Fast Recommendation Method of Personalized Tourism Big Data Information Based on Improved Clustering Algorithm

Yi-lin Feng(✉), He-qing Zhang, and Cai-ting Peng

Management (Tourism) School of Guangzhou University, Guangzhou 510006, China

Abstract. The conventional tourism big data information recommendation method does not reduce the search scope of the database, resulting in a long running time of the algorithm. Therefore, based on the improved clustering algorithm, a fast personalized tourism big data information recommendation method is designed. The improved clustering algorithm is designed and the mathematical model of clustering algorithm is established. The semantic similarity of tourism information is calculated. Based on the improved clustering algorithm, the retrieval range of database is reduced, and the classification model of tourist attractions is established, so as to improve the speed of tourism big data information recommendation method. In the experiment of testing the improved clustering algorithm and test information recommendation method, the experimental data of the control group are better than the three control groups. Therefore, the fast recommendation method of personalized tourism big data information based on the improved clustering algorithm is better than the three conventional methods.

Keywords: Digital technology · Civil engineering · CAD drawing · Teaching assistant system

1 Introduction

With the rapid development of social networks, a large number of tourism related communities and websites have emerged as carriers, resulting in a huge scale of tourism information and data. After traveling, tourists like to publish their travel experience on the Internet. These travel experiences are shared online in the form of text, pictures and videos. Travelers from all over the world combine their first-hand information to form a huge tourism knowledge base. These tourism data are not only huge in quantity, but also various in types, such as picture, video, text, audio, geographical location and so on. The structure of these tourism data has strong randomness. Because of user personality, preferences and other factors, it has a very irregular data structure. Internet tourism data has formed the characteristics of big data and cross media tourism big data. Cross media tourism big data contains great application value and social benefits, and its related search, storage and knowledge mining become the research focus, and become a hot issue in the field of information retrieval and data mining [1].

S. Liu and X. Ma (Eds.): ADHIP 2021, LNICST 417, pp. 284–296, 2022.
https://doi.org/10.1007/978-3-030-94554-1_23

Based on the global cooperative information optimization algorithm, the paper summarizes many travel information, including text mining, Geographic Image Mining, travel route mining, and so on, and labels in detail, so that it can be applied in various scenes [2]. This information recommendation method needs to make detailed statistics and analysis of the data in the network, which takes a long time. The paper [3] through a recommendation model which combines content and relationship, uses support vector machine to train the data set of landmark scenic spots, and obtains a picture retrieval algorithm with high accuracy. This information recommendation method mainly focuses on the calculation of picture information, and has no good processing ability for text information. The literature [4] compares the past tourism information comprehensively with the personalized adaptive hybrid recommendation algorithm, and obtains a better semantic model of big data. The accuracy of push information is optimized. This method depends on the past data. If the error rate of existing data is large, it is difficult to guarantee the accuracy of the information itself.

This paper designs an improved clustering algorithm based on the above literature, and designs a fast recommendation method for personalized tourism big data information. The improved clustering algorithm is designed and the mathematical model of clustering algorithm is established. Calculate the semantic similarity of tourism information, reduce the retrieval range of the database based on the improved clustering algorithm, and establish the classification model of tourist attractions, so as to improve the running speed of tourism big data information recommendation method.

2 Design of a Fast Recommendation Method for Personalized Tourism Big Data Information Based on Improved Clustering Algorithm

2.1 Improved Clustering Algorithm

Fuzzy clustering algorithm can not be directly applied to the data set with missing data, so this paper proposes a completion algorithm which can be applied to the missing value filling. Firstly, the geometric structure of the data set with missing values is analyzed. Taking a simple two-dimensional data set as an example, the two-dimensional data set can be reflected in the plane coordinate system. For the two-dimensional data set X with missing values, for example, $N_k = (N_1, N_n)$ represents the missing point in the ordinate, which is represented by a straight line passing through point x_{k1} parallel to the ordinate in the plane, $N_k = (N_n, N_2)$ represents the missing point in the abscissa, which is represented by a straight line passing through point N_{k1} parallel to the abscissa in the plane, and $X_k = (X_{n1}, X_{n2})$ is meaningless in the clustering algorithm [5]. Therefore, the data set should not contain all attribute missing data groups. Therefore, for two two-dimensional datasets with missing values, the black dot represents the data points that are not missing, the straight line represents the data with missing attributes like $N_k = (N_{k1}, N_n)$, and the circle represents the clustering boundary of the corresponding complete dataset. If the cluster number $c = 2$ of each two-dimensional dataset with missing values is known, the missing values can be reasonably estimated. With the increase of data volume and dimension, missing data can be estimated more accurately.

