Current Trends in Technology and Science
ISSN: 2279-0535. Volume: VI, Issue: II

User Behavior Pattern Mining Method Based On Adaptive Coding

Tong Zhang
Master degree candidate, Henan University of Science and Technology, China. zt524904041@126.com

Ruijuan Zhen
Doctor, Henan University of Science and Technology, China. 15036989996@163.com

Abstract — Mobile cloud computing is a new application model with the continuous development and integration of cloud computing and mobile internet. In the architecture of trusted mobile cloud services, the determination and identification of user's identity and behavior will be an important prerequisite for subsequent services. This project focuses on the user trust issues and behavior pattern mining. First, we give a formal description of the sequence of the user's temporal behavior, and the coding structure of the sequence of user behavior. Then, we model the user's normal behavior patterns based on context and timing functions. According to the characteristics of mining temporal behavior patterns, we draw on the theory of genetic algorithm, and propose an adaptive user behavior pattern coding method.

Keyword — Mobile cloud computing, user trust, temporal behavior, pattern mining.

1. INTRODUCTION

Mobile cloud computing is an inevitable product of the mobile Internet era under the background of big data [1]-[4]. In order to make the data available for the production industry services, computers must be able to efficiently dig out valuable information from massive data [5]. The purpose of data mining is to find hidden, unknown potential information from a large amount of data. Pattern mining is an important branch of data mining [6].

The behavior of mobile users using mobile cloud services is closely related to life [7], work, personal habits and interests. Mining the normal temporal behavior of mobile cloud users can provide rich and accurate reference samples for online behavior recognition and anomaly detection.

Although the user is a random use of cloud services, but in the long run, the behavior of a user to operate the cloud service is stable [8]. Due to the complexity of the environment factors of mobile cloud users, the traditional "solid state" pattern mining method can't improve the mining methods according to the environmental changes. This leads to a decrease in the accuracy of the mining results and a reduction in the mining efficiency. Therefore, the pattern mining algorithms in mobile cloud services also need to be able to maintain the high accuracy and robustness due to the continuous evolution of the environment.

The core of this paper is to explore the normal temporal behavior of the legitimate users. Firstly, we study the formal definition and coding method of the user's temporal operations, and use it as the basis for further analysis of the normal timing behavior. Then, based on the dependency and temporal characteristics of the nodes in the normal timing behavior, the user's normal behavior pattern mining method is studied, and the adaptive incremental mining of temporal behavior patterns is realized.

2. RELATED WORK

Data is stored in the storage media more and more frequently. All over the world brought together experts and scholars, on the basis of previous studies proposed a new branch of data mining. The subject draws on the ideas of statistics, artificial intelligence, signal processing and visualization and so on [9]. Data mining domain for other ideas as shown in figure 1.

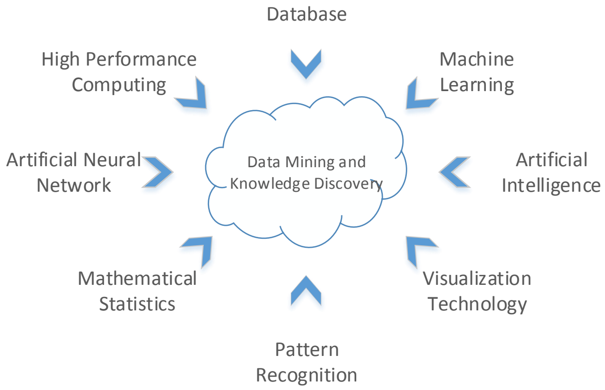

Fig. 1. Organization of mobile cloud computing content

Data mining is mainly used in the field of machine learning and artificial intelligence. The whole process of data mining also needs to cooperate with each other to get good results. The main research methods are as follows: neural network method, genetic algorithm, decision tree method, rough set method, positive and negative method, statistical analysis and fuzzy set method.

Which data mining method to use depends on the different objects. Due to the complexity of the mobile devices in mobile cloud services, the traditional pattern mining methods can’t be applied to mobile cloud services. In this paper, we propose an adaptive coding based user behavior pattern mining algorithm based on genetic algorithm. The algorithm can adjust the pattern mining strategy according to the change of the environment, so as to improve the accuracy and efficiency of the subsequent anomaly detection.

