Online dating research based on K-means and collaborative filtering

Zhongxian Zhu[1], Xiaoyu Han[2,a], Zhihong Liu[1], Yang Liu[1]

[1]Harbin Institute of Technology, Harbin, China

[2]Harbin University of Science and Technology, Weihai, China

[a]764685090@qq.com

Keywords: CF; SPSS; K-means; Online dating

Abstract: In online dating, it is difficult to grasp how many people should be matched at a time, and it is difficult to grasp who should be matched at a time, so we mainly study how many people should be recommended for online dating, and what kind of people should be recommended to ensure the quality of online dating. Finally, the relationship between the form design of the website and the success rate of online dating is also studied.

1. Introduction

Barry Schwartz, a professor at Swarthmore college of social theory and social action, believes that limiting our choices leads to better outcomes. When faced with so many choices, our attention is distracted from qualities that are important to most people, such as a partner's honesty, dependability and sense of humor. So less choice will make us pay more attention to our partner's inner quality and find the most suitable person. In 1995, match.com was established. At the beginning, the website only had 60,000 registered Users, but by 2004, it had 9 million Users [1]. We can conclude that the dating website has increasingly become the mainstream of today's social development.

Therefore, how to recommend partners to Users and how many people should be recommended at a time, so that Users can not only have a variety of choices, see different types of the opposite sex, but also be able to deeply understand each recommended User.

2. Related work

Our overall framework is shown in figure 1.

For the first part, first, we divide Users into two types of Users. The one is new Users who have just registered on the site. The other is old Users who have already registered and browsed. Since the new Users did not browse on the website, we used the fuzzy matching technology of the database to search the heterosexual Users who meet the criteria of the User's mate selection in the database, and then made recommendations from high to low according to the score of each User. For the old Users, we use the old Users' rating of the browsed opposite sex, based on the collaborative filtering recommendation algorithm, to find the Users with similar criteria for mate selection, and recommend the opposite sex browsed by another User to the User who has not browsed.

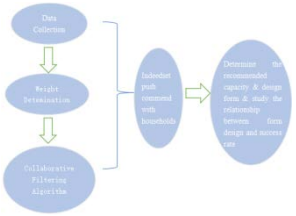

Figure 1 Overall framework

For the second part, we use the k-means to divide the heterogeneity that Users have browsed into two categories, one is the heterogeneity that Users are not very interested in, and the other is the heterogeneity that Users are more interested in. Then we use the sample size formula of parameter estimation to calculate how many Users should be selected as samples to calculate the

  DOI: 10.25236/mmmce.2019.123

recommendation number of each User, which is representative. After selecting the sample, we take the average of the samples as the recommendation number for each User.

We made the following assumptions.

1) Suppose that the data we get from the query is true and reliable;

2) It is assumed that all information of the selected data is complete and the influence of incomplete information on the data is ignored.

3) Ignoring the subjectivity of individual ratings;

4) Let's assume that the User we choose can represent the entire User.

3. Symbol Description

Table 1. Symbol Descriptio

Symbol	Description
Xi	The score of item i
Wi	The weight of item i
score i	The score of User i
P(x,y)	Pearson's correlation coefficient
n	Sample size -- recommendations

4. Algorithm Framework & Algorithm Verification

4.1 Data Definition

We collected data of different Users from domestic dating websites [2], including age, education background, with or without children, whether to buy a car, whether to buy a house, height, weight, monthly salary, housework level, pet preference, smoking degree, drinking degree, interest compliance degree, marital status and occupation.

We select 100 people to collect their information, including 50 males and 50 females. The remaining 50 members of the opposite sex were scored according to their own criteria for mate selection, and the scoring rules are shown as follow.

Age: Scores range from one to ten, the higher the score, the higher the satisfaction, the higher the age score near the median of the age range, and the lower the score

Education Background: Scores range from one to ten, The higher the score, the higher the satisfaction, the full score of the academic qualifications in the mate selection criteria, and the remaining academic qualifications are distributed according to the positive and the scores.

With or Without Children: Zero points for with children and ten points for without children.

Whether to Buy A Car: Zero points for with a car and ten points for without a car.

Whether to Buy A House: Zero points for with a house and ten points for without a house.

