ISSN 2616-5775 Vol. 7, Issue 1: 35-41, DOI: 10.25236/AJCIS.2024.070106

Library User Feature Data Generation Based on Genetic Algorithm

Chu Chenhui[1,a,*], Li Guohui[1]

[1]Department of Management Science and Engineering, Hebei University of Engineering, Handan, 056038, China
[a]huixxx0314@163.com
[*]Corresponding author

Abstract: *With the advent of the era of big data, libraries are faced with the need to make more use of user characteristic data to provide personalized services. However, due to the difficulty of acquiring real user characteristics data, it is difficult to carry out effective analysis and application. This paper proposes a genetic algorithm-based approach to generate simulated user feature data to enhance library service quality. Through genetic algorithms, we can simulate users' borrowing behavior, search behavior and participation behavior. The experimental results show that the generated simulated data is basically consistent with the real data in the distribution of behavior characteristics, which can provide accurate user characteristics data for the library, so as to better meet the needs of users and provide personalized services.*

Keywords: *Genetic algorithm, User characteristic data, Library personalized service*

1. Introduction

With the rapid development of science and technology and the explosive growth of information, mankind has entered the "big data era". In this digital era, libraries not only need to provide rich resources and services, but also need to customize services according to users' individual needs, so as to provide library experience that is closer to users' needs[1]. The effective use of user characteristic data becomes the key to realize personalized service[2]. However, the process of obtaining real user characteristics data and applying them to library services is not easy. The acquisition of user characteristic data involves privacy issues, and the collection of real data also faces many restrictions[3]. In addition, a large number of user characteristics data need to be analyzed and processed to extract useful information to provide libraries with guiding decisions and directions for improving services. Therefore, we urgently need an innovative method to generate simulated user characteristics data to make up for the deficiency of real data. General methods of data generation model include: (1) simulation sampling, represented by probability distribution model; (2) approximation methods, including random walk model[4], autoregressive model[5] and variational autoencoder[6]; (3) Implicit method, the representative model is generative adversarial network (GAN)[7].

This paper will use genetic algorithm to distinguish from the above three categories, which is an optimization method that mimics the process of natural selection and evolution[8]. Genetic algorithms are used in various fields: Feng Junchi et al studied the data generation application of genetic algorithms in computer software testing[9]. Zhang Yingli et al. developed a method of generating guqin music with genetic algorithm[10]. In this paper, by simulating users' borrowing behaviors, search behaviors and participation activities, we can generate simulated data similar to real data[11]. Specifically, we use genetic algorithm to cascade code each feature in the library user feature data, combine the codes together, form an initial population, and then obtain fitness function according to the probability distribution of the original data to optimize and upgrade the population, evaluate the fitness value, and obtain the optimal solution or reach the end condition of the iteration. Finally, the simulation generated data is compared with the original data to draw a conclu.

ISSN 2616-5775 Vol. 7, Issue 1: 35-41, DOI: 10.25236/AJCIS.2024.070106

2. Genetic algorithm

2.1. Genetic algorithm Overview

Genetic Algorithm is an optimization algorithm based on the theory of biological evolution. By simulating the process of biological evolution in nature, genetic algorithm searches and optimizes the optimal solution or approximate optimal solution of the problem[12]. Genetic algorithms were first proposed by John Holland in 1975 and have been widely studied and used in the following decades. The basic idea of genetic algorithm is to optimize the quality of the solution step by step by simulating the process of heredity, crossover, variation and fitness selection. In each generation, individuals with higher fitness will be more likely to pass on good genes to the next generation through selection operations, thereby gradually optimizing the quality of the solution. The advantage of genetic algorithm is that it can deal with complex optimization problems, but it also needs reasonable parameter setting and fitness function design to get better results.

2.2. Data generation process based on genetic algorithm

(1) Define questions, objectives, and data representation: define questions and objectives of library user feature data that need to be generated, and generate data that conforms to real user features. Determine how user profile data is represented, such as user age, type of action, gender, book category, and so on. The user profile data can be represented using an appropriate data structure or encoding.

(2) Initialize population and fitness evaluation: randomly generate a set of initial user feature data as a population, and each individual represents a user feature data. A fitness function is defined to assess the fitness of each individual in the population, i.e. the degree to which the user's characteristic data is reasonable or meets the goal. The fitness function can consider the diversity, frequency and timing of user characteristics.

