INTERNATIONAL JOURNAL OF CIRCUITS, SYSTEMS AND SIGNAL PROCESSING
DOI: 10.46300/9106.2021.15.72 Volume 15, 2021

User Behavior Identification and Personalized Recommendation Based on Web Data Mining

Ya Wang
School of Information Engineering, Xuchang University, Xuchang 461000, China
*Email: wangyiaya@yeah.net

Received: January 22, 2021. Revised: July 5, 2021. Accepted: July 16, 2021. Published: July 20, 2021.

Abstract—A good understanding of user behavior and consumption preferences can provide support for website operators to improve their service quality. However, the existing personalized recommendation systems generally have problems such as low Web data mining efficiency, low degree of automated recommendation, and low durability. Targeting at these unsolved issues, this paper mainly carries out the following works: Firstly, the authors established a user behavior identification and personalized recommendation model based on Web data mining, it gave the user behavior analysis process based on Web data mining, improved the traditional k-means algorithm, and gave the detailed execution steps of the improved algorithm; moreover, it also elaborated on the K nearest neighbor model based on user scoring information, the score matrix decomposition method, and the personalized recommendation method for network users. At last, experimental results verified the effectiveness of the constructed model.

Keywords—Web data mining; User behavior identification; Personalized recommendation; K-means algorithm; User preference nearest neighbor model

I. Introduction

Web user behavior data generally contain multi-dimensional attribute information such as space, time, and semantics, as an important part of social network information in the era of big data, it has the characteristics of fast update speed, large capacity, rich content, wide coverage, and closely related to user activities, and it is of great mining potential and value [1-4]. Based on the behavior data of a user in a fixed time period or time point, the user's behavior preferences or behavior patterns can be mined, which is important for the reproduction of the user's life trajectory [5-7]. From the perspective of providing Web-based network services, a better understanding of user behavior and consumer preferences could provide certain support to operators so that they can improve their service quality [8-10].

To overcome shortcomings in existing user behavior analysis strategies such as the low efficiency and poor reliability, field scholars have conducted various studies, for example, Li and Xu [11] introduced particle swarm optimization (PSO) and genetic algorithm (GA) into a BP neural network model of user behavior analysis, which has effectively improved the stability and accuracy of the algorithms, and provided more accurate recommendation evidences for online product sales or services. The data mining-based user operation and consumption behavior analysis method has directly determined whether the website operators' understanding of user requirements is fast and accurate enough [12-15]. Dhanalakshmi [16] used random forest and Xgboost algorithm to extract and classify the features of user consumption behavior and make predictions. Offutt and Thummala [17] conducted time and space mining analysis on the geographic data of social media obtained from the API data access interface, and analyzed the annual, holiday and workday features of user sign-in data; they employed the ArcGIS-based spatial analysis method to perform kernel density analysis and obtained results of the regional visitors' travel patterns. The rating and comment data of users constitute the users' feedback information of their preferred items or services [18-22]. Some research results merely mitigate the limitation of the recommendation solely based on user feedbacks. Walek [23] comprehensively considered the user feedback information and comment topics, and constructed a personalized preference prediction model that can measure the degree of matching between user preferences and product attributes. Some research results solve the small size and sparsity of user behavior data. For example, Bernaschina et al. [24] combined the factorization machine (FM) and the hidden Markov model (HMM) to effectively reduce the interference and improve the accuracy of the personalized recommendation algorithm for the situation. Van Woensel et al. [25] took the personalized electronic teaching resource recommendation method as the research object, and built a personalized exercise recommendation model based on students' comprehensive professional knowledge level and cognition level of specific knowledge points. Based on SPSS Modeler and K-means algorithm, Bernaschina et al. [26] realized the data collection, data mining, similar user classification and feature extraction of a personalized recommendation system of university library, and designed the B/S structure of the system. Alshreef Abed et al. and Weipeng Wang et al. [27-28] carried out text extraction using an SVM-based distinct stages framework for data mining,

and developed a novel clustering scheme based on improved k-means algorithm.

A personalized recommendation system centered on user preferences should be able to reflect the changes in user preferences. The idea of current solutions is to optimize the personalized recommendation results based on the score weighting mechanism and similarity calculation, and these methods generally have the shortcomings of low data mining efficiency, low degree of automation of recommendation, and low durability; therefore, in order to better solve problems with current recommendation systems in user cold start, data sparseness and recommendation drift, this paper attempts to construct a user behavior identification and personalized recommendation model based on Web data mining. The main contents include the following aspects: The second chapter of this paper gives the user behavior analysis process and the corresponding analysis model, improves the traditional k-means algorithm, and gives the detailed execution steps of the improved algorithm. The third chapter presents the personalized recommendation method of network users, expounds the idea of personalized recommendation algorithm, and constructs the neighborhood model of user preference. At last, experimental results are employed to verify the effectiveness of the constructed model. Overall, this paper summarizes the defects of the previous research, pointing out the connotations and advantages of online user behavior analysis. By analyzing the superiority of web data mining, the authors constructed a web behavior analysis model based on web data mining. Combined with the massive sales and transaction data gathered by actual enterprises, our method can promote the construction of the relevant web platforms, speed up production and sales, and enhance the competitiveness of logistics enterprises.

II. Network User Behavior Classification and Identification Method

Figure 1 shows the process of user behavior analysis based on Web data mining. According to the figure, it mainly includes: the potential user discovery and value prediction before entering the new market; the use cluster analysis, precision marketing direction classification, user behavior analysis and prediction, and user relationship network construction in the user stabilization stage; and the user loss time identification and user loss cause analysis in the user exit stage. Figure 2 shows the structure of the user behavior analysis model based on Web data mining, which contains the complete analysis process including business understanding, data collection, data preprocessing, data mining, and knowledge representation.

This paper used an improved web data mining k-means algorithm to analyze the differences in user behavior. Based on the distance function value that characterizes the differences in user behavior, the k-means algorithm for network user behavior identification and personalized recommendation will allocate user behavior data to *K* clusters that characterize the different user behaviors.