3. Algorithm Description

3.1. Genetic Algorithm

Genetic algorithm is a randomized simulation of biological evolution in nature search algorithm [10]. Genetic algorithm has been applied in many fields, such as optimization, machine learning and parallel processing [11]. Elements of basic genetic algorithm: chromosome coding method, individual fitness evaluation, genetic operator, and operating parameters.

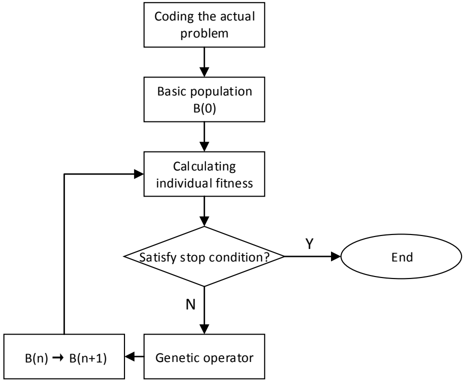

Fig. 2. Flow chart of genetic algorithm

As can be seen from Figure 2, the main genetic algorithm is as follows [12], [13]:

1) **Coding.** The genetic algorithm is used to search the solution space data before the search, and then the data of the solution space is represented as the genetic structure of the data.
2) **Initialization.** Set up the evolution algebra counter T; set the maximum evolution algebra T; randomly generate M individuals for the initial population B (0).
3) **Individual Evaluation.** The fitness of each individual in the population B (n) was calculated.
4) **Selection, Crossover, Mutation.** The selection operator, crossover operator and mutation operator are used to change the individuals in the population. The previous population P (T) after selection, crossover and mutation operation are obtained after the next generation P (t+1).
5) **Termination Conditions.** If t = T, t, t+1, then go to step three; if t>T, the evolutionary process obtains a maximum fitness of individuals as the optimal solution output, termination of operation.

3.2 Formal Description of User Timing Behavior

The formal definition of user timing behavior is the basis for the development of abnormal user behavior. In this study, a step formalization of user timing behavior is defined as the form shown in figure 3.

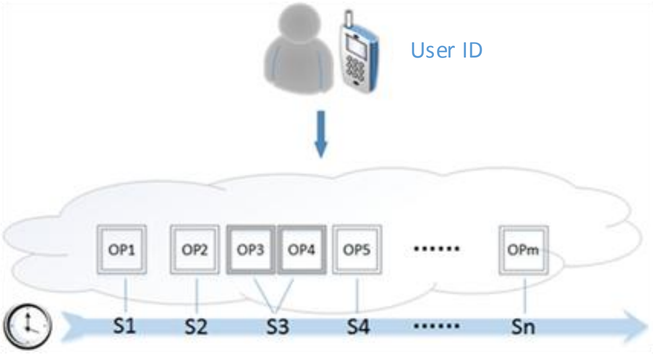

Fig. 3. Formalization of user's temporal behavior

As can be seen from the formal description of the user's temporal behavior step from the above, we can see that the A sequence of user actions in the user Ui can be expressed as a formula 1.

$$Sui ::=< UID \left| S_1 \ldots S_n \right| OP_1 \ldots OP_m > \qquad (1)$$

In general, the cloud server to provide services before the user's request for a cloud service contains lots of operations. These steps are stable due to the user's operating habits, and they also have a relatively fixed normal execution flow.