From this, four missing data filling methods based on improved fuzzy clustering algorithm can be obtained, namely: complete data strategy, partial distance strategy, optimized completion strategy and best model strategy [6]. Among them, the optimized completion strategy is the best way to fill in missing data. Assume that the data sample set is $N_n = \{N_1, N_2, \cdots, N_n\}$, where N_k is the k-th w-dimensional data vector in the data set N, and satisfies $1 \leq k \leq n$. Among them, N_j is the jth value in the N_k vector, and $1 \leq j \leq s$, $1 \leq k \leq n$. The data set N contains missing values, which can be represented by N_n, let:

$$N_M = \left\{ N_j = N_n | 1 \leq j \leq s, 1 \leq k \leq n \right\} \tag{1}$$

The N_M in the formula is extremely missing the data set [7]. For example, when $k1 =$ 3B, $k2 =$ 4CC, the data set is set as:

$$N = \left\{ \begin{bmatrix} 1 \\ N_{k1} \\ 7 \end{bmatrix} \begin{bmatrix} 2 \\ 4 \\ 8 \end{bmatrix} \begin{bmatrix} 3 \\ 5 \\ 9 \end{bmatrix} \begin{bmatrix} N_{k2} \\ 6 \\ N_{k3} \end{bmatrix} \right\} \tag{2}$$

At this time, $N_n = \{n_1, n_2, n_3\}$ is the mathematical model of the clustering algorithm.

2.2 Semantic Similarity Calculation of Tourism Information

Firstly, the tourism information is preprocessed, and a vector composed of multiple search keywords is used to represent the tourism information similarity vector model to improve the accuracy of retrieval. On this basis, the travel information in the database is classified into temporal semantics, and its temporal semantic feature vector is extracted [8]. At this point, a vector model should be established first, so that the vector can represent tourism information in this feature space, and all the temporal semantic sentences related to tourism information in the database are extracted to establish this vector model. Each feature in the vector model has its feature value, which is the weight of the feature. Each temporal semantic keyword is one of the characteristic dimensions. Assuming that there are i total temporal semantic keywords in the database, the characteristic dimension in the database is also i. For a tourism information F_i, the frequency of temporal semantic keyword F_j in the database can be used to calculate the weight of temporal semantic keyword, and the similarity is used as its continuous measurement index [9]. The formula for calculating the frequency of tense semantic keywords is as follows:

$$F_{ij} = \frac{\delta_{ij}}{\lambda_j} \tag{3}$$

Among them, F_{ij} represents the frequency of occurrence of tense semantic keywords, δ_{ij} represents the number of occurrences of tense semantic keywords in travel information, and λ_{ij} represents the number of occurrences of keywords with the most frequent occurrence of tense semantic keywords in travel information. This formula is mainly used to calculate the relative frequency of a tense semantic keyword in travel information. The temporal semantic keywords appearing in the travel information are converted into a vector in the order of frequency, the content of the travel information is extracted,

attributes and attribute values are extracted, and the temporal semantic vector describing the content of the travel information is described. Each travel information has different temporal semantic characteristics, so when calculating travel information similarity, the problem of conceptual attributes needs to be considered first. Suppose that the concept M has an instance m. At this time, the instance m can be represented as $m_y = S_{IN}[T_p]$, where $p = (p_1, p_2, \cdots, p_m)$ is the same in the instance n, $n_y = S_{IN}[T_q]$, $q = (q_1, q_2, \cdots, q_n)$ [10]. At this time, the similarity of instance n and m can be calculated. First, the attribute vector of instance m and n is a common attribute vector through the above method, and then the similarity of tourism information is calculated according to the attribute value, and the attribute values of the two instances are compared. And similarity, the obtained formula is shown below.

$$Sim_T(p, q) = \sum_{t=1}^{n} \frac{\delta_{ij} + \lambda_j}{2} \cdot Sim_T(p_1, q_1) \tag{4}$$

Among them, δ_{ij} and λ_j are the weight coefficients of attributes p_1 and q_1 in each vector, which are preset parameters. They are usually the statistical values obtained after preprocessing the tourism information. The final weight value of the tourism information is determined by this statistical value, and its value range is [0.1]. By substituting the above similarity into the semantics of tourism information, the similarity between the tourism information and the standard semantics can be calculated.