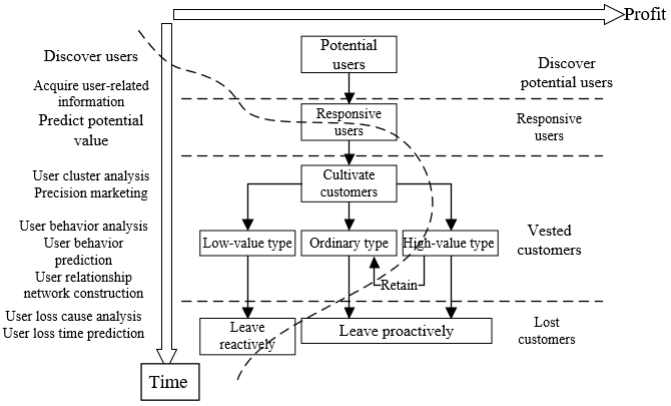


Fig. 1 Process of user behavior analysis based on Web data mining

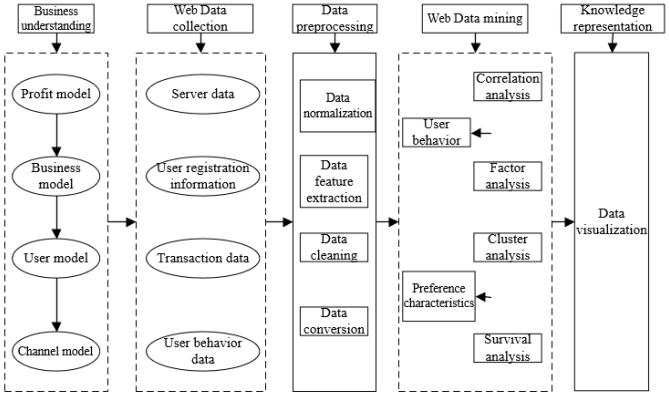


Fig. 2 Structure of the user behavior analysis model based on Web data mining

Suppose $C=\{C_1, C_2,\ldots,C_m\}$ represents the behavior data set of m users to be clustered, $C_a=\{c_{a1}, c_{a2}, \ldots, c_{as}\}$ represents the behavior data of the a-th user, the behavior data of this user is a vector in the real number space $C\in\mathbb{R}^s$, wherein s represents the data space dimension, namely the number of attributes of user behavior data; then, the Euclidean distance between the behavior data of any two users can be obtained from Formula 1:

$$D(c_a,c_b)=\sqrt{(c_{a1}-c_{b1})^2+(c_{a2}-c_{b2})^2+\ldots(c_{as}-c_{bs})^2} \quad (1)$$

Among the K clusters describing different user behaviors, suppose CEN_k represents the cluster center of the k-th behavior cluster, N_k represents the k-th behavior cluster set, $|N_k|$ represents the number of user objects to which the behaviors belong in N_k, then, the cluster center of each cluster can be calculated by Formula 2:

$$CEN_k=\frac{\sum_{c\in N_k} C_a}{|N_k|} \quad (2)$$

The distance between user behavior data point and the cluster center after clustering can be calculated by Formula 3:

$$D(c,CEN_k)=\sqrt{(c_{a1},n_{k1})^2+(c_{a2},n_{k2})^2+\ldots+(c_{as},CEN_{ks})^2} \quad (3)$$

The sum-of-square error (SSE) of the distance between the user behavior data point and the cluster center can be calculated by Formula 4:

$$SSE = \sum_{k=1}^{K} \sum_{c \in N_k} D\left(c, CEN_k\right)^2 \tag{4}$$

Before using the k-means algorithm to perform user behavior classification and identification, the Web data mining system randomly selects K user behavior data as the initial clustering centers, and calculates the distance between each user behavior data and the random initial clustering center based on Formula 3, and then selects the closest distance and completes the classification of user behavior data. Then, after the behavior clustering is completed, the cluster centers of the user behavior data set are recalculated according to Formula 2. The above steps are repeated until the cluster centers and user behavior classification results no longer update, and the SSE calculated by Formula 4 approaches zero. Taking a data set of 12 user behavior data points shown in Figure 3 as an example, 5 and 11 are selected as the initial cluster centers, and Table 1 gives the distances from the user behavior data points to the initial cluster centers.

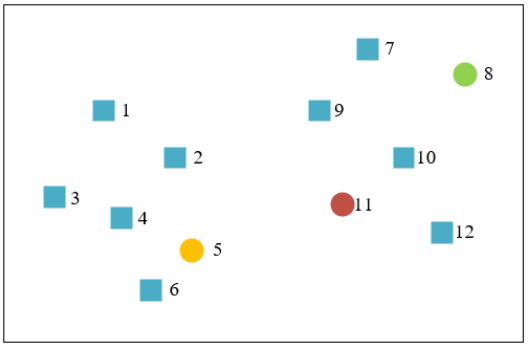


Fig. 3 An example of user behavior data set

Table 1. Distances from user behavior data points to initial cluster centers

Data point	Cluster center 5	Cluster center 11	Cluster result
1	0.72	0.65	5
2	0.32	0.37	5
3	0.51	0.76	5
4	0.34	0.72	5
5	0.00	0.00	5
6	0.42	0.53	5
7	1.52	1.37	11
8	1.73	2.51	11
9	1.89	2.03	11
10	1.35	2.32	11
11	0.00	0.00	11
12	1.45	1.58	11

The traditional k-means algorithm has been improved in this paper. In the algorithm, the average distance between the a-th and b-th user behavior data is given by Formula 5:

$$D_{AV} = \frac{1}{m} D(c_a, c_b) \tag{5}$$

The standard deviation of the two user behavior data can be calculated by Formula 6:

$$D_{SD} = \sqrt{\frac{1}{m} \sum \left[D_{AV} - D(c_a, c_b)\right]^2} \tag{6}$$

The density function value of the a-th user behavior data C_a can be calculated by Formula 7:

$$DF(C_a) = \sum_{j=1}^{m} \xi\left(D_{SD} - D(c_a, c_b)\right) \tag{7}$$

where, $\xi(*)$ is a binary function that characterizes whether the Euclidean distance between the behavior data is smaller than the standard deviation, if it is smaller, its value taken as 1, otherwise it takes 0.

$$\xi(x) = \begin{cases} 0 & x < 0 \\ 1 & x \geq 0 \end{cases} \tag{8}$$

The average density of the a-th user behavior data C_a can be calculated by Formula 9:

$$DF_{AV}(C) = \frac{1}{m} \sum_{i=1}^{m} DF(C_a) \tag{9}$$

Formula 10 can calculate the standard deviation of the density of behavior dataset samples:

$$DF_{SD}(C) = \sqrt{\frac{1}{m} \sum_{a=1}^{m} \left[DF_{AV}(C) - Density(c_a)\right]^2} \tag{10}$$

It should be noted that if $DF(c_a)$ is less than $DF_{SD}(C)$, then the behavior data point cannot be classified into the preset category and is judged as an outlier point.