(3) Selection, crossover and mutation operations: according to the fitness size, select a part of the individual as the parent for generating the next generation of individuals. Two individuals are selected from the parent generation and a pair of new individuals are generated by crossing operations. The user characteristic data of the new individual is modified to introduce some randomness and change.

(4) Judgment of termination conditions and output results: Repeat the above steps until the termination conditions are met, such as reaching the maximum number of iterations and the amount of generated data meeting the requirements. If the termination condition is satisfied, the generated user characteristic data is output as the solution of the problem.

3. Construction of data generation model based on genetic algorithm

3.1. Coding design

Table 1: Library user data binary coding table

User characteristic parameter	Gene location	Genotype	Phenotype
Gender	1	0,1	0:Female;1:Male
Age	2~7	000000~111111	Convert binary to decimal 0~64years old
User type	8~11	0000~1111	E.g. 1001: Citizen card reader; 0111: General reader
Book type	12~16	00000~11111	The National Library classification of China is 01000: I (Literature); 01001: J (Art)

Coding is the primary problem to be solved when applying genetic algorithm [13]. When designing coding, it is necessary to choose the appropriate coding method according to the characteristics of the problem. The coding method should be able to effectively represent the solution space of the problem, and take into account the diversity of the population and the search ability. In addition, the encoding should be feasible, that is, the decoding operation can convert the chromosome encoding back to the actual solution. Coding design will directly affect the performance and result quality of genetic algorithm. So far, many different coding methods have been proposed, the common coding methods include binary coding, real coding and permutation coding.

ISSN 2616-5775 Vol. 7, Issue 1: 35-41, DOI: 10.25236/AJCIS.2024.070106

One of the user characteristics data in this paper is an individual X, because for an individual X it contains different characteristics, such as the user's age, gender, the type of books and so on, so ordinary binary coding can not meet the needs of expressing individual characteristics. Therefore, this paper makes an improvement on the basis of binary and adopts the method of cascade coding to encode individuals. First, the characteristics of each variable are encoded into a fixed length binary string, and then the encoding is connected in order to become a new binary string, as shown in Table 1:

If one of the individuals 1011000100101000 is encoded as shown in the binary coding table of the population, its meaning is shown in Table 2.

Table 2: Examples of user characteristic data parameters

User characteristic parameter	Genotype	Phenotype
Gender	1	Male
Age	011000	24 years old
User type	1001	Citizen card reader
Book type	01000	Chinese library classification I

3.2. Fitness function design

In this paper, by simulating the probability distribution of the original data, the relationship between different features in the original data is found, and the probability formula is constructed by these existing relations. The user feature data processed in this paper is similar to a decision tree structure, each branch has its own probability, and the probability of each case is different according to the different path chosen. The formula for calculating the probability is shown in equation (1).

$$fitness(x) = \sum_{i=1}^{n} U_i \times V_i \qquad x = 1,2,\cdots,n \tag{1}$$

Where: fitness(x) represents individual fitness; U_i, V_i Represents the probability of decoding phenotypes corresponding to individual features.

3.3. Genetic operator

3.3.1. Crossover and variation

Cross operation is an important operation for generating new entities. It produces a pair of new individuals by combining the chromosomes of two parent individuals in a certain way. In this paper, a single point crossing is adopted, a random crossing point is selected, the chromosomes of two paternal individuals are cut at this point, and then the gene fragments after the cutting point are exchanged to generate two new individuals. By exchanging and combining genetic information, the diversity of the population is increased and new potential solutions are introduced.

A mutation operation is an operation used to introduce randomness and variation. It produces new individuals by randomly changing their chromosomes. In this paper, point mutation is used to select a random gene location, and then the gene value on the gene location is reversed according to a certain mutation probability. By introducing randomness and variation, it helps to avoid falling into local optimal solutions and provides an opportunity to search for global optimal solutions.

3.3.2. Selection operator

The selection operation is an operation used to select a superior individual as a parent. According to the fitness of the individual, a part of the individual is selected as a breeding pool for producing the next generation of individuals. This paper adopts the roulette method[14], the core idea of which is that the probability of each individual in the selection process is proportional to its fitness value according to its fitness value. Then, individuals from the breeding pool are selected by random selection. Individuals with higher fitness have a higher probability of being selected. Suppose there are n chromosomes X, calculate the fitness value of each individual according to the binary value of individual X, and then add the fitness value of each individual to obtain the sum, divide the fitness value of a single individual by the total fitness value, calculate the probability of individual selection, and then add up the probability of each individual.

$$p(x_i) = \frac{f(x_i)}{\sum_{i=1}^{n} f(x_i)} \qquad i = 1,2,3\ \cdots,n \tag{2}$$

Where: $p(x_i)$is the probability of the individual being selected, $f(x_i)$is the fitness of the individual.