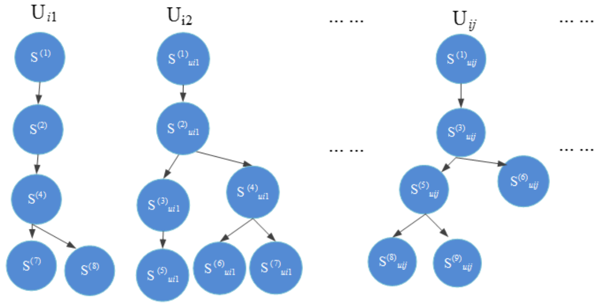

Fig. 4. User timing state transition

Thus, a temporal behavior of the user Ui can be represented as a series of actions of the user's sequence of actions when it is accessed by a particular cloud service Cj , as shown in formula 2,

$$< Ui \to Cj >= \left\{ S\left(Cj\right)\ \left(1\right)ui,\ S\left(Cj\right)\ \left(2\right)ui, \ldots,\ S\left(Cj\right)\ \left(n\right)ui \right\} \qquad (2)$$

which can be reduced to:

$$Uij = \left\{ S\ \left(1\right)uij,\ S\ \left(2\right)uij, \ldots,\ S\ \left(n\right)uij \right\} \qquad (3)$$

The user's set of actions for access to all services can be represented as:

$$< Ui \to C >= \left\{ Ui \to C1,\ Ui \to C2, \ldots \right\} = \left\{ Ui1,\ Ui2, \ldots \right\} \qquad (4)$$

3.3 Sequence Coding Structure of User Time Behavior

The coding structure of user timing behavior should be able to reflect the following elements:

- **Service Identification (C).** Used to indicate the specific cloud services invoked by this behavior. In order to improve the mining efficiency, the system can be classified according to the service identification.

- **Fitness (F).** The main function of fitness is to determine the probability of each candidate sequence from the next generation.
- **User ID (U).** It is used to represent the identity of the user.
- **Category Identification (A).** Is used to represent the meaning of this sequence, such as "abnormal login".
- **Sequence Behavior Sequence (S).** Used to store user timing behavior:

$$\left\{S(1)_{Uij}, S(2)_{Uij}, \ldots, S(n)_{Uij}\right\}.$$

The sequence structure of the user timing behavior is shown in figure 5.

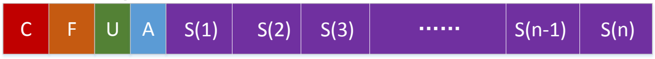

Fig. 5. Coding structure of user time behavior sequence

According to the coding structure, the user time sequence is represented as a character string based on (0, 1). Set the total number of reference sequences is n, x_t^i represents the i sequence of pattern mining in the t generation. It should be pointed out that the adaptive pattern mining can only change the sequence of the S sequence in the coding structure, and calculate the fitness F according to the new sequence. So that each individual can be represented as a formula 5, where m represents the number of actions in this sequence.

$$x_t^i = \{C_t^i F_t^i U_t^i A_t^i S_t^i(1) S_t^i(2) \ldots\ldots S_t^i(m)\} \quad (5)$$

The reference set of the T generation can be expressed as the formula 6.

$$X_t = [x_t^1 x_t^2 x_t^3 \ldots\ldots x_t^n]^t \quad (6)$$

3.4 An Algorithm for Mining User Temporal Behavior Patterns Based on Adaptive Coding

This algorithm uses the genetic algorithm to improve the traditional pattern mining algorithm, so that the system optimization in the complex environment is achieved. The genetic algorithm is used to increase the number and diversity of the reference sequence, and the efficiency and accuracy of the subsequent anomaly detection can be improved by updating the fitness parameters of the reference sequence.

Selection Algorithm. The main flow of the algorithm is divided into the following three steps: According to a certain proportion, it is the first choice to select the reference sequence with higher fitness to enter the next generation reference set directly. Then, all reference sequences in the previous generation reference set are classified according to the service identifier C. In each category, the higher the degree of adaptation of the sequence, the greater the probability of being selected by the system.

It is assumed that the sequence number of the previous reference set is n, and the n is large enough to meet the requirements of the fitness proportional selection method to the next generation. The selection ratio of the highest fitness sequence is M. The total number of cloud services is C. The fitness of each sequence is $f(x)$. The probability that each sequence is selected is:

$$P = \frac{f(x)}{\sum_{n=1}^{N} f(x)} \quad (7)$$

The expected number of the sequences which are selected into the next generation reference set from the previous generation reference set is:

$$E = n \times \frac{f(x_i)}{\sum_{i=1}^{n} f(x_i)}, i \in (1,2,3 \ldots n) \quad (8)$$

The total number of next generation reference sequence is:

$$N_2 = N_1 \cdot M + E \quad (9)$$

The above procedure not only ensures that the sequences with higher fitness can be saved to the next generation, but also improves the computational efficiency.