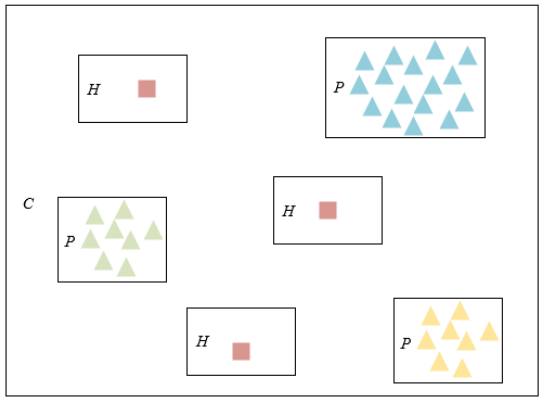


Fig. 4 A diagram of outlier point classification

The improved k-means algorithm takes the standard deviation of two user behavior data as the search radius, and completes accurate initial cluster center searching combining with the density function of user behavior data C_a. Based on the outlier point determination condition shown as Formula 10, the user behavior data set C to be clustered is divided into two parts: those meet the determination condition and those don't. The former ones are put into set H, and the latter ones are put into set P. Figure 4 gives a diagram of outlier point classification. Next, the density function value corresponding to the user behavior data that does not meet the judgment condition is calculated and compared with the density standard deviation, if it is greater than the density standard deviation, then the data is

put into set N, if it is smaller than the density standard deviation, then the data is put into set Q, and the rest are put into set F. Through the above operations, the search range and search time of the improved algorithm can be greatly reduced, and the probability of outlier point selection can be reduced as well. The execution steps of the improved algorithm can be summarized as follows:

Step1: Calculate the average distance and standard deviation of all user behavior data samples in the data set using Formulas 5 and 6;

Step2: Calculate the density function value of each data sample in the data set using Formula 7;

Step3: Calculate the average density and density standard deviation of user behavior data samples using Formulas 9 and 10;

Step4: Determine whether the data samples are outliers, put the points that meet the condition into set H, and put those that do not meet the condition into set P;

Step5: Put those user behavior data in set P that are greater than the density standard deviation into set N, select user behavior data points that correspond to the maximum value and determine them as the initial cluster centers;

Step6: In the circles with the standard deviation of the user behavior data sample as the radius and the initial cluster centers as the origins, perform 0 value assignment operation on the density function;

Step7: Return to step 4 until K ideal user behavior cluster centers are obtained, and output.

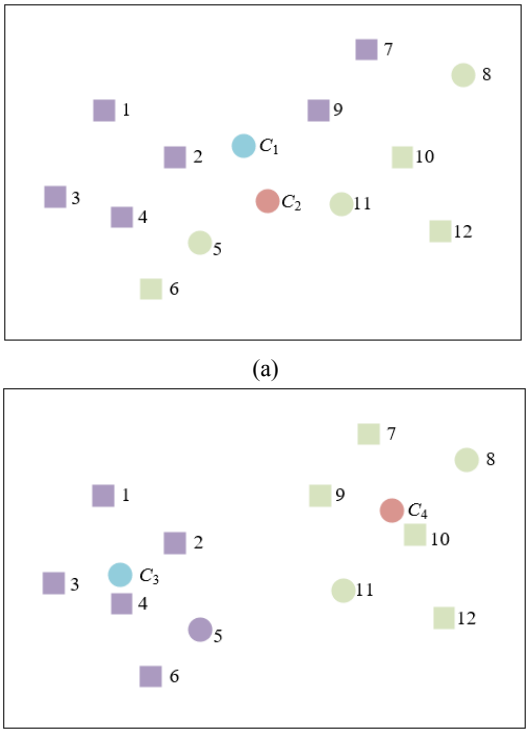


Fig. 5 User behavior clustering results

Table 2. Distances between behavior data points to cluster centers C_1 and C_2

Data point	Cluster center C_1	Cluster center C_2	Cluster result
1	1.35	1.72	C_1
2	0.82	1.09	C_1
3	1.37	1.52	C_1
4	0.72	0.99	C_1
5	0.75	0.63	C_1
6	1.23	0.75	C_2
7	1.39	1.38	C_2
8	1.35	1.59	C_2
9	1.26	1.83	C_2
10	1.71	1.75	C_2

Taking the data set of 12 user behavior data points shown in Figure 3 as an example, Figure 5 shows the results of user behavior clustering for 1 iteration and 2 iterations; Table 2 lists the distances between behavior data points to the cluster centers C_1 and C_2 after 1 iteration; Table 3 lists the distances between behavior data points to the cluster centers C_3 and C_4 after 2 iterations.