ISSN 2616-5775 Vol. 7, Issue 1: 35-41, DOI: 10.25236/AJCIS.2024.070106

Randomly generate a uniformly distributed random number r at [0,1], such as r<q_i, then chromosome x_i is selected. if q_{n-1}<r<q_{n+1} then chromosome x_n is selected. Where q_i is called the cumulative probability of chromosome x_i.

$$q_i = \sum_{k=1}^{i} p(x_k) \qquad k = 1,2,3 \cdots i \tag{3}$$

The termination condition of the algorithm is to reach the maximum number of iterations, which can avoid falling into a dead loop and ensure the distribution of the diversity of user feature data.

4. Empirical analysis

4.1. Experimental data source and parameter setting

The data set of this paper adopts a library data set, and calculates a total of 177556 library book borrowing records in 2022. Due to the large amount of data, we will preprocess these data first, and first put them into MATLAB calculation from the library. Then, from the user feature data, remove too small and missing data to prevent the impact on the proportion of data features, and select the required user features and resource features. Finally, from the processed data, find the relationship between various features, statistical rules. After many calculations and adjustments, the initial population number N=40000, crossover probability PC=0.6, mutation probability PM=0.15, and iteration times GER=300.

According to the original data, the proportion of each reader type in all reader types is calculated first, and then the proportion of the types of books borrowed by each reader type is calculated, as shown in Figure 1, in which the proportion of the types of books borrowed by citizen card readers is shown in Figure 2.