2.3 Reducing the Retrieval Range of Database Based on Improved Clustering Algorithm

The most important step of classifying text data is to retrieve the text data features. The mathematical model of text data classification established above will train any parameter in the model, which is also called text data preprocessing. In the process of preprocessing, the most difficult problem to be solved is the feature retrieval of text data, which can be solved efficiently by coding. In the process of coding, the way of improving the parameter connection coefficient and its behavior track in clustering model is studied by the form of reverse link. Although it is likely to ignore some sample data in calculating behavior trajectory, when learning efficiency is high to a certain extent, the number of missing samples can be ignored compared with the whole sample set. In the text data classification mathematical model based on improved clustering algorithm, corresponding parameters need to be established in advance. The change of the parameters selected after training is generally between 0.003–0.04. The parameter selection in this paper is 0.008. When the weight value attribute no longer changes, the maximum value of training parameter fitting can be achieved, that is, the retrieval of text data characteristics is realized. According to the different characteristics of the retrieved text data, the text data can be classified. Assuming that the data text of the input layer is $h_n = \{h_0, h_1, h_2, \cdots, h_n\}$, the text data of the input layer is converted to the hidden layer in turn, so that the data becomes $g_n = \{g_0, g_1, g_2, \cdots, g_n\}$, then, through the training model of the improved clustering algorithm, the hidden layer data is converted to the data text of the output layer through

calculation. The calculation process is as follows:

$$\begin{cases} \delta_i = f(H_x y_i + H_y k_i + b_g) \\ \beta_i = H_z k_i + b_x + 1 \end{cases} \tag{5}$$

Among them, H_x is the conversion function from the input layer to the hidden layer, H_y is the conversion function within the hidden layer, H_z is the function from the hidden layer to the output layer, b_g is the deflection vector from the input layer to the hidden layer, and b_x is the deflection vector from the hidden layer to the output layer. Through formula (5), the update mode of input layer, hidden layer and output layer can be obtained, so that the state information of input layer and output layer at the previous time can be obtained by training hidden layer, and then the input and output text information at the current time can be obtained by combining with improved clustering algorithm, so as to achieve the purpose of extracting text information classification features. The improved clustering algorithm is used to train the objective function and redefine the changed function after passing through the hidden layer

$$\psi_x = -\sum_{i=1}^{n} X_i \log(x_i) \tag{6}$$

Among them, ψ_x represents the total number of text information to be classified, X_i represents the product of probability distribution and prediction value of each category after text information classification, and x_i represents the probability distribution value of each category after text information classification. The above function is a mathematical model of text data classification established by improving clustering algorithm.

2.4 Establishing the Classification Model of Tourist Attractions

Before extracting the classification features of tourist attractions, it is necessary to calculate the classification statistics of tourist attractions, and recognize the effect of automatic statistics of various tourist attractions in the classification process. In this process, there are the following parameters that will inevitably be affected. This paper uses the standard frequency, initial frequency, optimal frequency and base variable frequency to determine the classification basis of a scenic spot relative to the classification model of tourist attractions.

