Personalized Diet Recommendation Based on K-means and Collaborative Filtering Algorithm

Zhicai Yuan\textsuperscript{1,a} Fang Luo\textsuperscript{1,b*}
\textsuperscript{1}School of Computer Science and Technology, Wuhan University of Technology
1178 Heping Road, Wuhan, China; +86 15171495467
\textsuperscript{2}School of Computer Science and Technology, Wuhan University of Technology
1178 Heping Road, Wuhan, China; +86 13071231823
\textsuperscript{a}E-mail: yuanzhicai@whut.edu.cn \textsuperscript{*}Corresponding author: \textsuperscript{b}E-mail: luof@whut.edu.cn

ABSTRACT: With the improvement of people's living standards, people pay more and more attention to the health of diet, and traditional dietary recommendations are difficult to meet the user's dietary preferences and nutritional balance. This paper first uses the k-means clustering algorithm to divide the food set into multiple disjoint subsets, then uses the user-based collaborative filtering algorithm to recommend the food that the user may like. The recommended food and food in standard recipes set by user's own situation are in the same cluster, which meets the user's nutritional balance. The experimental results show that the recommended effect of this method is effective, and the recommended accuracy rate is over 70%.

CCS Concepts
\begin{itemize}
\item Computing methodologies \rightarrow Artificial intelligence.
\end{itemize}

1. INTRODUCTION
With the continuous improvement of people's living standards, in terms of diet, people are more and more concerned about the health and nutrition balance of diet. Nutritional imbalances can easily cause many diseases such as obesity, high blood pressure and diabetes. Healthy Diet means a nutritional balance instead of a blind pursuit of high nutrition. It is difficult for people to balance between various nutrients and their contents depend on people's life experience. Nowadays, people have long been accustomed to using mobile phones to solve various life problems such as food, clothing, housing and transportation. Therefore, various dietary recommendation softwares came into being.

Most of the existing dietary recommendation methods begin with the user's dietary preferences or nutritional balance. The recommend algorithm of the user's from the user's dietary preferences mainly includes user-based collaborative filtering algorithms, content-based collaborative filtering algorithms [1], and so on. Yushan Wang et al [2] analyzed the user's dietary records, used the user-based collaborative filtering algorithm, selected neighbors to weight the food recommendation, and used the roulette gambling method to diversify the recommended food. Cheng et al [3] used a content-based collaborative filtering algorithm to analyze the similarity between recipes by saving the recipe records
viewed by the user, and recommending the recipes that the user might like. Wei et al [4] proposed a collaborative filtering algorithm based on Co-Clustering smoothing to solve the cold start and sparsity problems common to collaborative filtering algorithms. The algorithms recommended from the aspect of nutritional balance mainly include multi-objective particle swarm optimization algorithm [5], Apriori algorithm [6] and so on. Jixin Zhang et al [7] determined the three nutrient contents of energy, protein and calcium required by the user according to the user's age, gender and other information, and used the multi-objective particle swarm optimization algorithm to make the recommended food meet the needs of these three nutrients. Wanzhen Zhou et al [8] used k-means clustering algorithm and Apriori algorithm to realize the association rule mining between food nutrient content, so that users can choose certain foods in different “clusters” according to their own needs for nutrients.

In summary, most of the existing dietary recommendation schemes are difficult to satisfy the needs of the user's dietary preferences and nutritional at the same time. The recommendation scheme based on the user's dietary preference only recommends the favorite dishes to the user, which often causes the user to develop a single eating habit and results in malnutrition. The recommended scheme based on nutritional balance, although giving clear nutrients and food weight, does not take the individual needs of the user into account. This paper proposes a recommended scheme for both types of users, and can satisfy the needs of different types of users. According to the Chinese food ingredient list [9], 100 kinds of common foods are divided into k disjoint "clusters" according to the k-means algorithm. Users can set a standard recipe according to their own situation, or use traditional healthy recipes. According to the user's dietary record, a collaborative filtering algorithm based on the user's dietary preference is used to calculate the recommended value of the food in the "cluster" of each dish in the standard recipe, and recommend the food with the higher recommended value to the user. It not only meets the nutritional needs of users, but also meets the dietary preferences of users.