Height: Scores range from one to ten, the higher the satisfaction, and the higher the height score near the median of the height interval, while the lower the score.

Weight: Scores range from one to ten, the higher the score, the higher the satisfaction, the higher the body weight score near the median of the weight interval, and vice versa.

Monthly Salary: Scores range from one to ten, within the prescribed range, the higher the monthly salary, the higher the score.

Housework Level: Scores range from one to ten, the more skilled the housework, the higher the score.

Pet Preference: Scores range from one to ten, the more you love your pet, the higher your score.

Smoking Degree: Scores range from one to ten, the lower the level of smoking, the higher the score.

Drinking Degree: Scores range from one to ten, the lower the alcohol addiction, the higher the score.

Interest Compliance Degree: Scores range from one to ten, The more consistent the interest, the

higher the score.

Occupation: Scores range from one to ten, the more stable the career, the safer the job, the higher the score.

Marital Status: The score is zero for married and ten for unmarried

4.2 EWM to Determine The Weight

Step One: Data Standardization

Standardize the data of each index, It is assume that wo have k indices $X_1, X_2, ..., X_k$, there into $X_i = \{x_1, x_2, ..., x_n\}$. Suppose that the standardized value of each indicator data is $Y_1, Y_2, ..., Y_k$, then

$$Y_{ij} = \frac{x_{ij} - \min(x_i)}{\max(x_i) - \min(x_i)}$$

Step Two: Find The Information Entropy of Each Index

According to the definition of entropy in information theory, the information entropy of a set of data $E_j = -\ln(n)^{-1} \sum_{i=1}^{n} p_{ij} \ln p_{ij}$. Thereinto

$$p_{ij} = \frac{Y_{ij}}{\sum_{i=1}^{n} Y_{ij}}$$

if $p_{ij} = 0$, we have $\lim_{p_{ij} \to 0} p_{ij} \ln p_{ij} = 0$.

Step Three: Determine The Weight of Each Index

According to the calculation formula of information entropy, the information entropy of each index is calculated as $E_1, E_2, ..., E_k$. The weight of each index is calculated by information entropy

$$W_i = \frac{1 - E_i}{k - \sum E_i} (i = 1, 2, ..., k)$$

4.3 Comprehensive Weighting Results in Scores

The proportion of height, weight, age and other indicators obtained by EWM is multiplied by their scores, and finally the comprehensive evaluation score of the User on another member of the opposite sex is obtained.

$$Score_i = \sum W_i \bullet X_i$$

Where W_i is the weight obtained by the entropy weight method, X_i is the score of each index, and $Score$ is the total score given by the User to another User of the opposite sex.

4.4 New User Mate Matching

- Fuzzy Matching

Through browsing domestic and foreign dating websites, we find that each dating website will ask registered Users to fill in their own criteria for choosing a spouse, so we can use the fuzzy matching technology of the database (keyword search) to screen out the other half that meets the User's criteria for choosing a spouse.

- Match based on popularity

Users are selected based on fuzzy matching [3]. According to the research on the selected dating websites, it is found that the popularity of the Users can be measured by such indicators as charm value on dating websites. Therefore, we recommend the Users from high to low according to the charm value.

- Old User Companion Matching - Collaborative Filtering Algorithm

For old Users, we have used the mutual scores between Users and Users of the opposite sex to analyze the match situation of the 14th male User.

According to certain probability to select indicators for several 0 first, means that the User does not have read this female Users of any information, then we use Pearson correlation evaluation to

calculate the correlation coefficient, and determined the male of similarity between Users, according to the weighted scores for sorting, according to the score from high to low to male Users is recommended.

1) Looking for Similar Users

First, We use Pearson coefficient to calculate the similarity of mate selection criteria between male Users. Suppose the similarity between Users i and Users j is measured by Pearson correlation coefficient sim(i,j), and the similarity between them can be expressed as :

$$sim(i,j) = \frac{\sum_{c \in I_{ij}} (R_{i,c} - \overline{R_J})}{\sqrt{\sum_{c \in I_{ij}} (R_{i,c} - \overline{R_J})^2} \sqrt{\sum_{c \in I_{ij}} (R_{j,c} - \overline{R_J})^2}}$$

We analyzed the mate selection criteria of 50 male Users and obtained the Pearson correlation coefficient between them. Due to too much data, only 5 male Users were selected to list the similarity coefficient with other Users. The correlation coefficient is shown in the table below.