The first is about the change of standard frequency. In the process of eliminating standard frequency, signal overload often occurs. In order to eliminate the influence of this kind of signal, filter eliminator is usually used to reduce the signal frequency. If the short-time framing coefficient of a signal is set to $g(x)$, the adaptive function of the signal can be obtained

$$g(x) = \sum_{i=1}^{n} H_n(x) \cdot (x + g) \tag{7}$$

Where n is the number of frames of the signal; g represents the time parameter of signal elimination, and satisfies $k \in [0, n]$ generally, the value of k is between 60 Hz and 200 Hz;

$H_n(x)$ is the change range of signal windowing range function. The initial frequency is usually used to reflect the frequency frame of a tourist spot data sample, which plays a very important role in the expression of tourist spot data. When calculating the initial frequency, we can contact the classification model of tourist attractions through the specific tourism data information, and get the weighted sum of squares of the sampling points by calculating the energy mean

$$U_N = \sum_{i=1}^{N} H_i^2(\sigma_i) \tag{8}$$

Where U_N is the value of the initial frequency; N is the frame length of a tourist attraction data; i is the number of frames of this segment of scenic spot information; H_i is the average level of scenic spot information; σ_i is the electric energy calculation parameter of this section of scenic spot information. Since the average value of scenic spot information will be taken in the process of calculation, the data is an integer not less than zero. In daily calculation, the specific parameters of energy summation can be obtained in an open form.

The optimal frequency is usually due to the short-term continuous change of the scenic spot information in a certain period of time, resulting in a short-term zero energy phenomenon. This phenomenon can obviously lead to large differences between the data of various scenic spots, and this kind of phenomenon usually exists in the scenic spot information with large frequency fluctuation. The method of extracting the optimal frequency is very simple, which can be directly determined by the change of symbol

$$T_\beta = \sum_{i=1}^{n} |f_N(\beta) - \mathrm{sgn}(\beta_0 - 1)| \tag{9}$$

Where, T_β is the frame length of the scenic spot information data with short-time zero crossing; n is the number of frames of this segment of scenic spot information; $f_N(\beta)$ is the frequency coefficient of the scenic spot information. Where sgn is usually a sign function, when $\beta_0 > 0$, the sign function value is 1, when $\beta_0 < 0$, the sign function value is 0. Most of the base variable frequency is based on the analysis of the function of tourist attractions. If the computing power of the classification algorithm of tourist attractions can be linearly and positively correlated with the information of tourist attractions, the frequency of the information of tourist attractions is low; If the computing power of the scenic spot algorithm can not be correlated with the scenic spot information data, the frequency of the scenic spot information data is higher. When a frequency filter is constructed based on this concept, it can be concluded that the frequency calculation formula of the scenic spot information data is shown in (4)

$$g(a) = x_0 \cdot \ln(1 + \frac{g(b)}{x_n}) \tag{10}$$

Where $g(a)$ is the function value of frequency filter; $g(b)$ is the calculation function of frequency; x_0 and x_n represent the initial and final terms of the function values respectively. By taking the logarithm analysis of the function value, the cosine transform parameters

of the tourist attractions information data can be obtained. By accurately extracting the above four types of tourist attractions information features, we can establish a tourist attractions classification model.

2.5 Design Big Data Matching Algorithm for Tourist Attractions

To design a big data matching algorithm related to tourist attractions and tourist user information, we need to integrate the above model, and the specific algorithm structure is shown in Fig. 1.

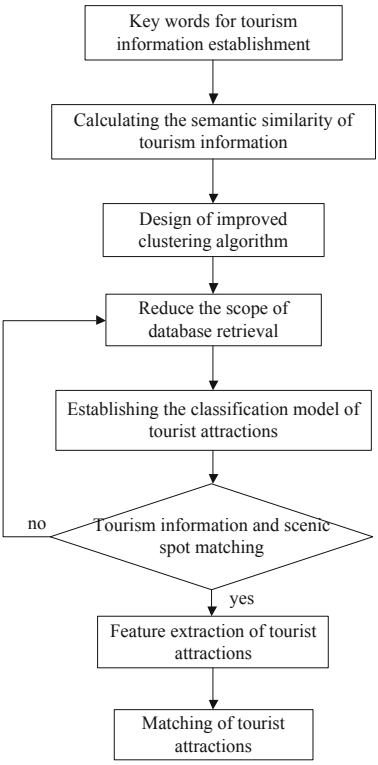


Fig. 1. Algorithm structure

As shown in Fig. 1, the classification model of tourist attractions and the similarity calculation method of semantic features of tourism information are established, and the two are matched to form the above big data matching algorithm of tourist attractions.