2. THEORETICAL BASIS OF K-MEANS ALGORITHM

The purpose of clustering is to divide a data set into a number of disjoint subsets according to the similarity of certain attributes. Each subset is called a "cluster", which makes the objects in the same "cluster" have higher similarity on these attributes, and the similarity between different "cluster" objects is lower.

2.1 Principle of k-means algorithm

The k-means algorithm is one of the most classical algorithms in clustering algorithms, proposed by Mac Queen J. in 1967. The algorithm divides the data set $X=\{x_1,x_2,\ldots,x_n\}$ of $n$ foods with $d$-dimensional properties into $k$ disjoint "clusters" $Y=\{y_1,y_2,\ldots,y_k\}$ according to the similarity degree of the attributes. Among them, $y_i \cap y_j = \emptyset$ and $X = \bigcup_{i=1}^{k} y_i (1 \leq i \leq k, 1 \leq j \leq k, i \neq j)$. The similarity of various foods in each "cluster" is measured by the sum of squared errors ($SSE$) of the "cluster". The $SSE$ is defined as:

$$SSE = \sum_{i=1}^{k} \sum_{x \in y_i} ||x - \mu_i||^2$$

(1)

Where $\mu_i = \frac{1}{|y_i|} \sum_{x \in y_i} x$ is the mean center of the i-th cluster $y_i$, and the smaller the $SSE$, the higher the similarity of the food in the cluster.

2.2 Using the elbow method to determine the k value

The number of clusters for different data sets is different. For a data set, when the $k$ value is smaller than the real cluster number of the data set, as the $k$ value increases, the degree of aggregation within the cluster increases faster, and $SSE$ decreases rapidly. Once the $k$ value exceeds the actual cluster number of the data set, the decrease of $SSE$ is rapidly slowed down, and the degree of the cluster’s return is reduced. Therefore, the relationship between $k$ and $SSE$ is an elbow-shaped line graph, in which the $k$ value corresponding to the "elbow" is the actual cluster number of the data set.

The elbow pseudo code is as follows:

Input: $n$ sets of foods $X$ with $d$ nutrients and a large initial $k$ value $m$. 

$$SSE = \sum_{i=1}^{k} \sum_{x \in y_i} ||x - \mu_i||^2$$
Output: The SSE of each k value is placed in the array SSE[m].

\[
\text{For } k \leftarrow 1 \text{ to } m \text{ do}
\]

\[
E \leftarrow 0
\]

\[
\text{For } l \leftarrow 1 \text{ to } k \text{ do}
\]

\[
\text{For } x \text{ in } y_l \text{ do}
\]

\[
dist \leftarrow \sqrt{(x - \mu_l)^2}
\]

\[
E \leftarrow E + dist
\]

\[
\text{End for}
\]

\[
\text{End for}
\]

\[
SSE[i] \leftarrow SSE[i] + E
\]

\[
\text{End for}
\]

2.3 k-means algorithm flow

The array SSE[k] is depicted on two-dimensional coordinates with the abscissa k and the ordinate SSE[k]. The best value of k is determined by observing the "elbow".

After obtaining the optimal value of k, k data is randomly selected in the data set X as the initial mean vector \(\{\mu_1, \mu_2, \ldots, \mu_k\}\), and the k data are put into the initialized \(y_i (1 \leq i \leq k)\) "cluster". Follow the steps below to perform clustering:

Input: data set X, initial mean vector \(\{\mu_1, \mu_2, \ldots, \mu_k\}\), the value of k

Output: clustering results

Step 1: Calculate the Euclidean distance \(\text{dist}_i \leftarrow \sqrt{(x_i - \mu_j)^2}\) of each object \(x_i (x_i \in X)\) in the data set X to the mean center.

Step 2: Divide each \(x_i\) into the \(y_j\) cluster with the smallest \(\text{dist}_i\) until all objects are divided.

Step 3: Update all values of \(\mu_i (1 \leq i \leq k)\) according to formula (2).

\[
\mu'_i = \frac{1}{|y_i|} \sum_{x \in y_i} x \quad (2)
\]

Step 4: Repeat the above steps until all \(\mu_k\) no longer changes.

