




Article

Collaborative Filtering Recommendation Algorithm Based on TF-IDF and User Characteristics

Jianjun Ni ^{1,*} , Yu Cai ¹ , Guangyi Tang ^{1,2}  and Yingjuan Xie ¹ 

¹ College of IOT Engineering, Hohai University, Changzhou 213022, China; caiyu0319@hhu.edu.cn (Y.C.); tang_gy@hhu.edu.cn (G.T.); yjxie@hhu.edu.cn (Y.X.)

² College of Computer and Information, Hohai University, Nanjing 211100, China

* Correspondence: njjhhuc@gmail.com; Tel.: +86-519-85191711

Abstract: The recommendation algorithm is a very important and challenging issue for a personal recommender system. The collaborative filtering recommendation algorithm is one of the most popular and effective recommendation algorithms. However, the traditional collaborative filtering recommendation algorithm does not fully consider the impact of popular items and user characteristics on the recommendation results. To solve these problems, an improved collaborative filtering algorithm is proposed, which is based on the Term Frequency-Inverse Document Frequency (TF-IDF) method and user characteristics. In the proposed algorithm, an improved TF-IDF method is used to calculate the user similarity on the basis of rating data first. Secondly, the multi-dimensional characteristics information of users is used to calculate the user similarity by a fuzzy membership method. Then, the above two user similarities are fused based on an adaptive weighted algorithm. Finally, some experiments are conducted on the movie public data set, and the experimental results show that the proposed method has better performance than that of the state of the art.

Keywords: collaborative recommendation; TF-IDF method; user characteristics; fuzzy membership function; weighted fusion



Citation: Ni, J.; Cai, Y.; Tang, G.; Xie, Y. Collaborative Filtering Recommendation Algorithm Based on TF-IDF and User Characteristics. *Appl. Sci.* **2021**, *11*, 9554. <https://doi.org/10.3390/app11209554>

Academic Editor: Rui Araújo

Received: 31 August 2021

Accepted: 12 October 2021

Published: 14 October 2021

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.



Copyright: © 2021 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the advent of the big data era, information on the Internet has grown exponentially. People have entered the era of information explosion from the past when information was scarce. However, most of this massive amount of information is worthless. The information explosion has made it more and more difficult for people to obtain valuable information from the Internet [1]. To improve the efficiency of production and life, people need some information filtering technologies to filter out useless information. The recommender systems are software tools and techniques providing suggestions for items which are useful to a user. As one of the effective information filtering tools, the personalized recommendation system can help users efficiently obtain information that meets their needs when their needs are unclear [2].

The core of a personalized recommendation system is the recommendation algorithm, which mainly includes the content-based recommendation algorithm, collaborative filtering recommendation algorithm, and hybrid recommendation algorithm [3,4]. Among them, because of the high efficiency, accuracy, and personalization, the collaborative filtering recommendation algorithm has become one of the most effective and extensive application recommendation algorithms [5]. For example, Nakagawa and Ito [6] proposed a recommendation system which can recommend interesting document files to users by collaborative filtering. Yu et al. [7] presented the application of a collaborative filtering algorithm in the field of E-commerce. Park et al. [8] presented a fast collaborative filtering algorithm with a k -nearest neighbor graph. Wu et al. [9] used a collaborative filtering algorithm to improve the prediction accuracy of large-scale recommendation system. Bartolini et al. [10] implemented a personalized recommendation. Although the collaborative filtering algorithm

has been widely used, there are still some problems such as data sparsity, cold start, and information expiration, etc [11].

To solve the problems above, a series of improvements based on the traditional collaborative filtering algorithm were made and achieved some success. For example, Piraste et al. [12] alleviated the sparsity and cold start problems of the matrix using the film type label and director genre. Kumar et al. [13] used matrix decomposition technology to reduce the dimension of the matrix and improve the accuracy of the recommendation result. Sun and Dong [14] proposed a dynamic time drift model considering the influence of user interest changes on similarity in different time periods. Wang et al. [15] proposed a collaborative filtering algorithm combining the KNN model and XGBoost model. Zarzour et al. [16] presented a new effective model-based trust collaborative filtering to improve the quality of recommendation. In addition, there are some collaborative filtering algorithms based on clustering [17], neural networks [18], and various probability models [19]. The above studies optimized the recommendation model to a certain extent and improved the accuracy of the recommendation results, but there are still some problems to be further studied. For example, most of the existing collaborative filtering algorithms only consider the rating information among users, but ignore the user characteristics and the impact of popular items on user similarity, which leads to poor recommendation results.