Taking the classification features in tourist attractions as the fourth-order dynamic error of the dynamic system, and returning the state of the system to zero, we can substitute formula (11) into the control equation to calculate its fault-tolerant approximation.

$$\lim_{x \to +\infty} a_x = b_i \tag{11}$$

Where, a_x is the node whose state is close to 0 in X system fault types; b_i is the deviation fault parameter of the system in the fault node. Through the above fuzzy logic approximation, the error of the basis vector function is calculated. In this process, the following conditions need to be ensured.

$$\frac{|s(a) - h(b)|}{2} \leq \lambda_n \tag{12}$$

Where $s(a)$ and $h(b)$ represent the approximation error of the basis vector function and the weight vector function; Represents the parametric properties of smooth functions. Based on the above formula, we can get the big data matching model of tourist attractions.

3 Experimental Study

3.1 Experimental Preparation

The main purpose of this experiment is to test the performance of the improved clustering algorithm designed above, and compare the data processing ability of the recommended method and the conventional method. Before the test, the environmental background of the experiment is explained. The test environment includes hardware environment and software environment. The details of hardware environment are shown in Table 1.

Table 1. Experimental test hardware environment

Hardware equipment	Parameter	Software required	Describe
Server under test	CPU (3.40 GHz), Memory is more than 4 G, and the total hard disk is more than 600 MB	Operating system	Windows Server2008&ubuntu 12.04
Network environment	Ethernet, or 100 m high speed network	Network protocol	IPV4
Test client	CPU (3.40 GHz), Memory is more than 4 G, and the total hard disk is more than 300 MB	Browser	Internet Explorer, Google browser, and Firefox

The software testing environment mainly includes the following professional test software, including software such as load runner, QTP, etc. The details are shown in Table 2.

Based on the above test environment, the data set is constructed. The big data of Tianchi tour guide competition is used in the data set, and several sub databases used in the algorithm test are constructed. The data set contains 1 million users' search information, 5 million scenic spots information and 30 million users' evaluation data. Delete the unqualified data, and establish a data set with 100000 user data, 1 million scenic spot data and 10 million evaluation information as the database of this experiment.

Table 2. Experimental test software environment

Name	Describe	Parameter
Operating system	Used to run the software	Include Windows xp/Windows 7/windows Server
LoadRunner	Test tools	Test model performance
QTP	Test tools	Test model security
JMeter	Test tools	Test the compressive strength of the model

3.2 Test of Improved Clustering Algorithm

For the evaluation of the improved clustering algorithm, we need to calculate the ability of data processing, the accuracy of information collection and the coverage of information search. The processing performance is used to evaluate the processing efficiency of the algorithm, which mainly measures the execution time of the algorithm, including the data preprocessing time. Accuracy is used to measure the proportion of correct items in the test set. The so-called correct items refer to the items that have appeared in the recommendation list and test set.

$$Precision = \frac{\sum_{i=1}^{n} \sum_{u \in U} |R(u) \cap T(u)|}{\sum_{i=1}^{n} \sum_{u \in U} |R(u)|} \tag{13}$$

Where *Precision* is the accuracy of travel information collection; $R(u)$ represents the travel information recommended by the user; U represents all user data sets; $T(u)$ represents the set of travel information that users like in the database.

$$Coverage = \frac{|U_{u \in U} R(u)|}{|I_i|} \tag{14}$$

Where $U_{u \in U}$ is the set of all user information in the database; *Coverage* is the coverage rate of information search; $R(u)$ represents the travel information recommended by the user; I_i is a collection of all travel information. Coverage is usually used to measure whether a recommender system has a strong ability to discover some unpopular items. It is generally calculated by the proportion of all recommended items and the probability distribution of all recommended items. The larger the proportion is, the more average the probability dispersion is. According to the above evaluation index, we can get the processing time of the algorithm, and then judge its performance. The algorithm is compared with the three conventional algorithms. The algorithm in this paper is taken as the experimental group, and the three conventional algorithms are taken as the control group. The optimization degree of the algorithm is judged as shown in Fig. 2.

