Content-Based Personalized Dating Recommendation System

Yujiao Hu, Xiaolin Gui*, Xinyue Hu, Cong Zeng, Youqi Wu
1, School of Software, Jiangxi Normal University, Nanchang, 330000, China
2, Foreign Languages College, Jiangxi Normal University, Nanchang, 330000, China
* xlgui@jxnu.edu.cn

Abstract: In order to improve the accuracy of recommendation on dating websites, a personalized recommendation algorithm aiming at dating objects is put forward. The algorithm firstly extracts the user's personal characteristic information and defines quantification criteria to quantify the characteristics. Then, the characteristics will be weighted according to the users’ different degree of emphasis on the eigenvalue and similarity between users will be calculated with Euclidean distance. Finally, possible candidates will be recommended according to similarity. The rationality of the algorithm is verified by some experiments.

1. Introduction
With the development of Internet technology, dating websites are playing an increasingly important role in people's daily life. According to the Research Report of China's Online Dating Industry in 2019, the market revenue of China’s online dating industry in 2018 is 4.99 billion yuan, and the penetration rate of online dating industry in the whole industry is 54.4% [1]. Recommendation aiming at dating objects is a key technology of dating websites. Therefore, it is significant for website operators to design a suitable recommendation system for users.

At present, there are some researches on the recommendation system of dating. Literature [2] proposed a dating matching method based on telecom big data. By expanding the dimension of users' personal information and increasing the matching information, the success rate of the match between users could be improved. Due to the fact that only basic information such as name, age and address was included in the telecom data, and there was no information on work, housing, users' preferences and so forth, the recommendation results indeed had limitations. Literature [3] used the matching data processing method of dating, which based on big data and deep learning, to obtain the matching results of its output. The images of both sides of users were inputted into the deep learning neural network, which had been trained according to the image data of married objects. This method only considered the facial factors of both users but did not take some important factors, such as personality, habits and values that affect dating, into account. Literature [4] recommended users of the opposite sex by combining the similarity of pictures selected with the user’s personal information and mate selection conditions. This test method was subjective and could not truly reflect the value orientation of both sides of users. And the recommended users might not have the similar values. To sum up, most of the above recommendation systems make recommendations according to users’ personal information and mate selection conditions. In the process of recommendation, all features are treated equally. Such a recommendation method ignores the different sensitivity of users to different features. For example, some users are more concerned about the height, while others are more concerned about the economic situation. Moreover, traditional content-based method does not capture the semantics of the user interests and cannot handle the ambiguity due to natural language [5] On the basis of summarizing and
analyzing the contents of previous researches, this paper proposes a personalized recommendation scheme for dating. Users can sort the given mate selection conditions, and then the system will weigh the ranked mate selection conditions according to the users' rank. At the same time, this paper also considers the preferences of both sides of users, calculates the similarity between users with Euclidean distance, and makes two-way recommendation for users with high similarity. Therefore, the precision of recommending users can be improved and the personalized recommendation function can be realized.

2. Content-based Recommendation

2.1. Content-based Recommendation

content-based filtering techniques where recommendations are based on the comparison of textual information between users’ information [6]. It finds relationship between two users by matching their contents [7]. Content-based Recommendation generally consists of three steps. First is Item Representation. The system extracts some features from each user's personal data to represent the project. Second step is Profile Learning. The system learns the users' preferences based on the feature data of people that users have followed in the past. The last is Recommendation Generation. By comparing the characteristics of users' preferences and candidates obtained in the above steps, a group of users of the opposite sex with the greatest relevance can be recommended for these users [8].

2.2. Matching Algorithm

Euclidean distance is the easiest method of distance calculation to understand. It refers to the distance between two points as a straight line. Where x and y are n-tuples, i is used to denote their n coordinates (or dimensions), and 1≤p≤∞ [9]. The Lp distance metric can be defined in an n-dimensional space [10]:

$$D_{EUC12} = \sqrt[n]{\sum_{i=1}^{n} (x_{1i} - x_{2i})^p}$$  \hspace{1cm} (1)