To further improve the accuracy of recommendation, a collaborative filtering algorithm based on the TF-IDF method and user characteristics is proposed in this paper. In the proposed method, both the rating information and the characteristics of the users are fully considered. The contribution of this paper can be summarized as follows: (1) Based on the rating data, the TF-IDF method is used to calculate the user similarity matrix to punish the impact of popular items on user similarity, and to improve the ability of mining unpopular items. (2) The user characteristics are fully considered in the proposed method, which is used to calculate user similarity based on a fuzzy membership function, to deal with the cold start problem by combining different dimension characteristics information of users. (3) An adaptive weighted algorithm is presented to fuse the two kinds of similarities of users obtained on the above two steps, to form a new user comprehensive similarity for recommendation algorithm. At last, experiments are carried out on real data sets to evaluate the accuracy of the proposed recommendation model. Experimental results show that the proposed algorithm is better than the state-of-the-art algorithms in accuracy.

This paper is organized as follows. Section 2 gives out an overview of related work. The proposed algorithm is presented in Section 3. Section 4 provides the experiments and results analysis. Discussions on the parameters and performance of the proposed algorithm are carried out in Section 5. Section 6 gives out the conclusions.

2. Related Work

The basic idea of the collaborative filtering algorithm can be simply summarized as recommending items of interest to target users who have similar interests [20,21]. As shown in Figure 1, the collaborative filtering algorithm is mainly divided into three steps, namely establishing the user-item rating matrix, finding other users with similar interests to the target users, and finally making recommendations by rating and predicting based on similar users. Traditional collaborative filtering (CF) algorithms are mainly divided into user-based collaborative filtering (UCF) and item-based collaborative filtering (ICF) (see Figure 2). There are many improvements in the collaborative filtering recommendation algorithms, to solve the data sparsity and cold start problem. These existing methods give a good research basis for the recommendation system.

In this paper, the user-based collaborative filtering algorithm is focused on, which is more suitable for responding to the favorite items for groups with similar interests, and the recommendation results are more social. The proposed collaborative recommendation algorithm is similar to those existing TF-IDF-based methods. However, there are many differences between the proposed methods with those existing methods. In the proposed method, the TF-IDF method is applied to rating data, and the user characteristics are

fused to optimize the user similarity and improve the accuracy of rating prediction. It is different from those methods that use a time-dependent similarity measure to compute the user similarity without considering user characteristics [22]. It is also different from those methods that directly calculate the user similarity through the TF-IDF method [23].

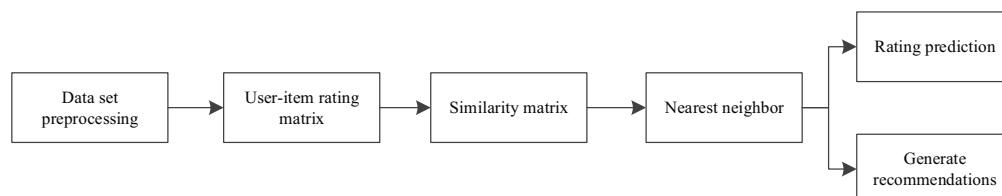


Figure 1. The main architecture of CF model.

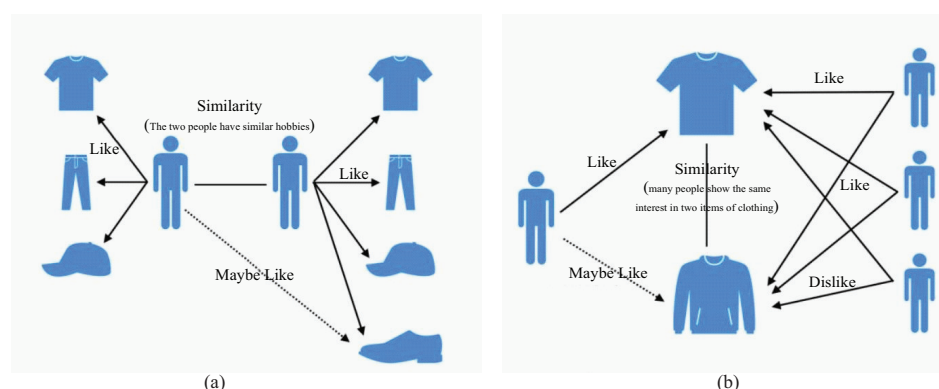


Figure 2. The collaborative filtering algorithms: (a) user-based; (b) item-based.

