products to enhance the strength of enterprise hardware, on the other hand, must also be customer-centric, understanding customer needs, enhance customer service quality, enhance the soft power[6]. In the course of this paper, the use of visual programming techniques and database management systems, as well as data mining and other related methods, to achieve the user call behavior analysis system, which can be carriers for the user class, business options, the development of personal operations in accordance with the decision analysis.

REFERENCES

[1] "Announcement on Deepening the Reform of Telecom",The Ministry of Industry and Information Technology,The National Development and Reform Commission, The Ministry of Finance,2008.5(Chinese)

[2] Kunjiang Lin,Minggao She,Xiufeng Jia,"Data mining technology and its application in customer behavior analysis of consumer",Fu Jian Computer,2007.2(Chinese)

[3] Xiaoyu Lei,Zhihua Qu,Yaming Zhang,Xiaoping Fan,Zhifang Liao,"APPLICATION OF THE ANALYTICAL METHOD OFFUZZY CLUSTER IN the Forecast and Analysis of IT Market",Computer System,2008.3(Chinese)

[4] "The basic concepts introduction of SQL Server Data Transformation Services",China Network Information Center. http://www.sudu.cn/info/html/edu/20070422/321611.html (Chinese)

[5] Yi Yuan,Visual C++ practice and improve - database development and application of engineering,Bei Jing:China Railway Publishing House,2005(Chinese)

[6] Tengjiao Wang,Ziyu Lin,"APPLICATION OF Data Mining in the field of telecommunications customer behavior analysis",Technology of Telecom,2008.1(Chinese)