Table 3. Distances between behavior data points to cluster centers C_3 and C_4

Data point	Cluster center C_3	Cluster center C_4	Cluster result
1	0.35	2.13	C_3
2	0.48	1.85	C_3
3	0.36	2.56	C_3
4	0.31	1.79	C_3
5	1.95	1.68	C_3
6	2.39	1.61	C_4
7	1.72	0.59	C_4
8	2.12	0.37	C_4
9	2.36	0.25	C_4
10	2.73	0.68	C_4

III. Personalized Recommendation Methods for Network Users

In the field of personalized recommendation systems based on Web data mining, the collaborative filtering recommendation algorithms based on the calculation of user behavior preference similarity are quite popular. However, when the traditional collaborative filtering algorithms construct the nearest neighbor set of user behavior preferences, they generally have a high requirement for the accuracy of the constructed user behavior preference score matrix. Therefore, for cases in which the score matrix is sparse, this paper improved the traditional collaborative filtering algorithm to ensure the recommendation quality of the personalized recommendation system. A complete personalized recommendation system mainly consists of a user behavior recording module that collects the user behavior logs, a model analysis module that examines user preferences, and a personalized recommendation algorithm. Figure 6 explains the structure of personalized recommendation system. Specifically, the user behavior recording module mainly collects the various online behaviors of users. The model analysis module examines the historical user behaviors recorded by the previous module,

and uses the examination results to build up a user preference model that describes the user interests. The personalized recommendation algorithm is the key component of the personalized recommendation system. The algorithm generally predicts the current needs of users according to the historical ratings given by the users and their registered information. Once the algorithm predicts the user ratings on the commodities not yet purchased, the recommendation system will recommend the high-rating commodities to the users.

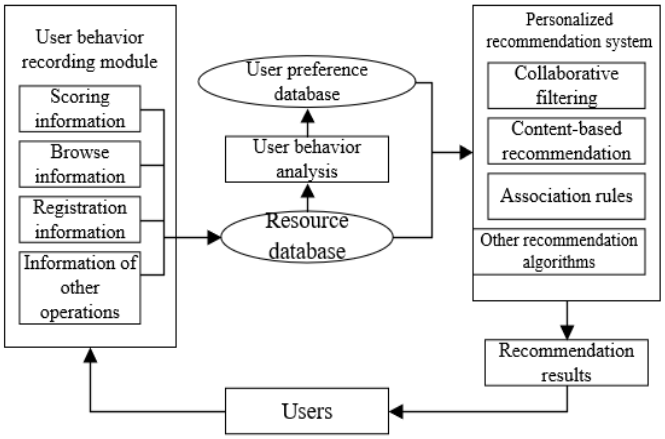


Fig. 6 Structure of personalized recommendation system

A. *Idea of personalized recommendation algorithm*

After experiencing goods, services, and other items, the users' perceptions or preferences can be described in scores. Suppose e_{vi} and e'_{vi} represent the existing score and the predicted score of user v for the i-th evaluation object, E presents the score matrix composed of e_{vi}, $V=\{v_1,v_2,\ldots,v_N\}$ and $I=\{i_1, i_2, \ldots, i_M\}$ are respectively the set of N users and the set of M evaluation objects. Table 4 shows the user-evaluation object scores.

Table 4. User-evaluation object scores

User \ Evaluation object	i_1	i_2	...	i_M
v_1	r_{11}	r_{12}	...	r_{1M}
v_2	r_{21}	r_{22}	...	r_{2M}
...	...	...		...
v_N	v_{N1}	v_{N2}	...	v_{NM}

In actual situations, the user's scoring habits or scoring factors for goods, services and other items are not directly related to the evaluation objects. Suppose λ represents the average value of the score records of all user behavior preferences in the training set of the personalized recommendation system, it describes the extent to which the user behavior preference score is affected by the website itself. The factors in user scoring habits that are not directly related to the evaluation objects are called user bias items and are represented by PZ_v; the factors in the evaluation object scores that are not directly related to users are called evaluation object bias items and are represented by PZ_i. In order to better characterize user scoring preferences, the set global bias item PZ_{vi} is expressed by Formula 11:

$$PZ_{vi} = \lambda + PZ_v + PZ_i \tag{12}$$

The matrix factorization method can extract a set of potential factors that describe users and evaluation objects from the scoring pattern of the personalized recommendation system. The singular value decomposition (SVD) that can achieve matrix dimensionality reduction can realize the transformation of the independent high-dimensional matrix reflecting the complex characteristics of the original matrix and the continuously multiplied simple low-dimensional matrices reflecting each single characteristic of the original matrix. Suppose there're $N\times N$ orthogonal matrix V, $M\times M$ orthogonal matrix U, and $N\times M$ diagonal matrix O; the elements on the diagonal of O, namely the singular values, are represented by $O = diag(\varphi_1, \varphi_2,\ldots,\varphi_M)$, and they satisfy two conditions at the same time: $\varphi_1 \geq \varphi_2 \geq \ldots \geq \varphi_M \geq 0$, and all elements that are not on the diagonal line are equal to 0, then, the decomposition method describes the $N\times M$ scoring matrix E in Formula 13:

$$E = V \times O \times U \tag{13}$$

According to above formula, SVD can construct a K-dimensional space based on the K largest diagonal elements of the left singular simplified matrix, thereby generating a new diagonal matrix. Through SVD operations, V becomes the $N\times K$-dimensional matrix V_k, O becomes the $K\times K$-dimensional matrix O_k, and U becomes the $K\times M$-dimensional matrix U_k, then the approximate matrix of score matrix E can be described by Formula 14:

$$E_k = V_k \times O_k \times U_k \tag{14}$$

Based on above method, the user score matrix is processed; suppose q_v and w_i represent the g-dimensional hidden feature vectors of user v and the i-th evaluation object, then the predicted score of user v for the i-th evaluation object can be expressed as:

$$e'_{vi} = q_v * w_i^T \tag{15}$$

B. *User preference nearest neighbor model construction*

Generally speaking, while saving the user scoring information, the Web data mining-based personalized recommendation system will save user registration information such as occupation, gender, and age of the user, and the description of the attributes of evaluation objects together. Some websites also allow users to register and log-in using their social network terminal accounts such as WeChat, Alipay, and Weibo. Based on the existing user registration information and evaluation object attribute descriptions, this paper improved the traditional collaborative filtering algorithm. Suppose $O(v, k)$ represents the behavior data set of K nearest neighbors of user v obtained in Chapter 2, PZ_{vi} and ω_{vu} represent the global bias item and the corresponding weight, then, the user's nearest neighbors can be selected through the existing user registration information, that is:

$$e'_{vi} = PZ_{ui} + \sum_{u\in O(v,k)} (e_{ui} - PZ_{vi}) * \omega_{vu} \tag{16}$$

The realization of Formula 16 is mainly based on finding the solutions of two problems: finding $O(v, k)$ and calculating ω_{vu}.