Table 2. Pearson correlation coefficient

User	**correlation coefficient**	**User**	**correlation coefficient**
User1&User2	0.87716	User1&User3	0.87494
User1&User6	0.39458	User1&User8	0.5221
User1&User9	0.11035	User1&User13	0.62408
User1&User15	0.21973	User1&User18	0.92985
User1&User23	0.89561	User1&User26	-0.95105
User1&User33	-0.90511	User1&User35	0.93985
User1&User38	0.92671	User1&User40	-0.87648
User1&User42	0.48343	User1&User46	0.13898
User1&User48	-0.86947	User4&User2	0.08471
User4&User3	0.11099	User4&User6	-0.88979
User4&User8	0.39098	User4&User9	-0.89273
User4&User13	0.89257	User1&User15	-0.89436
User4&User18	0.9206	User4&User23	-0.88393
User4&User26	0.14989	User4&User33	0.12015
User4&User35	0.02096	User4&User38	0.47335
User4&User40	0.69017	User4&User42	0.22749
User4&User46	0.12002	User4&User48	0.18929
User5&User2	0.30229	User5&User3	0.51109
User5&User6	0.90377	User5&User8	-0.91076
User5&User9	0.00125	User5&User13	0.89827
User5&User15	0.4543	User5&User18	0.33815
User5&User23	0.17658	User5&User26	0.19866
User5&User33	0.93012	User5&User35	0.33569
User5&User40	0.14674	User5&User42	0.32335
User5&User46	0.19994	User5&User48	0.60694

User1 has a high correlation coefficient with User2 and User3, and the standard is similar; User4

is similar to User13 and User18; The User5 standard is similar to the User6, User13, and User33 standards. Therefore, the female User who browsed by User with similar standards in mate selection can be recommended to these several Users.

2) Similar User Matches Partner

The similarity of calculation was used to score every other female User, and the rating of the female User not browsed by the 14th male User was obtained. The rating was sorted from high to low according to the rating, and the top 20 female Users were recommended for the 14th male User. The male User rated the unbrowsed female User from high to low. as shown in table3.

Table 3. Ratings of male Users for female Users

Number	Female Users	Grade
1	User5	4.8419
2	User31	4.8400
3	User32	4.8376
4	User28	4.8361
5	User36	4.8339
6	User39	4.8294
7	User50	4.8257
8	User11	4.8243
9	User34	4.8211
10	User37	4.8211
11	User44	4.8201
12	User19	4.8200
13	User24	4.8193
14	User12	4.8191
15	User7	4.8169
16	User43	4.8140
17	User29	4.8138
18	User25	4.8131
19	User27	4.8117
20	User41	4.8114

5. Recommendation Number Determination

5.1 K-means Confirms Basic Recommended Number

Taking the score of the 14th male User for the rest of Users as an example, the k-means clustering algorithm is used to divide the score into two categories by using SPSS software. The one has a high score, indicating that the User is interested in these female Users. The other has a lower rating, indicating that the User is less interested in these women. The clustering figure is shown in figure 2.

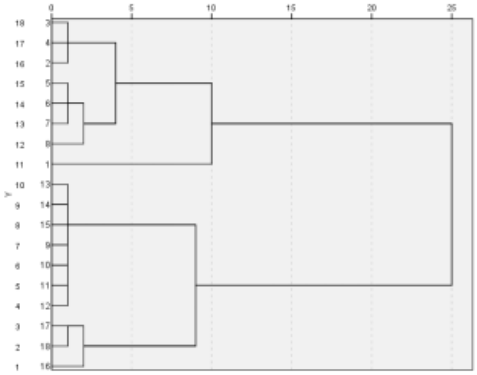

Figure 2 k-means clustering result diagram

In this figure, the 25 female Users who have not browsed this User are divided into two categories. In one category, there are 10 Users, whose score is relatively low, indicating that this User has little interest in these female Users. There's a class of 15 people, which means that this User is interested in these female Users. We conduct research on the basis of these 15 female Users to ensure that this male User is more interested in the recommended female User and can have an in-depth understanding of the recommended female User.

We conducted k-means for all the 50 male Users, and recorded the number of female Users that each male User was interested in.