In this paper, the octopus collector was used to get 420 city links in Zhenai.com. And then the users’ information in different cities was collected in a circular manner. A total of 15,897 pieces of specific personal information data were gotten, including the users' personal condition information and mate selection information. As the acquired data is irregular and unstructured, it is necessary to extract keywords from the data and delete users with little characteristic information or serious information loss, leaving 4000 pieces of user data with relatively complete information in the end. The data was stored in form of table, and the form of cell data are user's id, gender, personal information, and mate selection condition in turn. As for the features in a users' personal information cell, they are extracted in the way of keywords. For example, personal information in the record of formal schooling characteristic quantitative is as TABLE I:

| Quantitative table of educational characteristics | Feature Description |
|-----------------------------------------------|---------------------|
| Definition                                     | value               |
| Education Background                           |                     |
| High school or blow                           | 0                   |
| Professional School                           | 1                   |
| College                                       | 2                   |
| Bachelor                                      | 3                   |
| Master                                        | 4                   |
| Doctor                                        | 5                   |

For the numerical characteristics of height, age and monthly salary, if it is a value, its value will be reduced by 10 times. If it is in form of numerical interval, the minimum and maximum value will be obtained through the segmentation of '–', and its value will be replaced by the intermediate value
between the minimum and maximum value. The range between age and monthly salary in the dating data is too large, so the Euclidean distance is different from other features in the calculation. The absolute value $S_{Duv}$ as of the monthly salary difference between user u and user v was calculated, and the equation is as follows:

$$S_{Duv} = |S_u - S_v|$$  \hspace{1cm} (2)

$S_u$ and $S_v$ are respective user u’s and user v’s monthly characteristic values. $S_{Duv}$ is the absolute value of difference of salary characteristics between user u and user v. Processing method of $S_{Duv}$ is as TABLE II follows:

![TABLE II](#)

**Monthly Salary characteristic distance processing form**

| Monthly Salary | Feature Description |
|----------------|---------------------|
| processing form | Definition | value |
| $S_{Duv}$      |           |       |
| <3              | 0         |
| 3-6             | 1         |
| 6-9             | 2         |
| 9-12            | 3         |
| 12-15           | 4         |

Age characteristics are processed in the same way as monthly salary characteristics.

Personal information includes ten characteristics, such as marital status, age, height, monthly salary, education, body shape, whether smoking, whether drinking, whether having children and whether wanting children. Mate selection information is consistent with the personal information in data field, and its processing method is also the same. The form of final data is as TABLE III follows:

![TABLE III](#)

**User’s information**

| User’s id | User’s information | Conditions of mate selection |
|-----------|--------------------|------------------------------|
| u         | g                  | m1,m2,m3…m10               | o1,o2,o3…o10               |

3. Experimental Process

The data extracted by octopus is irregular and unstructured Chinese data, but the eigenvalue range of corresponding characteristics of Chinese data is fixed. For example, educational record characteristics have a few characteristic values like high school and below, technical secondary school, junior college, undergraduate course, master and doctor.

The characteristics can be queried through python’s “jieba” segmentation library as a keyword eigenvalue query. It locates the eigenvalue of the corresponding feature and quantifies and identifies according to the custom quantization criteria. Both the personal information in the users’ information lists and the Chinese data information in the mate selection condition were processed in this way.

The data was divided into training data and test data, and the number ratio of training data and test data is 7:3. Then we matched the users in the test data with the users in the training data, for example, we matched the user A in the test data with the user B in the training data. The weighted Euclidean distance $d_1$ between the personal information of user A and the mate choice information of user B is calculated first, and then the weighted Euclidean distance $d_2$ between the mate choice conditions of user A and the personal information of user B is calculated. The total distance $d$ between the two users is obtained by adding $d_1$ and $d_2$. Finally, the distance $d$ is normalized to get $dn$. Combine table I with the above process, the equation of calculating weighted Euclidean distance between user u and user v is as follows:

$$D_{uv} = \sqrt{\sum_{i=1}^{n} \lambda_i (u_i - v_i)^2}$$

The weighted Euclidean distance between two users was calculated by equation (3), and then the
weighted Euclidean distance between each user in the test data and the training data was calculated to form the user distance matrix \( D_{m,n} \). The figure follows as:

\[
\begin{pmatrix}
  u_1 & u_2 & \cdots & u_n \\
  v_1 & d_{11} & d_{12} & \cdots & d_{1n} \\
  v_2 & d_{21} & d_{22} & \cdots & d_{2n} \\
  \vdots & \vdots & \vdots & \ddots & \vdots \\
  v_n & d_{n1} & d_{n2} & \cdots & d_{mn}
\end{pmatrix}
\]

Fig. 1. Weighted distance matrix
d_{ij} represents the weighted Euclidean distance between user vi and user u_j.

