International Journal of Information Technology and Applications
Volume 2, Issue 2, June 2025.

ISSN 3007-7397

A Method for Mining Popular Search Behaviors in Tourism Networks based on Immune Clonal Algorithm

Fan Feng*

The 15th Research Institute, China Electronics Technology Group Corporation, Beijing, China
fengfan_0113@163.com

Abstract. In view of the high dimensionality, dynamics and noise interference of tourism network search data, this paper proposes a method for mining tourism popular search behaviors based on immune clonal algorithm. By simulating the clonal selection mechanism of biological immune system, the search behavior pattern is abstracted as an antigen-antibody matching problem, and a multi-dimensional affinity function, dynamic parameter adjustment strategy and hybrid model enhancement mechanism are designed. By optimizing the number of algorithm antibodies and classifiers, the model can be effectively used for mining tourism data. Experimental data show that the rule extraction rate of the model reaches 99.6%, far exceeding the effect of the comparison algorithm. It also performs well in terms of computing time, which is 34.25s, lower than other comparison models. It is further proved that the algorithm is particularly suitable for deep mining of association rules in low support scenarios.

Keywords: Immune clonal algorithm, Tourism search behavior, Pattern mining, Dynamic optimization, Multimodal data.

1. Introduction

According to statistics, the global online travel market is expected to reach 1.2 trillion US dollars in 2025 (Statista, 2023), and the popular search data on the tourism network contains various user needs and market trends [1]. Tourists around the world record travel data in various ways and use online searches to provide themselves with stable and reliable travel references [2]. Due to user personal preferences and other reasons, online search data has various irregular characteristics, so it is necessary to use data mining, classification and other methods to summarize the key information [3]. For unfamiliar travel destinations, it is difficult to accurately express personal wishes through vague search keywords, and more efficient algorithms are needed to help users obtain more accurate travel information.

In terms of tourism data mining, Zheng et al. [4] analyzed many geo-tagged photos, summarized users' personal preferences, summarized representative photos from the data of travel photos, and recommended scenic spots to users, achieving good results. Hao Qiang et al. [5] mined tourists' textual knowledge conducted topic mining on users' travel notes, and recommended information to users by calculating similar questions of different scenic spots. In terms of scenic spot information push, Jiang [6] mined user trajectories to form a variable memory Markov model and achieved rapid push of tourist

attractions. However, the algorithm ignored the influence of unpredictable contextual information in the calculation, and the push effect needs to be further improved. In terms of recommendation algorithm research, Branko Mihaljevic [7] studied the recommendation system of large websites, combined the immune clone algorithm with the danger theory to form a recommendation algorithm for the website, and then recommended personal preferences to users on personalized pages through learning the user's personal characteristics.

For the mining of various popular tourism search data on the Internet, traditional cluster analysis methods have problems such as insufficient adaptability to domains, high sensitivity to noisy data and data quality, and weak adaptability to dynamically changing data. The contribution of this research is: (1) Propose a data analysis and processing method in the tourism field based on the immune cloning algorithm; (2) Adopt algorithmic optimization design to mine potential tourism knowledge; (3) The effectiveness of the algorithm was verified through experiments, providing new methodological ideas for the analysis and mining of domain data.

2. Data Mining Based on Immune Clonal Algorithm

2.1. Overall Architecture

The Immune Clonal Algorithm (ICA) is based on the clonal selection theory and immune memory mechanism of the biological immune system. It solves optimization problems by simulating the interaction between antibodies and antigens. Its core theories include: (1) The principle of clonal selection. The clonal selection theory proposed by Burnet holds that the immune system screens antibodies through affinity. Antibodies with high affinity are cloned, amplified and mutated to form diverse antibody groups. In optimization problems, antigens correspond to the objective function, antibodies correspond to candidate solutions, and affinity represents the matching degree between the solution and the objective. High-affinity solutions are preferentially cloned and mutated, while low-affinity solutions are eliminated. (2) Immune memory and diversity maintenance. By retaining high-affinity antibodies (memory cells), the algorithm converges rapidly to the optimal solution. Through mutation operations and antibody concentration inhibition (to avoid excessive homologous antibodies), population diversity is maintained, and premature convergence is prevented. (3) Affinity maturation and variation strategies. The cloned antibody conducts local fine search through high-frequency mutation (the mutation probability is inversely proportional to the affinity) to improve the accuracy of the solution. Meanwhile, the variation step size and clone scale are dynamically adjusted with iterations to balance global exploration and local development.