By comparing the algorithms in Fig. 2, we can get the data results of running time. The initial running time of the experimental group is 8 s, which can reach 22 s when the number of users is 6 W. The running time of control group 1 was 4–58 s, that of control group 2 was 10–36 s, and that of control group 3 was 15–49 s. Thus, although

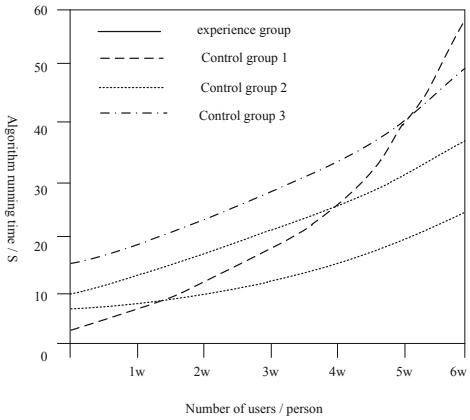


Fig. 2. Algorithm execution time test

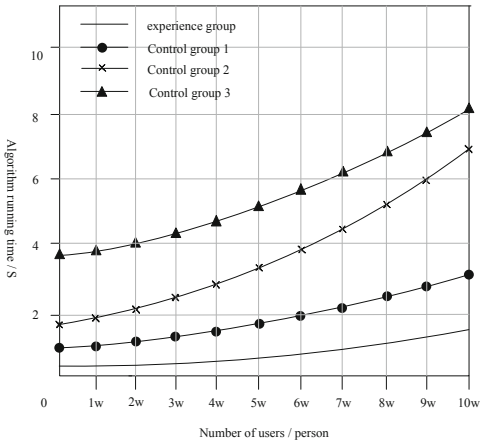
the initial running time of the experimental group is less than that of the control group 1, the overall running time of the experimental group is less than that of the three control groups. With the increase of the number of users, the running speed of the experimental group will show a trend of rapid increase. Therefore, the improved clustering algorithm is better than the three conventional algorithms, and the running time is less than the conventional algorithm.

3.3 Actual Retrieval Effect Test

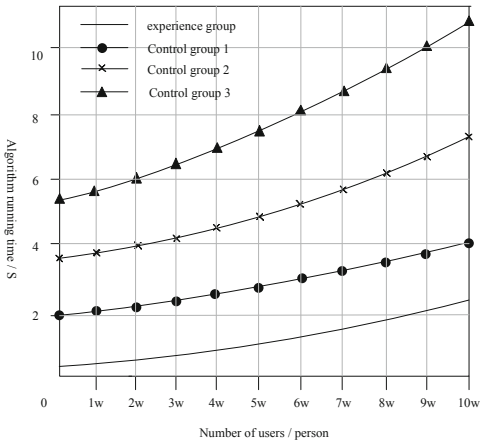
In the process of personalized tourism big data information search, the method designed in this paper makes use of the computing resources of big data platform, distributes the image feature extraction and visual vocabulary tree training process of consuming resources on the cloud computing platform, and provides users with search cloud service, which speeds up the search speed and improves users' experience of searching tourism information, To provide convenient tourism services for users. In order to test the performance of the tourism information search service designed in this paper and the three conventional methods, four data sets are designed to calculate the search time under different concurrent users and calculate the search speed. The test results are shown in Fig. 3.

As shown in Fig. 3, the retrieval time of the four data sets can be summarized as the data information shown in Table 3.

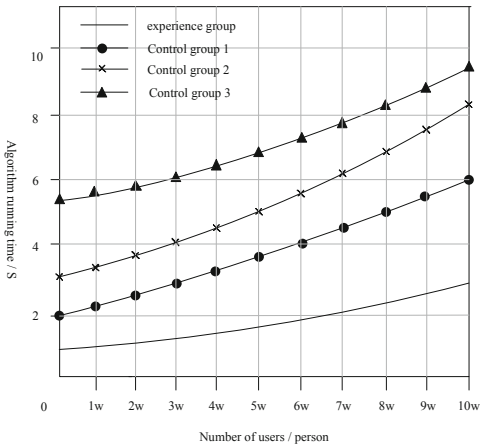
As shown in Table 3, in the four data sets, the retrieval time of the experimental group was less than that of the three control groups. It can be seen that the personalized tourism big data information fast recommendation method based on improved clustering algorithm designed in this paper has better retrieval efficiency and can quickly get the recommendation results.