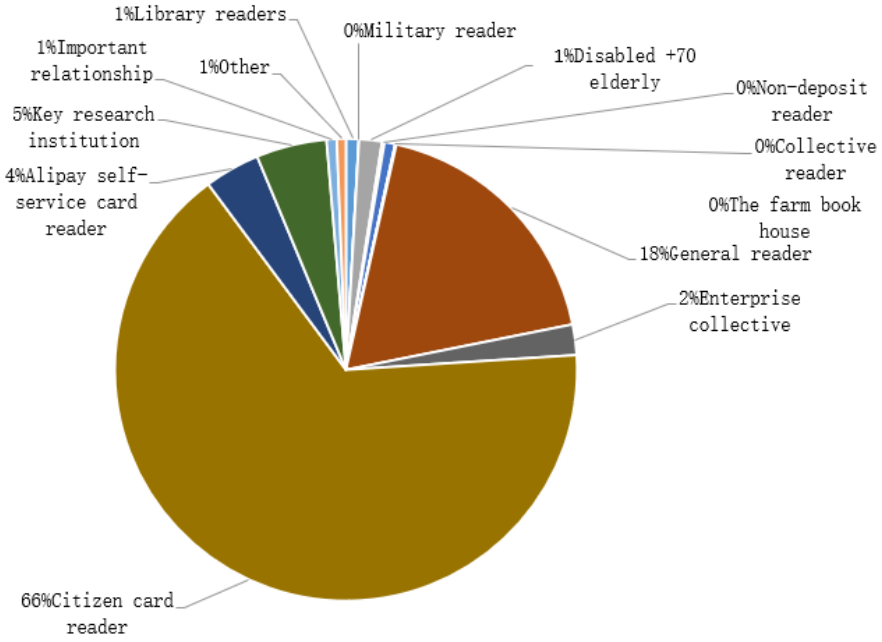

Figure 1: Readers type proportion

ISSN 2616-5775 Vol. 7, Issue 1: 35-41, DOI: 10.25236/AJCIS.2024.070106

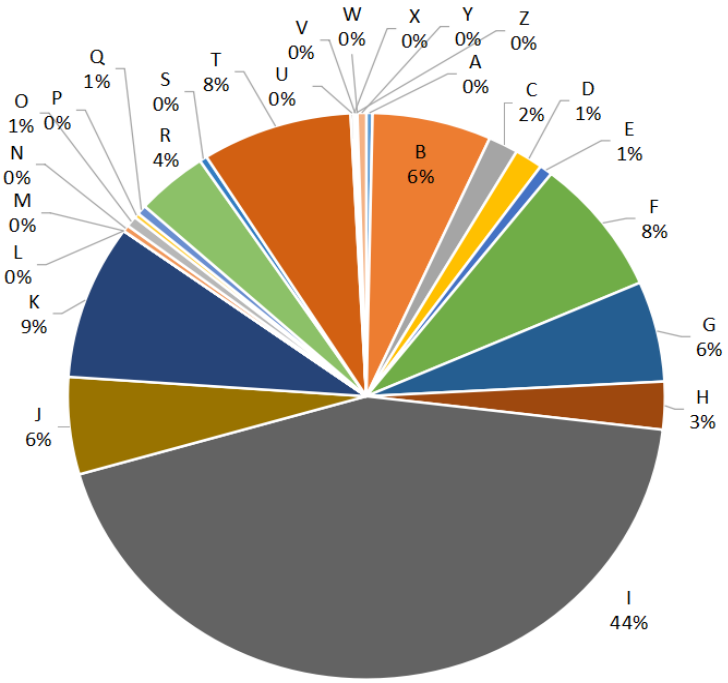

Figure 2: The proportion of books borrowed by citizen card readers

The ratio of generated data for different initial populations is shown in Table 3.

4.2. Generated data analysis

Genetic algorithm was used to conduct data generation test, and the statistically generated simulation data was shown in Figure 3. According to the simulated data obtained by statistical genetic algorithm, it is found that the top 3 types of books borrowed by citizen card readers are I (literature) type 40%, K (history and geography) type 10%, and F (economy) type 7%. As can be seen from Figure 2, the top three types of books borrowed by original citizen card readers are I (literature) type 44%, K (history and geography) type 9%, and F (economy) type 8%. By comparing the original data with the simulated data generated by genetic algorithm, the feasibility of genetic algorithm in generating user feature data is proved.

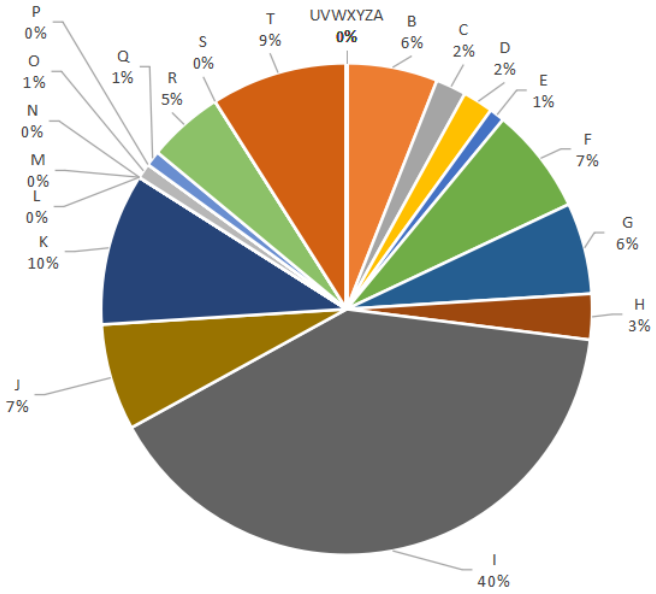

Figure 3: Generate data citizen card reader borrowing percentage

ISSN 2616-5775 Vol. 7, Issue 1: 35-41, DOI: 10.25236/AJCIS.2024.070106

The ratio of generated data for different initial populations is shown in Table 3.

Table 3: Proportion of data generated by different initial populations

	10000 pieces	20000 pieces	40000 pieces	Raw data
I	36%	38%	40%	44%
K	10%	11%	10%	9%
F	9%	9%	9%	8%

In the experiment, with a fixed number of iterations, when the initial population size is small, the data generated by genetic algorithm differs greatly from the original data in terms of the proportion of the types of books borrowed by users. However, by increasing the number of initial populations, the proportion of data generated gradually approaches the distribution of the original data. This is because when the initial population number is small, the fitness function has no obvious influence on the proportion of population individuals in the whole, but with the increase of the initial population number, the role of fitness function gradually appears. This experiment verifies the practicability of genetic algorithm to generate data, which can extract joint distribution probability from a small amount of data and generate a large amount of data. At the same time, the generated data can well protect the user's privacy.