Data mining using the immune clonal algorithm mainly includes two parts. The first is the model training part. This stage is mainly based on the clone proliferation, high-frequency mutation and immunosuppression of the immune clonal algorithm to achieve the update of the antibody group in the data mining process and finally achieve the feature selection of the model. The second is the result division stage. For the constructed model, the data samples are divided according to the results combined with under-sampling ensemble learning to finally get the output results. The data mining model process

based on the immune clonal algorithm is shown in Figure 1.

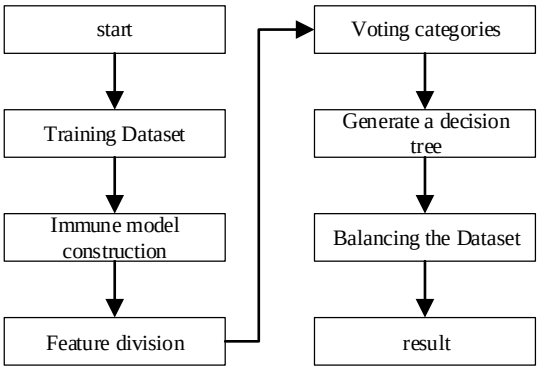

Figure 1. Research framework.

2.2. Algorithm Flow

In the immune clone feature selection algorithm, random mechanisms are used to generate antibodies, and the affinity between antibody groups is used to achieve antibody updates [8]. During the iteration process, new antibodies are generated through mutation operations, and the antibody group is optimized until all conditions are met [9].

The immune algorithm workflow is as follows:

2.2.1. Producing initial antibodies

The antibody population is generated using a random method, and the generation method is expressed as the Eq. (1):

$$G = round(rand(g, s)) \tag{1}$$

In the formula, G is the total antibody population; s is the number of features; g is the number of antibodies; rand (g, s) is a function; round is a precision counting method with a value of 0 or 1.

2.2.2. Calculating affinity

The affinity of population antibodies is used to measure the classification effect of antibody population characteristics. Each antibody is decomposed into two parts, one into the memory set and the other into the candidate set. Generally, 10% to 20% of the individuals in the population are selected to enter the memory set, while the candidate set stores the other antibodies outside the set. The affinity between antibodies represents the similarity between two antibody individuals [10]. Its calculation methods mainly include: the calculation method based on antibody-antigen affinity, the calculation method based on information-based antibody-antibody affinity, and the calculation method based on Hamming distance. This experiment uses binary representation, so the Hamming distance is used, expressed as the Eq. (2):

$$aff(ab_i, ab_j) = \sum_k^{L-1} \partial_k \tag{2}$$

$$\partial_k = \begin{cases} 1, ab_{ik} = ab_{jk} \\ 0, ab_{ik} \neq ab_{jk} \end{cases} \tag{3}$$

In Eq. (3), ab_{ik} and ab_{jk} are the k-th position of antibody i and antibody j respectively, and L is the antibody coding length.

2.2.3. Inhibition Steps

In the process of antibody group iteration, the number of antibodies with high affinity will effectively increase with calculation. To maintain the diversity of the antibody group, it is necessary to suppress it. In the artificial immune system, antibody concentration is used to characterize the diversity of antibody populations, and a high antibody concentration means that there are many very similar individuals in the population. Therefore, individuals with too high concentrations need to be suppressed during optimization to ensure individual diversity. Antibody concentration is defined as the Eq. (4):

$$den(ab_i) = \frac{1}{n} \sum_{j-0}^{L-1} aff(ab_i, ab_j) \tag{4}$$

where n is the population size, ab_i is the i-th antibody in the population, and $aff(ab_i, ab_j)$ is the affinity between antibody i and antibody j.

2.2.4. Mutation produces new antibodies

Through mutation and crossover, antibodies entering the next generation are generated. During the calculation process, mutation and crossover are repeated until the convergence judgment data is satisfied. The antibody recombination and mutation process simulate the biological cloning selection process and performs mutation and recombination operations on the cloned antibodies. The mutation is performed randomly according to a certain probability, and the affinity of the antibody is recalculated. If the affinity of the antibody in the set after the operation exceeds the affinity of the original antibody, the new antibody replaces the original antibody.