Crossover Algorithm. In nature, the two chromosomes of the organism can be exchanged between each other, and then into a new chromosome of the two. Crossing behavior can ensure the diversity of species. Crossover operator is also introduced in genetic algorithm. Before the crossover operation, all the individuals in the population need to be matched randomly, and then each pair of individuals will be cross operated according to the preset crossover probability parameters. The following is single -point crossover operator which is used in this paper.

The value of the intersection K range of $[1, L-1]$, L represents the number of bits in a sequence, With this point as the boundary of the exchange of variables. Such as:

Parent entity 1 **1 0 0 1 0 1 0 0 1 1**
Parent entity 2 **1 1 0 1 1 1 0 1 1 0**

The position of the intersection point is 5, and then the two sub groups are generated:

Child entity 1 **1 0 0 1 0 1 0 1 1 0**
Child entity 2 **1 1 0 1 1 1 0 0 1 1**

After that, the system calculates the fitness of each individual, and retains the improved sequence to the next generation.

Mutation Algorithm. A gene on a chromosome in a living organism may mutate because of environmental factors, and it becomes a new chromosome. In genetic algorithm, mutation operator is also introduced. It can improve the overall diversity of the population, and ensure the convergence of the algorithm, and to prevent the occurrence of premature.

The probability of sequence variation is P_v. For sequence $C = \{x1x2x3 \ldots xn\}$, its mutation algorithm is as follows:

$$x' = \left\{ \begin{array}{l} 1-x, r_x \le P_v \\ \quad x, r_x > P_v \end{array} \right\} \qquad (10)$$

x' is the result of variation of coding bit, r_x is a random quantity of each bit in a sequence, $r_x \in [0,1]$. In general, the value of P_v is between 0 and 0.001, so most of the sequences do not change.

4. Experimental Simulation

The main purpose of this experiment is to verify the dynamic performance and convergence performance of adaptive pattern mining algorithm [14]. In this experiment, the Java language is used to encode, and the experiment is carried out on the sequence data set synthesized by the IBM data set generator. The operating environment of the experiment is: the operating system is Win7 with 2GB memory. The detailed parameter settings of the experiment are shown in table 1.

Table (1) Simulation parameter settings

Parameter	Value
Minimum confidence	0.5
Initial population size	100
Crossover probability	0.5
Mutation probability	0.005

The dynamic performance of the pattern mining algorithm is that the average fitness of the offspring sequences generated by the algorithm runs all the time [15]. Its formula is shown in formula 11. $f(t)$ is the sum of the sequence fitness of t moments.

$$F(t) = \frac{1}{T}\left(\sum_{t=1}^{T} f(t)\right) \qquad (11)$$

Figure 6 shows the comparisons between the results of simulation of this algorithm and traditional genetic algorithm in dynamic performance of the. After 80 generations, the group adaptability of this algorithm is slightly lower than the traditional algorithm. These results indicated that there were more species and higher diversity in the population. The algorithm can guarantee the individual to develop in many directions, so as to avoid premature convergence.

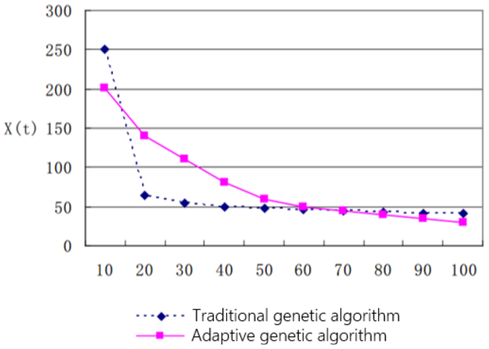

Fig. 6. Dynamic performance of algorithm

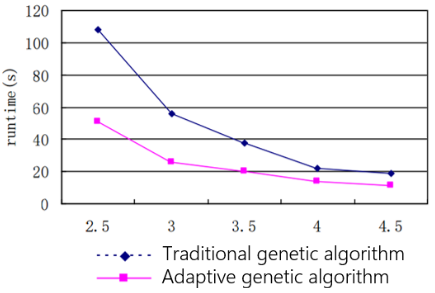

Fig. 7. Algorithm operating time

The experimental results shown in Figure 7 shows the execution time of the two algorithms at different confidence levels. It can be found that with the increase of the minimum confidence, the running time becomes smaller. We can see that the adaptive genetic algorithm is faster than the standard genetic algorithm in the case of the minimum confidence.