In order to verify the accuracy of the generated data, 100 borrowing records of citizen card readers were randomly selected from the original data, and the distribution of data resource features with the same user characteristics as the selected records were compared. Taking the borrowing record of a user as an example, the original data indicates that the user borrows two books of type F and type I respectively, while the characteristic data generated by the genetic algorithm indicates that the reader borrows three books, including 1 book of type F and 2 books of type I. Then the record is 100% accurate. If the characteristic data generated by the reader is one book of type F, one book of type I, and one book of type T, then the accuracy of the record is 66.7%. Further extract the borrowing records whose user characteristics are the same as those of 100 randomly selected records in the generated data, and make statistics on their borrowing data characteristics, as shown in Table 4.

Table 4: Statistical comparison of data characteristics

User type	Gender	Age group	Raw data	Generated data	Same feature type	Accuracy rate
Citizen card reader	Male and female	0~64	100	2000	1461	73%

Although the user characteristics in the generated data are the same as the selected records and the type of book is the same, the records are not exactly the same due to differences in other information. According to the statistical results, the accuracy of the data generated by the genetic algorithm is about 73%.

5. Suggestions and conclusions

By using genetic algorithm to generate user characteristic data, the personalized service of library can be enhanced. The experimental results show that the genetic algorithm can extract the joint distribution probability from a small amount of original data and generate a large number of borrowing data conforming to user characteristics. This can avoid the direct exposure of real user data and protect users' personal privacy information. The generated data can reach a high level of accuracy in user characteristics. This means that the data generated by genetic algorithms can be used as an effective way of data enhancement, providing libraries with more personalized service possibilities.

This paper focuses on the application of genetic algorithm, but there is still a lot of room for improvement. We can optimize the parameters of genetic algorithm, consider more feature information, try other algorithms, and consider the balance of data distribution. Through the improvement of the above aspects, the effect and application of genetic algorithm to generate user feature data are further improved.

References

[1] Li Lin.Research on the development status and optimization direction of personalized service of smart Library [J]. Media Forum, 2019,6(16):112-114
[2] Yu Xiaoji. Research on the protection of user privacy in Personalized Library Service in the era of

ISSN 2616-5775 Vol. 7, Issue 1: 35-41, DOI: 10.25236/AJCIS.2024.070106

Big Data [J]. Information Science,202,40(09):147-153
[3] Jiang Panpan. Discussion on ways to protect readers' Personal information in the era of Big Data [J]. Library Work and Research,2019(06):11-15
[4] Reimer P J, Bard E, Bayliss A, et al.IntCal13 and MARINE13 radiocarbon age calibration curves 0-50,000 years cal BP[J].Radiocarbon, 2013, 55(4):1869–1887
[5] Babatunde O T , Oranye H E , Nwafor C N .Volatility of Some Selected Currencies Against the Naira Using Generalized Autoregressive Score Models[J].International Journal of Statistical Distributions and Applications, 2020(3)
[6] Sun Ling, Han Lixin, Gou Zhinan. Dynamic subject Model based on Variational autoencoder [J]. Hebei Industry Science and Technology, 2017, 34(06):421-427
[7] Pan Z , Yu W , Yi X ,et al.Recent Progress on Generative Adversarial Networks (GANs): A Survey[J]. IEEE Access, 2019:36322-36333
[8] Chahar V, Katoch S, Chauhan S S.A Review on Genetic Algorithm: Past, Present, and Future[J].Multimedia Tools and Applications, 2020(4)
[9] Feng Junchi, Yu Lei. Improvement of Genetic Algorithm in Test data Generation [J]. Journal of Computer-Aided Design and Graphics, 2015, 27(10):2008-2014
[10] Zhang Yingli, Liu Hong. A method of generating Guqin music based on Genetic Algorithm [J]. Information Technology and Informatization, 2018(09):28-30
[11] Fang Wenhui, Hu Zhulin, Zhu Xinjuan. User behavior data generation based on Genetic Algorithm [J]. Foreign Electronic Measurement Technology, 2021, 40(09):154-159
[12] Zheng Liping, Hao Zhongxiao. Review of Genetic algorithm theory [J]. Computer Engineering and Applications, 2003, (21):50-53+96
[13] Zhang Chaoqun, Zheng Jianguo, Qian Jie. Genetic algorithm coding scheme comparison [J].Application Research of Computers, 2011, 28(03):819-822
[14] Li Minghui, Meng Xiankun. Optimization Design of production scheduling Model for papermaking enterprises using Roulette Method [J]. Journal of Shaanxi University of Science and Technology (Natural Science Edition), 2012, 30(02):44-48