Properly increasing or decreasing a small amount of new and old features may result in better classification performance compared to the original feature subset. Special attention should be paid to avoiding unnecessary and repeated antibody types, so as to maintain the diversity of antibodies. Therefore, new antibodies must be selected from the candidate antibody group each time a variant antibody operation is performed. The rule of clonal expansion is an increasing function of antigen affinity measurement. The greater the affinity, the larger the cloning scale. Generally, the cloning scale is carried out as the Eq. (5):

$$Q_i = int[N_c \cdot aff(a_i) / \Sigma_{j-1}^{n} aff(a_j)] \tag{5}$$

N_c is the total clone size of the antibody population, and $int()$ is a round-up function.

In the cloning algorithm, cloning mutation refers to the mutation operation on the antibody cloning results obtained by cloning amplification to produce and degree mutations and realize local research. It

is an important operator in cloning algorithms to generate potential new antibodies and realize regional search and has a great impact on the performance of the algorithm.

Generally, the probability of clonal mutation of each antibody is calculated as the Eq. (6):

$$T_m(ab_{i,j,m}) = \begin{cases} !\,ab_{i,j,m} & rand() < p_m \\ ab_{i,j,m} & else \end{cases} \tag{6}$$

where $ab_{i.j.m}$ is the j -th dimension of the m -th clone of the antibody: $!\,ab_{i,j,m}$ is the result of negating $ab_{i,j,m}$, $rand$ is a function that generates random numbers in the range of (0.1), and a random number generator based on uniform distribution is usually used: p_m is the mutation probability. Since the mutation probability p_m is generally small, the smaller the distance $d(a,b)$, the larger $P\{a \to b\}$ is, so that the mutation causes the antibody to change within its field, expanding the search range.

2.2.5. Clone selection

Antibodies that can be retained are selected based on affinity calculations: high-affinity antibodies are promoted, and high-density antibodies are suppressed, thus reflecting the diversity of immune control.

The antibody with the highest affinity is cloned, and the number of clones is proportional to its affinity. The formula for calculating the number of clones as the Eq. (7):

$$N = \sum round\left(\frac{al}{i}\right) \tag{7}$$

where N is the total number of antibodies produced by the clone; a is the coefficient; l is the number of antibodies used for cloning; and i is the serial number of the preceding antibody.

Ensemble learning is a classification algorithm, or more precisely, a system of classification model collections. Its basic idea is to combine multiple weak learners into a strong learner. The steps of the ensemble algorithm are shown in Figure 2.

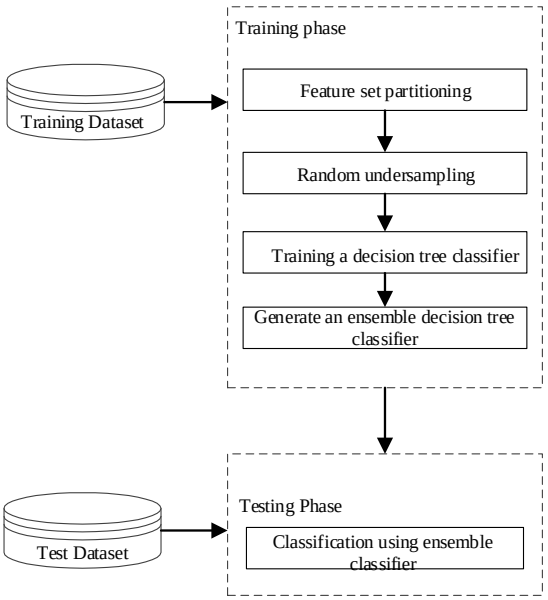

Figure 2. Block diagram of the integrated algorithm steps.

3. Experimental results and data analysis

3.1. Experimental Setup

The hardware environment of the experiment is: CPU is Intel® Core™ i7-9705H 2.60GHz, memory capacity is 16G, and graphics card configuration is NVDIA GeForce gtx 1650 4G.

The experimental platform of this study uses Windows 10 64-bit operating system, and the algorithm is implemented in Matlab R2020b integrated development environment. The maximum number of iterations of the algorithm is set to 100 generations, and the experimental sample data comes from the image social platform Panoramio. To improve the validity of model evaluation, the repeated sampling strategy is used to divide the data set into multiple rounds during the training process, and the comprehensive evaluation index is obtained through cross-validation of multiple independent training-test sets.