3. Relevant Theory of Collaborative Filtering Algorithm Based on User Preference

3.1 Collaborative filtering model based on user preference

(1) User preferences

User-based collaborative filtering calculates the similarity between users through analyzing the user's diet-frequency matrix. Then it finds a collection of users with similar dietary records and recommends foods that the target users have not eaten but may like to eat. The user-diet frequency matrix \(R(m, n)\) is defined as follows:

\[
R(m, n) = \begin{bmatrix}
R_{11} & R_{12} & \ldots & R_{1n} \\
R_{21} & R_{22} & \ldots & R_{2n} \\
\vdots & \vdots & \ddots & \vdots \\
R_{m1} & R_{m2} & \ldots & R_{mn}
\end{bmatrix}
\]

Where \(R_{ij}\) is the frequency of the diet of the user \(U_i (1 \leq i \leq m)\) versus the j-th \((1 \leq j \leq n)\) food \(F_j\).

Because some users have more dietary record data while others are fewer, especially between new and old users, the dietary data varies greatly. This paper introduces the concept of user's dietary preference and uses the user-diet preference matrix to calculate the similarity between users. The user's dietary preferences are defined as follows:
\[ L_{ij} = \frac{F_{ij}}{N_{i}} \quad (3) \]

Where \( L_{ij} \) is the dietary preference of the user \( U_{i} \) for the \( j \)-th food \( F_{j} \) which is the frequency of the diet of the user \( U_{i} \) for the food \( F_{j} \), and \( N_{i} \) is the total dietary frequency of the user \( U_{i} \).

This paper characterizes the user's dietary preferences by the ratio of the user's dietary frequency to their total diet frequency. Using the formula (3), the user-diet frequency matrix \( R(m,n) \) is transformed into a user-diet preference matrix \( L(m,n) \), which is defined as Table 1.

| user    | food      |
|---------|-----------|
| \( U_{1} \) | \( F_{1} \) | \( ... \) | \( F_{i} \) | \( ... \) | \( F_{n} \) |
| \( L_{1j} \) | \( ... \) | \( L_{ij} \) | \( ... \) | \( L_{in} \) |
| \( \ldots \) | \( \ldots \) | \( \ldots \) | \( \ldots \) |
| \( U_{m} \) | \( L_{m1} \) | \( \ldots \) | \( L_{mj} \) | \( \ldots \) | \( L_{mn} \) |

The user-diet preference matrix \( L(m,n) \) also includes the above \( m \) users and \( n \) kinds of food, and \( L_{ij} \) is the dietary preference of the user \( U_{i} \) (1 \( \leq \) \( i \) \( \leq \) \( m \)) to the food \( F_{j} \) (1 \( \leq \) \( j \) \( \leq \) \( n \))

(2) Model definition

The Pearson correlation coefficient is used to describe the degree of linear correlation between two variables. In order to obtain the user set similar to the target user's dietary preference, the similarity between users needs to be calculated. This paper uses the Pearson correlation coefficient to calculate the similarity between users. When calculating the similarity of dietary preferences among users, each user is treated as the variable, the user's dietary preference for \( n \) different foods as a set of sample values for that variable. The Pearson similarity between the two variables is the value of the similarity between the two users.

**Definition 1**: Assuming that the target user is \( U_{a} \), the similarity \( Sim(U_{a}, U_{b}) \) between user \( U_{a} \) and user \( U_{b} \) is defined as follows:

\[ Sim(U_{a}, U_{b}) = \frac{n \sum (L_{aj} - \bar{L}_{a}) \sum (L_{bj} - \bar{L}_{b})}{\sqrt{n \sum (L_{aj} - \bar{L}_{a})^2} \sqrt{n \sum (L_{bj} - \bar{L}_{b})^2}} \quad (4) \]

Among them, \( Sim(U_{a}, U_{b}) \) is the Pearson similarity between the user \( U_{a} \) and the user \( U_{b} \), \( L_{aj} \) is the dietary preference of the user \( U_{a} \) for the \( j \)-th food \( F_{j} \), and \( L_{bj} \) is the dietary preference of the user \( U_{b} \) for the food \( F_{j} \).