5.2 Deterministic Recommendation

In order to replace the overall standard of the overall male User with the standard of the selected male User, we use the sample size formula of parameter estimation

$$n = \left(\frac{\sigma Z_{\alpha/2}}{d} \right)^2$$

Where σ is the overall standard deviation, selecting samples of male Users interested in female Users of a standard deviation, as the allowed error range, we shall not exceed ± 3 personal error limits prescribed, made the first mistake, as we assume that a general standard deviation, but the standard deviation is not within the scope of using Z - score to determine a probability.

The 10 users we selected can replace the recommendation number of the 50 male users, so the average of the recommendation number is taken as the final determined recommendation number, which is 8.

6. Form design

According to the indicators listed in the first part, in order to solve the excessive information filled in by dating websites, we use principal component analysis to screen the indicators, so that our indicators can not only describe the basic information of this person, but also enable our algorithm to recommend to other users of the opposite sex based on this information.

We use SPSS for principal component analysis, and select indicators with cumulative percentage over 75%, and the results are shown in table 4 and 5.

Table 4. principal component analysis results of 6 principal components

Component	Initial Eigenvalue			Extract The Sum of The Squares of The Loads		
	Total	Percentage Variance	Accumulation %	Total	Percentage Variance	% Accumulation %
1	3.182	21.214	21.214	3.182	21.214	21.214
2	2.877	19.181	40.395	2.877	19.181	40.395
3	2.080	13.868	54.263	2.080	13.868	54.263
4	1.553	10.352	64.615	1.553	10.352	64.615
5	1.093	7.288	71.903	1.093	7.288	71.903
6	1.016	6.772	78.675	1.016	6.772	78.675

The proportion of each indicator in the component in table 5 was analyzed. In the first component, age, height, degree of smoking addiction and degree of drinking were relatively large. In the second component, education background, monthly salary and occupation account for a larger proportion; In the third component whether to have a house, whether to have a car and marital status accounted for a larger proportion. Therefore, these indicators are selected as the indicators in the form design.

Table 5. Scores for each indicator

	ingredient					
	1	**2**	**3**	**4**	**5**	**6**
Age	.172	.167	-.014	-.036	.376	.026
Education	.020	.262	-.189	.172	.015	.115
Child	.044	-.010	.143	.397	.163	-.595
Car Buying	-.054	-.059	.256	-.431	.098	.179
House Buying	-.088	-.014	.117	.343	-.007	.551
Height	.210	.120	-.125	-.017	-.385	.010
Weight	-.002	.096	.418	.010	.075	-.203
Salary	-.036	.294	-.016	-.128	.226	-.002
Household Level	-.247	.062	.119	.039	.256	.067
Pet preference	-.185	-.141	-.217	.014	.218	.003
Smoking status	.232	-.088	.098	-.053	.124	.309
The degree of alcohol	.243	-.004	.131	.073	.254	.157
Interest coincidence degree	.021	.020	-.246	-.171	.502	-.105
Marital status	-.057	.183	.061	-.283	-.288	-.220
Occupation	-.119	.268	.052	.125	-.065	.238

According to the variance contribution rate of the six common factors in table 6 as the weight, a comprehensive evaluation model can be established by combining factor scores [7].

$$F = 0.21214 f_1 + 0.19181 f_2 + 0.13868 f_3 + 0.10352 f_4 + 0.07288 f_5 + 0.06772 f_6$$

The six common factors have a contribution rate of 78.675%, that is, the six extracted common factors can contain all the indicators and can describe these indicators well, which can be further used for multifactor regression analysis.

7. Success rate analysis

Check The Success Rate of Dating Websites, We search and find that the domestic success rate is 3.9%.

7.1 Use regression models to determine relationships

We conduct Multiple regression analysis, and find the significance level of the overall result is 0.014<0.05, indicating that the model is significant. Table 4 is the coefficient table of multiple regression equation.