First, suppose that the registration information of each user contains N items, including basic personal information such as user name, gender, date of birth, location, and occupation, etc.; then, the user feature attributes of each item are divided into gradients and described in numbers, after data preprocessing, a N-dimensional registration information (o_1, o_2,…,o_N) of user v could be obtained. After that, for the behavior preferences of user v and other users u, calculate the Pearson Correlation Score, and sort out the similarity calculation results $S(v, u)$, and the set $O(v, k)$ constituted by K users with highest $S(v, u)$ value is defined as the K nearest neighbors of user v.

In this paper, the basis for the personalized recommendation of goods, services and other items that meet the cognition and preferences of target user is the prediction of the score of the targe item based on the scoring information of K nearest neighbor users, if all scores given by K nearest neighbor users to the target item are relatively high, then the item is recommended. When predicting the preferences of user v, it is necessary to make full use of the history score records of user v and other users u. Therefore, the following steps need to be added to the process of preference prediction: 1) If the website platform has stored the history scoring information of user v, then the weight parameter ω_{vu} of the personalized recommendation system is adjusted and updated based on the stochastic gradient descent method, at this time, the objective function is defined by Formula 17:

$$\mu\left(PZ_v^2+PZ_i^2+\sum_{u\in O(v,k)}\omega_{vu}^2\right)+\sum\left(\begin{matrix}e_{vi}-\lambda-PZ_v-PZ_i\\-\sum_{u\in O(v,k)}\left(e_{ui}-PZ_{ui}\right)*\omega_{vu}\end{matrix}\right)^2 \tag{17}$$

The first term in Formula 17 can effectively avoid overfitting of the weight parameter training of the personalized recommendation system. To minimize the above formula, it needs to take the partial derivatives of PZ_v, PZ_i, and ω_{vu} using the stochastic gradient descent method and adjust and optimize the parameters through multiple iterations. Suppose $Error_{vi} = e_{vi}-e'_{vi}$ represents the evaluation error of user v for the i-th evaluation object, β represents the learning speed, and μ represents the regularization coefficient, then the update process of PZ_v can be described by Formula 18:

$$PZ_v \leftarrow PZ_v+\beta\left(Error_{vi}-\mu PZ_v\right) \tag{18}$$

The update process of PZ_i can be described by Formula 19:

$$PZ_i \leftarrow PZ_i+\beta\left(Error_{vi}-\mu PZ_i\right) \tag{19}$$

User u is one among the K nearest neighbors of user v, then there is:

$$\forall u\in O\left(v,k\right) \tag{20}$$

The update process of ω_{vu} can be described by Formula 18:

$$\omega_{vu} \leftarrow \omega_{vu}+\beta\left\lfloor r_{vi}\left(e_{ui}-PZ_{ui}\right)-\mu\omega_{vu}\right\rfloor \tag{21}$$

2) If the website platform doesn't have the history scoring information of user v, then replace ω_{vu} with $S(v, u)$.

C. Recommendation algorithm training and implementation

In order to effectively reduce the space complexity while extracting the potential g-dimensional vector of the hidden features of user preferences, this paper combines the K nearest neighbor model established based on user scoring information in previous chapter with the score matrix decomposition method; according to the form of $E=V^T*W$, the score matrix E is decomposed into the product of user set V and object set W (objects to be evaluated) to further complete the prediction of the missing values of the matrix. The loss function of the personalized recommendation system can be described by Formula 22:

$$\sum\left(e_{vi}-q_v^T*w_i\right)^2+\mu\left(\left\|q_v\right\|^2+\left\|w_i\right\|^2\right) \tag{22}$$

where, $\mu(\|q_v\|^2+\|w_i\|^2)$ can be used to avoid overfitting of the personalized recommendation system training. Correspondingly, according to the stochastic gradient descent method, the update process of q_v can be described by Formula 23:

$$q_{vt} \leftarrow q_{vt}+\beta\left[Error_{vi}*w_{it}-\mu q_{vt}\right] \tag{23}$$

The update process ofw_i can be described by Formula 23:

$$w_{it} \leftarrow w_{it}+\beta\left[Error_{vi}*w_{vt}-\mu w_{it}\right] \tag{24}$$

In the training process of the constructed personalized recommendation system, at first, it needs to initialize q_v and w_i, namely the g-dimensional hidden feature vectors of user v and the i-th evaluation object, and fill in with random numbers. For a known score e_{vi}, the predicted score $e'_{vi}=q_v^T*w_i$ and the prediction error $Error_{vi} = e_{vi}-e'_{vi}$ can be obtained through calculation. After several iterations, matrices V and W that are decomposed according to the form of $E=V^T*W$ can be obtained. During this process, the predicted score can be obtained by Formula 25:

$$e'_{vi}=PZ_{vi}+\sum_{u\in O(v,k)}\left(e_{ui}-PZ_{ui}\right)*\omega_{vu}-q_v^T*w_i \tag{25}$$

The objective function at this time is:

$$\sum\left(e_{vi}-\lambda-PZ_v-PZ_i-\sum_{u\in O(v,k)}\left(e_{ui}-PZ_{ui}\right)*\omega_{vu}-q_v^T*w_i\right)^2+\mu\left(PZ_v^2+PZ_i^2+\sum_{u\in O(v,k)}\omega_{vu}^2+\left\|q_v\right\|^2+\left\|w_i\right\|^2\right) \tag{26}$$

To minimize the above formula, similarly, the stochastic gradient descent method could be adopted to take the partial derivatives of PZ_v, PZ_i, q_v, w_i, and ω_{vu}, and the parameters could be adjusted and optimized through several iterations. That is, construct the training set of the personalized recommendation system based on history scoring information of users, perform learning training on the system according to the above-mentioned learning process, and obtain each user's user bias items and hidden feature vector in V, and the object bias items and hidden feature vector of each evaluation object in W, and the weight ω_{vu} of K nearest neighbors of the user v. It should be noted that, before solving the parameters such as PZ_v, PZ_i, q_v, w_i, and ω_{vu}, it is necessary to determine the regularization coefficient, the maximum number of iterations, and the learning

speed of the personalized recommendation system according to the experience values of the stochastic gradient descent method.