5. Conclusion

The main content of this paper is the research of user behavior pattern mining algorithm based on adaptive coding. The core idea of this algorithm is to introduce the genetic algorithm to realize the self- optimization of pattern mining algorithm. This paper first introduced the concept of pattern mining and genetic algorithm. Then, the formal description and coding rules of user behavior sequences were proposed, which laid the foundation for pattern mining. Secondly, the selection, crossover and mutation algorithms in adaptive coding were introduced. Finally, the simulation results showed the efficiency of the algorithm.

In our study, there are still a lot of work waiting for further study, such as the function coupling, the main mining, the modeling, and optimization methods.

Reference

[1] KMI Media Group, “Global information grid enterprise engineering”, 2008 DISA Year in Review, pp. 8-12, 2009,

[2] KMI Media Group, “RACE and cloud computing”, 2008 DISA Year in Review, pp. 14-16, 2009,

[3] NASA, “NASA’s nebula cloud computing technology to play key role in new open source initiative”, National Aeronautics and Space Administration, 2010,

[4] F Sardis, G Mapp, J Loo, et al., “On the investigation of cloud-based mobile media environments with service-populating and QoS-aware mechanisms”, IEEE transactions on multimedia, pp.769-777, 2013,

[5] Park J-S, Chen M-S, Yu P.S, “Efficient parallel data mining for association rules”, ACM, International Conference on Information and Knowledge Management, pp. 31-36, 1997,

[6] Ng R, Han J, “Efficient and effective clustering method for spatial data mining”, Morgan Kaufmann Publishers Inc. International Conference on Very Large Data Bases. Vol. 88, pp. 144-155, 1994,

[7] Lee D, Lee H, Park D, et al., “Proxy based seamless connection management method in mobile cloud computing Title”, Cluster Computing, pp. 733-744, 2013,

[8] Champati J P, Liang B, “Energy Compensated Cloud Assistance in mobile cloud computing”, University of Toronto, Canada, Department of Electrical and Computer Engineering, pp. 392-397, 2014,

[9] Agrawal R, Srikant R, “Mining sequential patterns”, IEEE Xplore, Eleventh International Conference on Data Engineering, pp. 3-14, 1995,

[10] Wang X, “Genetic Algorithm and Its Application”, Publisher, Mini-Micro Systems, 1995,

[11] Ming Z, Shudong S, “Principle and application of genetic algorithm”, National Defense Industry Press, 2009,

[12] Zhengjun Pan, Lishan Kang, Yuping Chen, “Computation of the evolutionary”, Tsinghua University Press, 07 2008,

[13] Vonk E, Jain L C, Johnson R P, “Mathematical Foundations of Genetic Algorithms”, Automatic generation of neural network architecture using evolutionary computation, pp. 60-78, 2015,

[14] Wang J, Han J, “BIDE: efficient mining of frequent closed sequences”, Proceedings. IEEE, International Conference on Data Engineering, pp. 79-90, 2004,

[15] Davis L, Orvosh D, “The Mating Pool: A Test Bed for Experiments in the Evolution of Symbol Systems”, Pittsburgh, Pa, Usa, July. DBLP, International Conference on Genetic Algorithms, pp. 405-412, 1995,

AUTHOR’S PROFILE

Tong Zhang, Master, who was born in February, 1992. His research interests cloud computing, network security theory, etc.

Ruijuan Zheng, Doctor, Master's Supervisor, who was born in January, 1980. Her research interests involve bio-inspired multi-net security technology, network security theory and method based on autonomic computing, etc.