4. Recommendation Accuracy

After obtaining the recommendation list, we decided to set a threshold \( \text{min} \) to count the numbers of features that users in the recommendation list meet the requirements of target users since there is no label data. When \( C > \text{min} \), the recommendation is deemed to be successful. Otherwise, the recommendation will fail. When judging whether they meet the standards, there are three situations. First, if the value of feature \( f \) in the mate selection information of user A is "#", then the value is null. We argue that user A has no requirement for this feature, then no matter the corresponding characteristic value of personal information of user B is, it is believed that user B meets the needs of user A in this feature. Second, if the value of characteristic \( f \) in the mate selection information of user A is a specific value, and the corresponding characteristic value of personal information of user B is "#", that is, the corresponding characteristic is null. It is considered that user B does not meet the needs of user A in this feature. Third, if the value of characteristic \( f \) in the mate selection information of user A is an interval, and the corresponding eigenvalue of personal information of user B is a numerical value, and the difference between the corresponding eigenvalue of user B and the median value of the corresponding eigenvalue of user A is less than three, it is deemed that the feature meets the requirements. Otherwise, the requirements are not met. The accuracy equation is as follows:

\[
\text{precision} = \frac{\sum_{i \in S} 1}{\sum_{j \in \text{Rec}} 1}
\]

Different users have diverse recommendation results under their respective weight ranking. If the target users put age, monthly salary and body shape in the first three, the system will give priority to the users with age, monthly salary and body shape that are consistent with the target users. Compared with no weight, the algorithm with weight ranking will give priority to the features with greater weight given by the target user, so as to achieve the effect of personalized recommendation.

5. Conclusion

This paper proposes a new recommendation algorithm to solve the problem that traditional dating websites have low accuracy in personalized recommendation of dating objects. The recommendation algorithm first weights the eigenvalues of the user's independent selection and sorting. Then, the Euclidean distance between users is calculated. Finally, the marriage objects for users are recommended according to Euclidean distance. Experimental results show that the accuracy of this recommendation algorithm is 78% when the number of features selected by users is 13 and the number of recommended
users is 5. Due to the data in this paper originated from the Internet, there are some blank values and the lack of user feedback information, so the results indeed have certain limitations. The next step is to consider improving information integrity and further improving recommendation accuracy.

Acknowledgement
This work is supported by the planned project of Jiangxi key research and development (20181ACE50030).

References
[1] China Online Dating Industry Research Report 2019 [A]. Iresearch Consulting Series Research Report (the second issue of 2019) [C].
[2] X.X. CAI, Y. Wang, A Dating Matching Method and Device Based on Telecom Big Data [P]. Beijing: CN108875069A, 2018-11-23.
[3] D.J. Zhu, Data Processing Methods, Devices, Computer Equipment and Storage Media for Matching of Love Objects Based on Big Data and Deep Learning [P]. Guangdong: CN108038496A, 2018-05-15.
[4] J.C. Wang, A Marriage and Dating Matching Method [P]. Liaoning: CN107689012A, 2018-02-13.
[5] C. Bhagavatam, S. Feldman, R. Power, W. Ammar, Content-based citation recommendation, 2018, [online] Available: https://www.arxiv.org/abs/1802.08301.
[6] R. Dong, L. Tokarchuk, A. Ma, "Digging friendship: Paper recommendation in social network”, Proc. Netw. Electron. Commerce Res. Conf. (NAEC), pp. 21-28, 2009.
[7] Y. Ding, G. Zhang, T. Chambers, M. Song, X. Wang, C. Zhai, "Content-based citation analysis: The next generation of citation analysis", J. Assoc. Inf. Sci. Technol., vol. 65, no. 9, pp. 1820-1833, 2014.
[8] L. Li, A.H. Zhu, T. Su, Research and Implementation of an Improved VSM-BASED Text Similarity Algorithm [J]. Computer Applications and Software, 2012, 29 (2):282-284.
[9] D. Coeurjolly, A. Montanvert, "Optimal separable algorithms to compute the reverse Euclidean distance transformation and discrete medial axis in arbitrary dimension", IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 3, pp. 437-448, Mar. 2007.
[10] T. E. Schouten and E. L. v. d. Broek, "Fast Exact Euclidean Distance (FEED): A New Class of Adaptable Distance Transforms," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 36, no. 11, pp. 2159-2172, 1 Nov. 2014.