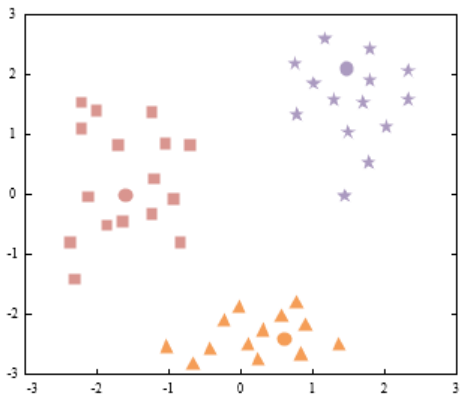
IV. Experimental Results and Analysis

The experimental data are the evaluations made by lots of users on websites. The data have been widely and successfully used to simulate personalized recommendation algorithms. There are 1,054 users, 1,743 items to be evaluated, and 140,000 plus evaluations in the dataset.

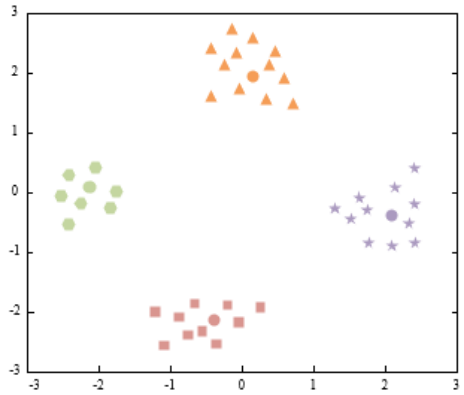
Table 5. Performance evaluation of the clustering algorithm before and after improvement

Algorithm	Dataset No.	Cluster No.	Error rate	Correct rate
Before improvement	1	1	2.3	97.1
		2	1.5	
		3	4.3	
	2	1	1.5	96.2
		2	1.8	
		3	1.9	
		4	1.1	
After improvement	1	1	0.3	99.8
		2	0.5	
		3	0.1	
	2	1	1.4	99.1
		2	0.6	
		3	0	
		4	1.9	

The original and improved k-means algorithms were run on two randomly generated Web user behavior data sets, Table 5 gives the comparison of the clustering performance evaluation results of the k-means algorithm before and after the improvement. According to the table, before improvement, the error rates of clustering categories 1, 2 and 3 of dataset 1 are 2.3%, 1.5% and 4.3%, respectively, and the correct rate of clustering is 97.1%. The error rates of clustering categories 1, 2, 3, and 4 are 1.5%, 1.8%, 1.9%, and 1.1%, respectively, the correct rate of clustering is 96.2%. After the algorithm is improved, the error rates of clustering categories 1, 2 and 3 of dataset 1 are 0.3%, 0.5%, and 0.1%, respectively, and the correct rate of clustering is 99.8%. The error rates of clustering categories 1, 2, 3, and 4 of dataset 2 are 1.4%, 0.6%, 0%, and 1.9%, respectively, and the correct rate of clustering is 99.1%.



(a)



(b)

Fig. 7 The clustering effect of the improved k-means algorithm based on Web data mining

Table 6. Comparison of the performance of the clustering algorithm before and after improvement

Item	Dataset No.	Before improvement			After improvement		
		Poor	Good	Average	Poor	Good	Average
Correct rate	1	63.5	88.2	85.62	88.5	87.5	82.1
	2	53.1	71.6	68.9	71.6	71.6	71.6
Number of iterations	1	4	12	8	7	6	5
	2	3	10	7.2	5	5	5
Running time	1	27	62	45	1039	1039	1039
	2	41	112	71	1972	1972	1972

After calculation we can see that, the clustering error rate of dataset 1 after algorithm improvement is reduced by 1%-4% compared with that before improvement, and the clustering correct rate has increased by 2.7%. The clustering error rate of dataset 2 after algorithm improvement is reduced by 0.1%-1.9% compared with that before improvement, and the clustering correct rate has increased by 2.9%. These results can verify that the improved algorithm does have advantages in terms of clustering error rate and correct rate. Figure 7 gives the clustering effect of the algorithm on the two datasets.

Table 6 compares the performance of the clustering algorithm before and after improvement. According to the table, the improved k-means algorithm that runs once has higher clustering correct rate and less convergence iteration times than the traditional k-means algorithm that runs 10 times. Although the calculation of parameters such as the mean, standard deviation, and density function value of the samples has resulted in higher time consumption of the improved algorithm when determining the initial cluster centers, its advantage of achieving the ideal clustering effect with only one-run is quite obvious.

This paper measured the recommendation accuracy of the constructed personalized recommendation system based on the mean absolute error (MAE). The experiments mainly deal with two situations: (1) Fixing the number of nearest neighbors, and observing the error changes as the number of hidden features gradually increases from 20 to 400; (2) Fixing the number of hidden features, and observing the error changes as the number of nearest neighbors gradually increases from 30 to 50.

Table 7 shows the changes in the prediction results when the number of hidden features of the user is 20, 50, 100, 200, and 400 respectively, and Figure 9 gives the histogram of the MAE of the constructed system under different hidden feature numbers in case of different numbers of nearest neighbors, which can reflect the changes more intuitively.