The k-fold cross-validation technique was used, with the experimental parameter k=5. This method equally divides the original data set into five mutually exclusive subsets and uses each subset as a validation set in turn through five iterations, while the remaining four groups constitute the training set. Finally, the evaluation results obtained through five independent experiments were statistically analyzed to effectively improve the reliability of the model performance evaluation.

3.2. Evaluation Indicators

The evaluation indicators and formulas used in this study are as the Equations:

$$Precision = TP / (TP + FP) \tag{8}$$

$$Recall = TP/(TP + FN) \tag{9}$$

$$Specificity = TN/(TN + FP) \tag{10}$$

$$Accuracy = TP + TN/(TP + FP + FN + TN) \tag{11}$$

$$F1score = 2 * TP/(2 * TP + FP + FN) \tag{12}$$

In the Equations, TP stands for True Positive; FP stands for False Positive; FN stands for False Negative; and TN stands for True Negative.

3.3. Experimental Results

3.3.1. Effect of Antibody Quantity

For the immune clonal algorithm, the most important parameter is the number of optimal feature subsets, which is also the number of optimal antibodies selected for cloning. This parameter is set to an odd sequence of 3, 5, 7, 9, 11, and 13. Each odd parameter is tested 50 times, and the values of the evaluation indicators are shown in Table 1.

Table 1. Effect of different antibody quantities.

Number of antibodies	Precision	F1-score	AUC Value
3	0.8653	0.9087	0.9354
5	0.8798	0.9125	0.9369
7	0.8815	0.9207	0.9487
9	0.8809	0.9215	0.9527
11	0.8860	0.9275	0.9589
13	0.8814	0.9234	0.9529

From the results in Table 1, we can see that when the evaluation indicators are accurate to 4 decimal places, all evaluation indicators first increase and then decrease with the number of antibodies. When the optimal number of antibodies is 11, the accuracy, F1-score and AUC value are in a relatively optimal state. Therefore, this study sets the optimal number of antibodies to 11.

3.3.2. Impact of Classifiers

After the sub-classifier selection algorithm, this experiment finally chose to use common classifiers such as SVM, logistic regression, C4.5 Decision Tree, Bayes, etc. to classify the data set. The classification results are shown in Table 2.

Table 2. The impact of different sub-classifiers.

Subclassifier	Precision	F1-score	AUC Value
SVM	0.7128	0.7286	0.7683
Logistic regression	0.6265	0.7051	0. 6108
C4.5 Decision Tree	0.7212	0.7258	0.7941
Bayes	0.6219	0.7674	0.6413

In the comparison of classifier performance, C4.5 and SVM are more outstanding in comprehensive performance, while Bayes and Logistic regression are better than C4.5 and KNN classifiers in F1-score, but there is a significant gap in their AUC scores. In view of the imbalanced distribution of categories in the test set of the experimental data set, combined with the analysis of F1-score and AUC values, it can be seen that the recognition accuracy of the current classifier for minority class samples still needs to be improved. Therefore, in the imbalanced data scenario, it is recommended to adopt an evaluation strategy that combines the dual indicators of F1-score (comprehensive accuracy) and AUC (model discrimination), which also constitutes the theoretical basis for selecting these two evaluation dimensions in this study. Finally, through multi-dimensional indicator analysis, it is shown that the C4.5 decision tree has a relative advantage in the comprehensive performance of the model.

3.3.3. Comparison of network search behavior mining effects of different algorithms

To verify the effectiveness, scalability and advancement of the above algorithms, the Apriori algorithm, the reference 10 algorithm and the mining algorithm based on clonal selection were experimented respectively. The support degree was 0.5% and the confidence degree was 60% to mine the data records. The extraction rate comparison results of the mined association rules are shown in Table 3.

Table 3. Comparison of association rule extraction effects of different algorithms.

Algorithm Used	Extraction rate of rules %
Apriori algorithm	92.8
Algorithm in Reference [10]	91.4
Immune clonal algorithm	99.6

Under the same conditions, the calculation time of different algorithms is compared, and the comparison results are shown in Table 4.

Table 4. Data mining time comparison.