After calculating the similarity between the target and other users, it is necessary to calculate the weighted recommendation value of each food, and select the food with higher food recommendation value to recommend to the target user.

**Definition 2**: Assuming that the target user is \( U_{a} \), then for the target user \( U_{a} \), the weighted recommended value \( R_{aj} \) of the food \( F_{j} \) is defined as follows:

\[ R_{aj} = \sum_{i=0}^{v} Sim(U_{a}, U_{i})*L_{ij} \quad (5) \]

Where \( v \) is the number of neighbors of the target user \( U_{a} \) diet preference, \( Sim(U_{a}, U_{i}) \) is the Pearson similarity between the users \( U_{a} \) and \( U_{i} \), and \( L_{ij} \) is the dietary preference of the user \( U_{i} \) for the food \( F_{j} \). \( R_{aj} \) is a recommended value of the target user \( U_{a} \) for food \( F_{j} \).

3.2 User-based collaborative filtering algorithm

According to the collaborative filtering model based on user preference, the collaborative filtering algorithm based on user preference is described as follows:

Input: user-diet frequency matrix \( R[m][n] \), threshold \( p \), neighbor number \( v \), food clustering result \( Y=\{y_{1}, y_{2}, \ldots, y_{k}\} \) and standard recipe dish \( F=\{f_{1}, f_{2}, \ldots, f_{w}\} \)

Output: Recommended food set \( foodRecmd[w] \)
According to the collaborative filtering model based on user preference, the steps of designing collaborative filtering algorithm based on user preference are as follows:

Step 1: Set the threshold as $p$ and read the user's diet record from the database, ignoring the user whose total diet frequency $N$ is less than $p$. Then construct a user-diet frequency matrix.

Step 2: Calculate the user's diet-preference matrix using Equation 3 and the user's dietary preference matrix.

Step 3: Calculate the Pearson similarity between the target user and other users by using Equation 4 with the user's dietary preference matrix. then select $v$ neighbors with the highest similarity, and save the similarity between $v$ neighbors and the target user $U_i$ in the array $Sim[v][Sim[f]=Sim[U_i][U_j], 0 \leq j < v]$.

Step 4: Find the "cluster" $y_i$ of the standard recipe for each dish $f_i$, that is to say $f_i \in y_i$. The dishes in $y_i$ are sorted according to the recommended values from high to low, and the dishes with higher recommended values are recommended to the user. At the same time, randomly select a part of the "cluster", in which the recommended user's diet record total diet frequency is 0 but the recommended value is higher. Then store these foods in the recommended food set $foodRecmd[w]$.

4. EXPERIMENTAL PROCESS AND ANALYSIS

4.1 Experimental process

First of all, 40 students were used as user objects to count the diet of the past month. The user's daily diet was recorded as a dietary record and stored in the database, which contains nearly 1,000 effective dietary records. From the 1000 records, we randomly selected dietary data of ten consecutive days of student A, including the recommended data and actual dietary records of student A in ten days. Then we counted the amount of foods which were recommended successfully. Finally we calculated the recommended success rate. From the Chinese food ingredient list, 100 common foods and their nutrients were obtained as the initial food library. The attribute set $d = \{protein, calories, fat, carbohydrate, dietary fiber, moisture\}$, and the healthy recipe in the book $F= \{rice, pork, cabbage, fish, milk, potatoes, apples, tomatoes\}$ as the standard recipe for this experiment.

(1) The elbow method determines the value of $k$ and performs cluster analysis on the food.

The elbow method was used to calculate the food set according to the $d$ properties of protein, heat, fat, carbohydrate, dietary fiber and moisture, and the results in Figure 1 were obtained. It can be seen that the optimal value of $k$ is 5. The food was then clustered to divide the initial food set into five "clusters" by the k-means algorithm.

Figure 1. The graph of the squared error sum with k

(2) Recommendations for student A based on dietary preferences of users and the results are shown in Table 2.