Table 7. Prediction results of the algorithm under different hidden feature numbers

K \ g	20	50	100	200	400
10	0.921	0.913	0.821	0.871	0.872
20	0.926	0.921	0.846	0.823	0.876
30	0.937	0.928	0.865	0.759	0.881
40	0.929	0.920	0.796	0.805	0.880
50	0.927	0.906	0.753	0.876	0.879

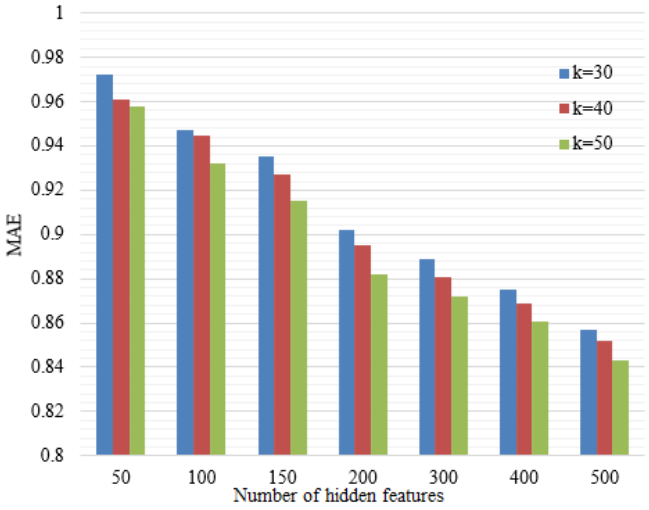


Fig. 8 Histogram of MAE under different hidden feature numbers in case of different numbers of nearest neighbors

According to Figure 8, when the user's hidden feature number remains unchanged, as the number of nearest neighbors increases, the prediction accuracy of the score of user preference increases with it. When the number of nearest neighbors is fixed and the number of hidden features increases, the prediction accuracy of the score of user preference decreases with it. These results indicate that, in the personalized recommendation system, the score matrix after SVD is greatly affected by the number of user's hidden features, and this further affects the predication accuracy of the personalized recommendation algorithm.

Table 8. Prediction results of different algorithms

K \ Algorithm	Similarity calculation	Traditional matrix factorization	Proposed algorithm
10	1.082	0.972	0.947
20	1.058	0.967	0.925
30	1.057	0.951	0.872
40	1.036	0.938	0.879
50	0.975	0.939	0.871

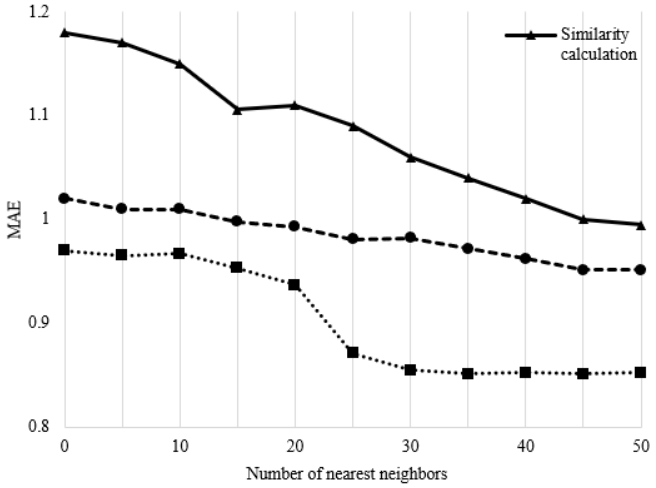


Fig. 9 Prediction errors of different algorithms

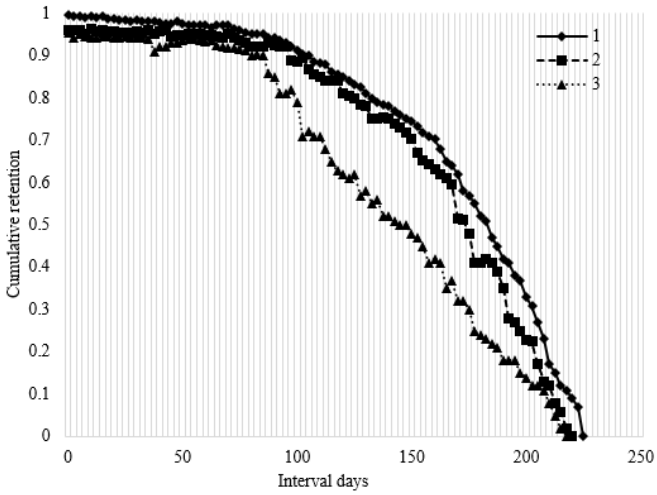


Fig. 10 Changes in the cumulative retention value of users

Then, the proposed algorithm was compared with the similarity calculation and the traditional matrix factorization algorithm. Table 8 and Figure 9 respectively show the prediction results and the prediction error curves of these algorithms. In case of fixed number of hidden features, as the number of nearest neighbors increases, the MAE values of the three algorithms decrease, but the MAE of the proposed algorithm is the smallest, which has verified that the proposed algorithm has better personalized recommendation performance and accuracy than the other two algorithms. Figure 10 shows the changes in the cumulative retention value of users after the recommendation algorithms are implemented. According to the figure, there are certain differences in the cumulative retention of female users 1, male users 2, and unknown gender users. Users who do not provide complete registration information leave more quickly, while female users leave more slowly under the condition of implemented algorithm. For this situation, this paper suggests that the website operators need to increase the attention of male users via measures such as rewards and promotions, and encourage users to give complete information.

V. Conclusion

This paper constructed a network user behavior identification and personalized recommendation model based on Web data mining, gave the process of user behavior analysis based on Web data mining, improved the traditional k-means algorithm, and gave the detailed execution steps of the improved algorithm. Then, through experiments, the performance of the k-means algorithm before and after the improvement was compared, and the results proved the advantages of the improved algorithm in the clustering error rate and correct rate. After that, this paper elaborated on the K nearest neighbor model based on user scoring information, the score matrix decomposition method, and the personalized recommendation method for network users, and experimental results gave the changes in MAE of the proposed personalized recommendation system under different numbers of nearest neighbors. After singular value decomposition, the scoring matrix of the personalized recommendation system is greatly affected by the number of hidden features of users. This further validates the effectiveness and accuracy of the constructed model. Compared with different algorithms, our recommendation algorithm was found to outperform the others in personalized recommendation quality and accuracy. Finally, the change in cumulative retention of users was tested, and several suggestions on personalized recommendation were presented for relevant website operators.

Based on the traditional web data mining technology, this paper derives a web behavior analysis model, which achieves clear and highly accurate personalized recommendation. Combined with the massive sales and transaction data gathered by actual enterprises, our method can promote the construction of the relevant web platforms, speed up production and sales, and enhance the competitiveness of logistics enterprises.