The user-based collaborative filtering algorithm first needs to calculate the similarity between the target user and other users. Then, some users with high similarity are selected as the nearest neighbor set. Finally, aim at items in the neighbor set and predict all ratings of the target user. The main process of the traditional user-based collaborative filtering algorithm will be described as follows.

2.1. Data Preprocessing

Suppose the data set of a recommender system is $D\{U, I, R\}$, where $U = \{u_1, u_2, \dots, u_n\}$ is the user set of the system, $I = \{i_1, i_2, \dots, i_m\}$ is the item set of the system, and R is a user-item rating matrix. For a data set with m users and n items, the data are preprocessed to obtain a $m \times n$ user-item rating matrix $R(m \times n)$, which is shown as follows:

$$R = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1j} & \dots & r_{1n} \\ r_{21} & r_{22} & \dots & r_{2j} & \dots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ r_{i1} & r_{i2} & \dots & r_{ij} & \dots & r_{in} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ r_{m1} & r_{m2} & \dots & r_{mj} & \dots & r_{mn} \end{pmatrix} \quad (1)$$

where r_{ij} represents the rating data of user U_i for item I_j .

2.2. Similarity Calculation

In the recommender system, there are three main methods used to calculate the similarity between two users: the cosine similarity, adjusted cosine similarity, and pearson similarity [24]. In this study, the pearson similarity will be used, which is calculated

through the common rating items between any two users. The Pearson similarity is shown as follows:

$$Sim_{User}(u, v) = \frac{\sum_{i \in I} (R_{u,i} - \bar{R}_u) \times (R_{v,i} - \bar{R}_v)}{\sqrt{\sum_{i \in I} (R_{u,i} - \bar{R}_u)^2 \sum_{i \in I} (R_{v,i} - \bar{R}_v)^2}} \quad (2)$$

where $R_{u,i}$ and $R_{v,i}$ represent the ratings of user u and user v on the i -th item, respectively; \bar{R}_u and \bar{R}_v represent the average of all the ratings of user u and user v , respectively.

2.3. Generate Recommendation Set

Before rating prediction to generate recommendation, it is necessary to determine the target user's similar neighbor set. A similar neighbor set refers to the set of users who have similar preferences with the target user. In the recommendation system, the most K similar users are usually selected as the nearest neighbor set to form the similar neighbor set of the target user [25].

After the neighbor set of the target user is selected, it combines with all the neighbors' ratings of the items and the similarity between the users to predict the target user's ratings on the test set. The rating prediction is calculated as follows [26]:

$$P_{u,i} = \bar{R}_u + \frac{\sum_{v \in S(u,K) \cap N(i)} Sim_{User}(u, v) \times (R_{v,i} - \bar{R}_v)}{\sum_{v \in S(u,K) \cap N(i)} |Sim_{User}(u, v)|} \quad (3)$$

where $P_{u,i}$ represents the prediction rating of user u for unknown item i ; $S(u, K)$ is the set of K users most similar to user u ; and $N(i)$ represents the set of users who have rated item i . After rating prediction, select the N items with the highest rating from the predicted rating set as the recommendation results to the target users, and the recommendation process ends [27].

3. Proposed Method

As introduced in Section 2, the traditional user-based collaborative filtering algorithms usually only use the user's rating information, but ignore the impact of other aspects of user information and popular items on user similarity. To deal with these problems, an improved collaborative recommendation algorithm (defined as ICFTU) is proposed by combining the Term Frequency-Inverse Document Frequency (TF-IDF) method and user characteristics model. The overall framework of the proposed algorithm is shown in Figure 3, which has three main parts, namely the improved TF-IDF-based method, the improved user characteristics model, and the proposed fusion strategy. The proposed method will be presented as follows.