(a) data set 1

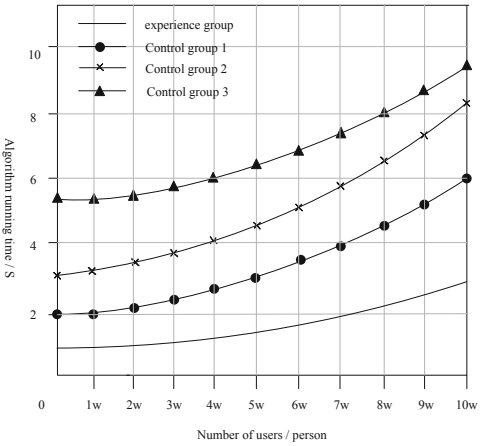


(b) data set 2



(c) data set 3

Fig. 3. Actual retrieval effect test



(d) data set 4

Fig. 3. continued

Table 3. Experimental results

Data set		Experience group	Control group 1	Control group 2	Control group 3
1	Maximum	0.36 s	1.98 s	1.86 s	3.76 s
	Minimum value	1.75 s	3.26 s	7.02 s	8.14 s
2	Maximum	0.36 s	2.01 s	3.67 s	5.35 s
	Minimum value	2.67 s	4.02 s	7.42 s	10.92 s
3	Maximum	1.09 s	2.00 s	3.14 s	5.47 s
	Minimum value	3.07 s	4.00 s	8.62 s	9.26 s
4	Maximum	0.84 s	2.02 s	3.26 s	5.65 s
	Minimum value	3.04 s	4.07 s	8.37 s	9.29 s

4 Conclusion

The research work of this paper mainly focuses on the problem of fast search and active push for big data personalized tourism. By fusing the travel information of tourists and the recorded scenic spot information, the travel information is matched with the scenic spot information on the semantic level to realize the rapid recommendation of personalized tourism big data information. However, this method has some shortcomings in tourism big data matching, such as the matching quantity and characteristics of tourism big data, which still need further research and improvement.

Fund Projects. National Planning Office of Philosophy and Social Science Foundation of China: Research on Cultural Heritage Conservation and the Activation of South China Historical Trail (NO: 19FSHB007).

References

1. Ma, R.X., Guo, F.Q., Liu, Z.J., et al.: Collaborative filtering recommendation algorithm for fusion context information and kernel density estimation. Comput. Technol. Dev. **31**(4), 34–39 (2021)
2. Tong, L.Y., Zhang, B.: Personalized recommendation algorithm based on global collaboration information. Value Eng. **40**(11), 233–234 (2021)
3. Wang, Y., Liu, L.: Construction of a classification model of network software test data based on feature expansion. Electron. Design Eng. **29**(8), 29–32, 37 (2021)
4. Liu, Y., Wei, M.: Personalized adaptive network hybrid information recommendation. Comput. Simulat. **38**(4), 399–402, 416 (2021)
5. Liu, S., Liu, D., Muhammad, K., Ding, W.: Effective template update mechanism in visual tracking with background clutter. Neurocomputing **458**, 615–625 (2020). https://doi.org/10.1016/j.neucom.2019.12.143
6. Liu, S., Liu, X., Wang, S., Muhammad, K.: Fuzzy-aided solution for out-of-view challenge in visual tracking under IoT assisted complex environment. Neural Comput. Appl. **33**(4), 1055–1065 (2021)
7. Wang, Z., Wang, S., Du, H.: Improved fuzzy C-means clustering algorithm based on density-sensitive distance. Comput. Eng. **47**(5), 88–96, 103 (2021)
8. Gao, P., Li, J., Liu, S.: An introduction to key technology in artificial intelligence and big data driven e-Learning and e-Education. Mob. Netw. Appl. **26**(5), 2123–2126 (2021). https://doi.org/10.1007/s11036-021-01777-7
9. Sun, Q., Chen, H., Li, C.: Clustering algorithm of big data based on improved artificial bee colony algorithm and MapReduce. Appl. Res. Comput. **37**(6), 1707–1710, 1764 (2020)
10. Feng, J., Yao, Y.: An optimization clustering algorithm based on multi-population genetic simulated annealing algorithm. Comput. Simulat. **37**(9), 226–230 (2020)