Algorithm Used	Calculation time (s)
Apriori algorithm	73.55
Algorithm in Reference [10]	47.68
Immune clonal algorithm	34.25

The experimental data analysis of Table 3 and Table 4 shows that although the traditional Apriori algorithm can effectively identify basic association rules, its high computational complexity leads to a large time overhead. In contrast, the immune clonal association mining model shows significant performance advantages in terms of rule extraction rate and execution efficiency, especially when processing large-scale data. Experimental data proves that the rule extraction rate of this model reaches 99.6%, far exceeding the effect of the comparison algorithm. It also performs well in terms of calculation time, which is 34.25s, lower than other comparison models. It is further proved that this algorithm is particularly suitable for deep mining of association rules in low support scenarios.

From the perspective of algorithmic mechanism, the computational complexity of the immune clonal model is mainly constrained by three factors: the population size parameter directly affects the number of iterations of the antibody concentration calculation and affinity mutation operation, while the support threshold adjustment indirectly regulates the affinity calculation by changing the antibody length. It is worth noting that when the population size is kept constant: as the support threshold decreases, the running time of the model shows a decreasing trend - this is closely related to its unique architecture without database scanning; while the Apriori algorithm has a linear growth trend in running time due to the surge in candidate item sets, which leads to an increase in the frequency of database scanning.

In-depth research on the algorithm characteristics revealed that the immune clonal model has dual technical advantages: first, by integrating affinity gradient search and population concentration control

mechanism, a dynamic balance between global optimal solution exploration and local fine search is achieved; second, an adaptive parameter coupling architecture is used to optimize support and confidence parameters.

4. Conclusion

To mine the internal characteristics of popular search behaviors in tourism networks, this paper studies a data mining model based on the immune clone algorithm. The immune clone model can effectively achieve a dynamic balance between global optimal solution exploration and local fine search and use an adaptive parameter coupling architecture to collaboratively optimize the support and confidence parameters. By optimizing the number of algorithm antibodies and classifiers, the model can be effectively used for tourism data mining. Compared with other classic algorithms, experimental data shows that the rule extraction rate of the model reaches 99.6%, far exceeding the effect of the comparison algorithm. It also performs well in terms of computing time, with a computing time of 34.25s, which is lower than other comparison models. It is further proved that the algorithm is particularly suitable for deep mining of association rules in low support scenarios.

References

[1] Ruixiang L. Development of a travel recommendation algorithm based on multi-modal and multi-vector data mining. [J]. PeerJ Computer science, 2023, 9:e1436e1436.

[2] Lampropoulos V, Panagiotopoulou M, Stratigea A. Unfolding the Potential of the Culture-Tourism Nexus in Greece: A DataEnabled Methodological Approach[J]. Journal of Tourism & Hospitality,2021,10(6):1-3.

[3] Chen X. Data mining-based ecotourism visitor evaluation management model in the context of sustainable development[J]. International Journal of Environment and Sustainable Development, 2024, 23(2-3):142-157.

[4] Zheng V W, Zheng Y, Xie X et al. Collaborative Location and Activity Recommendations with GPS History Data[C]. Proceedings of the 19th International Conference on World Wide Web. ACM, 2010:1029-1038.

[5] Qiang Hao, Rui Cai, Changhu Wang et al. Equip Tourists with Knowledge Mined from Travelogues[C]. In Proc. WWW, New York, NY, USA. ACM Press, 2010:401-410.

[6] Jiang Kai, Yu Nenghai, and Weihai Li .Online Travel Destination Recommendation with Efficient Variable Memory Markov Model[J]. Multimedia and Expo Workshops (ICMEW), 2013 IEEE International Conference on. IEEE, 2013:1-4.

[7] B. Mihaljevic, A. Cvitas and M. Zagar, recommender system model based on artificial immune system[J]. Information Technology Interfaces, 2006,367-372.

[8] Jun L, Luyu Y, Haiyue Z et al. Impact of climate change on hiking: quantitative evidence through big data mining[J]. Current Issues in Tourism, 2021, 24(21):3040-3056.

[9] Hou M. Research on Upgrading of Cultural Tourism Management Information System Based on Data Mining Technology[J]. Tourism Management and Technology Economy, 2022, 5(1): 242-251.

[10] Yanhong Z. Retraction Note: Coastal tourism resource development based on big data mining and environmental sustainability[J]. Arabian Journal of Geosciences,2021,14(23): 579-588.