REFERENCES

[1] Z. Wu, L. Tian, Z. Wang, Y. Wang, "Web User Behavior Trust Evaluation Model Based on Fuzzy Petri Net," In 2021 IEEE 6th International Conference on Big Data Analytics (ICBDA), pp. 344-348, 2021.

[2] P. Dhanalakshmi, "A novel frequent pattern mining technique for prediction of user behavior on web stream data," Ingenierie des Systemes d'Information, vol. 24, no. 1, pp. 51-56, 2019.

[3] H. Kawazu, F. Toriumi, M. Takano, K. Wada, I. Fukuda, "Analytical method of web user behavior using Hidden Markov Model," In 2016 IEEE International Conference on Big Data (Big Data), pp. 2518-2524, 2016.

[4] R. X. Zhang, "Design and application of a prediction model for user purchase intention based on big data analysis," Ingénierie des Systèmes d'Information, vol. 25, no. 3, pp. 311-317, 2020.

[5] R. Ojino, "User's profile ontology-based semantic model for personalized hotel room recommendation in the web of things: student research abstract," In Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing, pp. 2314-2316, 2019.

[6] S. Kundu, L. Garg, "Web log analyzer tools: A comparative study to analyze user behavior," In 2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence, pp. 17-24, 2017.

[7] S. Bulla, B. B. Rao, "Performance and cost analysis of web application in elastic cloud environment," Ingénierie des Systèmes d'Information, vol. 24, no. 4, pp. 385-389, 2019.

[8] G. Deepak, B. N. Shwetha, C. N. Pushpa, J. Thriveni, K. R. Venugopal, "A hybridized semantic trust-based framework for personalized web page recommendation," International Journal of Computers and Applications, vol. 42, no. 8, pp. 729-739, 2020.

[9] J. Bhavithra, A. Saradha, "Personalized web page recommendation using case-based clustering and weighted association rule mining," Cluster Computing, vol. 22, no. 3, pp. 6991-7002, 2019.

[10] M. M. Rahman, N. A. Abdullah, "A personalized group-based recommendation approach for Web search in E-learning," IEEE Access, vol. 6, pp. 34166-34178, 2018.

[11] Y. Q. Li, B. D. Xu, "Personalized Recommendation Method Based on Web Log Mining," In 2018 International Conference on Smart Grid and Electrical Automation (ICSGEA), pp. 416-419, 2018.

[12] H. Liu, X. Zhang, J. Li, B. Wang, "A Novel Collective User Web Behavior Simulation Method," CMC-Computers Materials & Continua, vol. 66, no. 3, pp. 2539-2553, 2021.

[13] A. Rosyidah, I. Surjandari, "Mining Web Log Data for Personalized Recommendation System," In 2018 6th International Conference on Information and Communication Technology (ICoICT), pp. 441-446, 2018.

[14] D. Herath, L. Jayaratne, "A personalized web content recommendation system for E-learners in E-learning environment," In 2017 National Information Technology Conference (NITC), pp. 89-95, 2017.

[15] K. Su, B. Xiao, B. Liu, H. Zhang, Z. Zhang, "TAP: A personalized trust-aware QoS prediction approach for web service recommendation," Knowledge-Based Systems, vol. 115, pp. 55-65, 2017.

[16] P. Dhanalakshmi, "A Novel Frequent Pattern Mining Technique for Prediction of User Behavior on Web Stream Data," Ingénierie des Systèmes d Inf., vol. 24, no. 1, pp. 51-56, 2019.

[17] J. Offutt, S. Thummala, "Testing concurrent user behavior of synchronous web applications with Petri nets," Software & Systems Modeling, vol. 18, no. 2, pp. 913-936, 2019.

[18] J. Sachse, "The influence of snippet length on user behavior in mobile web search," Aslib Journal of Information Management, 2019.

[19] G. B. Vin'cius, S. L. Corr, V. J. D. S. Rodrigues, K. V. Cardoso, "Characterizing User Behavior on Web Mapping Systems Using Real-World Data," In 2018 IEEE Symposium on Computers and Communications (ISCC), pp. 01056-01061, 2018.

[20] W. G. Siqueira, L. A. Baldochi, "Leveraging analysis of user behavior from Web usage extraction over DOM-tree structure," In International Conference on Web Engineering, pp. 185-192, 2018.

[21] N. Mahyavanshi, M. Patil, V. Kulkarni, "A realistic study of user behavior for refining web usability," In 2017 International Conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC), pp. 450-453, 2017.

[22] Y. Yamamoto, S. Shimada, "Analysis on Effect of Disputed Topic Suggestion for User Behavior in Web Search," Transactions of the Japanese Society for Artificial Intelligence, vol. 32, no. 1, pp. 1-12, 2017.
[23] B. Walek, "Creating adaptive web recommendation system based on user behavior," In Journal of Physics: Conference Series, vol. 933, no. 1, pp. 012014, 2017.
[24] C. Bernaschina, M. Brambilla, A. Mauri, E. Umuhoza, "A big data analysis framework for model-based web user behavior analytics," In International Conference on Web Engineering, pp. 98-114, 2017.
[25] W. Van Woensel, W. Baig, S. S. R. Abidi, S. Abidi, "A Semantic Web Framework for Behavioral User Modeling and Action Planning for Personalized Behavior Modification," In Swatls, 2017.
[26] C. Bernaschina, M. Brambilla, T. Koka, A. Mauri, E. Umuhoza, "Integrating modeling languages and web logs for enhanced user behavior analytics," In Proceedings of the 26th International Conference on World Wide Web Companion, pp. 171-175, 2017.
[27] Alshreef Abed, Jingling Yuan, Lin Li, Based SVM Distinct Stages Framework Data Mining Technique Approach for Text Extraction, WSEAS Transactions on Information Science and Applications, ISSN / E-ISSN: 1790-0832 / 2224-3402, Volume 16, 2019, Art. #12, pp. 100-110.
[28] W. P. Wang, S. S. Tu, X. Y. Huang, "IKM-NCS: A Novel Clustering Scheme Based on Improved K-Means Algorithm," Engineering World, vol. 1, pp. 103-108, 2019.