Table 6. Regression equation coefficient table of multiple equations

Model	**Unstandized Coefficients**		**Standardized Coefficients**	**t**	**Sig**
	B	Std.Error	Beta		
(Constant)	1.452E-17	0.054		0.000	1.000
REGR factory score 1 for 3	0.079	0.054	0.079	1.469	0.143
REGR factory score 2 for 3	0.189	0.054	0.189	3.530	0.000
REGR factory score 3 for 3	-0.060	0.054	-0.060	-1.115	0.266
REGR factory score 4 for 3	-0.030	0.054	-0.030	-0.568	0.570
REGR factory score 5 for 3	-0.002	0.054	-0.002	-0.044	0.965
REGR factory score 6 for 3	0.001	0.054	0.001	0.013	0.990

In table 6, we may safely draw the factor 2 relatively large impact on the success rate of degree, degree, monthly salary and career and the success rate of correlation is larger, and a significant impact, the factor of 1 second, age, height, degree of addiction and the influence of alcohol degree is a bit weak, factor 3, factor and factor 4 5 negative correlation with the success rate, whether to have a

car, if there is a room and household level degree of the influence degree of the success rate of weak negative correlation. Factor 6 has the least impact on the success rate, indicating that the pet love degree has the weakest impact here. Due to the strong significance of factor 2, the large significance probability of other indicators and the weak significance, the sub-regression equation is not accurate and the ideal effect is not achieved, so we use stepwise regression for analysis.

And we get that by stepwise regression:

Table 7. coefficient table of stepwise regression equation

Model	Unstandized Coefficients		Standardized Coefficients	t	Sig
	B	Std.Error	Beta		
(Constant)	-1.991E-16	0.053		0.000	1.000
X_2 Education	-0.171	0.067	-0.171	-2.566	0.011
X_4:salary	0.153	0.057	0.153	2.686	0.008
X_5 occupation	0.222	0.071	0.222	3.151	0.002

According to the coefficient table of the regression equation, since the significance probability of the constant term is 1, that is, the specific significance, the ternary linear regression equation affecting the success rate of online dating can be obtained as follows:

$$Y = -0.17x_2 + 0.153x_4 + 0.222x_5$$

In the significance test of the above regression equation, the statistical significance probability of the test is 0.000<0.05, indicating that the ternary linear regression equation is highly significant. In the significance test of regression coefficient, the significance rate of degree is 0.011<0.05, the significance rate of monthly salary is 0.008<0.05, and the significance rate of position is 0.002<0.05, indicating that these three indicators are all significant for the success rate of the project.

Among the factors influencing the success rate of online dating, education background, monthly salary and position have a significant impact on the improvement of success rate.

Therefore, we can conclude that if these three indicators are detailed in the form design of online dating sites, the success rate of online dating can be improved.Strengths and Weaknesses

7.2 Strengths

In the first question, we considered the difference between new users and old users to make recommendations, so as to adapt to more situations and have better flexibility.

The timeliness of our recommendations for users is relatively strong. For users with different mate selection criteria, collaborative filtering algorithms can be used to find the right partner more accurately.

It is not possible to make personalized recommendations to a new user immediately after he or she ACTS on a small number of items, because the user similarity table is calculated offline at regular intervals

After a new item has been online, once a user has acted on the item, the new item can be recommended to other users who have similar interests to the user who generated the behavior.

7.3 Weaknesses

In the calculation of the number of recommended users, we only selected 50 male users for analysis because we did not have much data, which was limited and could not reflect the characteristics of all people on the dating website.

In our model, we only analyze the users with complete information, ignoring the impact of incomplete information on the results. However, in real life, users' personal information and mate selection information are not complete, so it may have some impact on the results of the model.

References

[1] Mei peng. Successful enlightenment of foreign online dating sites [J]. Computer knowledge and technology, 2004(06):77-79.

[2] The data is available at http://www.jiayuan.com/

[3] Chen Wei.An Auditing Method Based on Fuzzy Matching in Big Data Environment[J].Chinese Certified Public Accountant,2016(11):84-88+3.

[4] The entropy weight method (EWM) https://blog.csdn.net/qq_32942549/article/details/80019005.

[5] Li xiaoyu. A review of collaborative filtering recommendation algorithms [J]. Journal of shangqiu normal university, 2008, 34(09):7-10.

[6] The code of Recommend function https://blog.csdn.net/google19890102/article/details/28112091

[7] Zhao Ying, Shao Feifei. What factors determine the success of online crowdfunding?——An empirical study based on Jingdong crowdfunding[J].Financial Theory and Practice,2017(03):89-95.