Table 2. Ten-day diet data of Student A

| Time (days) | Recommended food ($R_f$)          | Actual food ($A_f$)          |
|-------------|----------------------------------|------------------------------|
| 1           | Porridge, Fritters, Rice,…,Duck  | Soy milk, Fritters, Rice,…,Duck |
According to the records of the 40 students, a collaborative recipe model based on user preferences was used to recommend a recipe for student A. The recommended results are: rice, chicken, eggplant, fish, carrots, tomatoes, cauliflower and apples.

4.2 Analysis of experimental results

This paper uses precision to measure the quality of the recommendation system. The higher the degree of coincidence between the food actually eaten by the user and the recommended food, the higher the accuracy of the recommendation, the better the recommended effect. The accuracy rate is defined as follows:

$$\text{precision} = \frac{\sum \text{count}_i(\mathcal{R}_f \cap \mathcal{A}_f)}{\sum \text{count}_i(\mathcal{R}_f)}$$ \hspace{1cm} (6)

Among them, \(\text{count}_i(\mathcal{R}_f)\) is the number of food recommended by the user on the i-th day, and \(\text{count}_i(\mathcal{R}_f \cap \mathcal{A}_f)\) is the number of foods that the user recommends the food and the actual food on the i-th day, that is, The number of foods successfully recommended on the i-th day. Precision is the accuracy of this ten-day food recommendation.

The calculation yields precision = 0.82. We repeated the experiment to obtain the dietary recommendation data and actual dietary records of the seven students for one month, and calculated the accuracy of the recommendation. The results are shown in Figure 2.

![Figure 2. Precision of the seven students' dietary recommendation.](image)

It can be seen from Fig. 2 that the recommended accuracy rate is almost above 0.70, and the average accuracy rate is 0.77, which indicates that the collaborative filtering algorithm based on user preference is stable, and the recommended recipes satisfy the user's nutritional needs.

5. CONCLUSION

This paper considers the user's dietary preferences and nutritional balance needs to meet the user's personalized dietary needs. Firstly, the k-means algorithm is used to divide the food cluster into independent clusters, and then it analyzes the user's diet records to find \(v\) neighbors whose dietary preferences are close to the target users'. Then, according to the standard recipe, it looks for foods with higher food recommendation values in the “cluster” as recommended foods. After experimental tests, it is verified that the recommended algorithm constructed in this paper can better meet the user's personalized dietary needs.

Acknowledgment

This research was supported by National College Students Innovation and Entrepreneurship Training Program (Item Number: 201810497226)

REFERENCES

[1] Xiaolan Weng, Zhijian Wang. Research Process of Collaborative Filtering Recommendation Algorithm. Computer Engineering and Applications, 2018, 54(01):25-31.
[2] Yushan Wang, Zecong Lin. The Research of Recommendation Algorithm of Food Based on Pearson’s Similarity. China Computer & Communication. 2017(04):100-102.

[3] Teh Lee Cheng, Umi Kalsom Yusof, Mohd Nor Akmal Khalid. Content-Based Filtering Algorithm for Mobile Recipe Application[C]. Pulau Pinang: MySEC, 2014.

[4] Suyun Wei, Jingjing Xiao, Ning Ye. Collaborative Filtering Algorithm Based on Co-Clustering Smoothing. Journal of Computer Research and Development, 2013, 50(2):163-169.

[5] Cheng D, Isidori A, Respondek W, et al. Exact linearization of nonlinear systems with outputs. Math Systems Theory, 1986, 21: 247-255.

[6] Agrawal R, Imielinski T, Swami A. Mining Associations Between Sets of Items in Massive Databases. In: Proc. Of the 1993 ACM SIGMOD Int’l Conf. on Management of Data. NewYork: ACM Press, 1993.

[7] Jixin Zhang, Dexiang Zhou, Hao Zhang, et al. Nutrition Decision Method based on Multi-Objective Particle Swarm Optimization Algorithm. Journal of Henan University of Technology(Natural Science Edition), 2010,31(03):82-85.

[8] Wanzhen Zhou, Jingsen Kan. Analysis of Food Nutritional Components Based on k-means and Apriori Algorithm. Science Technology and Engineering, 2018, 18(17): 211-216.

[9] Yaguang Wang. China Food Composition. Peking University Medical Press, 2009.