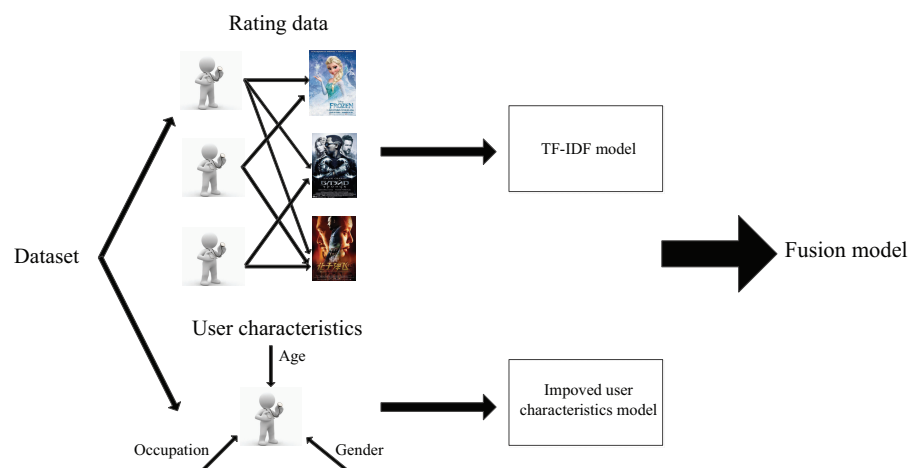


Figure 3. Framework of the proposed collaborative filtering algorithm.

3.1. Improved TF-IDF Based Method

The traditional collaborative filtering algorithm calculates the user similarity matrix based on the user's rating data of items, which is easily affected by popular items. For example, "Shawshank Redemption" is a very good movie. If the user *A* and user *B* both gave the movie "Shawshank Redemption" 5 points, the traditional collaborative filtering algorithm will come to the conclusion that the user *A* and user *B* have high similarity. However, the fact is not necessarily the case. As we know, the same behavior of users on popular items does not mean that they have similar interests. On the contrary, if two users have taken the same behavior on unpopular items, it is more likely that their interests are similar. For example, if both users *A* and *B* have watched a relatively small number of movies, such as musicals, then they can be considered to have similar interests. Therefore, in order to eliminate the impact of popular items on the user similarity, the Term Frequency-Inverse Document Frequency (TF-IDF) method is applied to the traditional collaborative filtering algorithm in this paper, which is used to punish the popular items in the user behavior list. The main reason to use the TF-IDF method is that it is suitable for the problem of weight extraction. In addition, the TF-IDF method is simple and easy to calculate [28].

TF-IDF is a statistical method, which is often used to evaluate the importance of a word to a file. The importance of a word is directly proportional to the number of times it appears in the file, but at the same time, it is inversely proportional to the frequency it appears in the file library [28]. Based on the principle of TF-IDF, an improved user similarity calculation method is proposed to reduce the weight of the impact of popular items on the user similarity. If an item appears in the user's behavior list, but it also appears many times in other users behavior list, this item is regarded as a popular item, and its impact on the user similarity should be punished. The weight of the *i*-th item in this paper is calculated as:

$$W_i = TF_i \times IDF_i = \frac{freq(i, u)}{|u|} \times \lg \frac{|U|}{(1 + popular(i))} \quad (4)$$

where $freq(i, u)$ represents the number of times that the *i*-th item appears in the behavior list of user *u*; $|u|$ represents the length of behavior list of user *u*; $|U|$ represents the total number of users; and $popular(i)$ represents the number of times that the *i*-th item appears in all of the user behavior lists.

Then the weight of the item is introduced into the equation of Pearson similarity (see Equation (2)), and an improved similarity calculation method is obtained as:

$$Sim_{User}(u, v) = \frac{\sum_{i \in I} (R_{u,i} \times W_i - \bar{R}_u) \times (R_{v,i} \times W_i - \bar{R}_v)}{\sqrt{\sum_{i \in I} (R_{u,i} \times W_i - \bar{R}_u)^2 \sum_{i \in I} (R_{v,i} \times W_i - \bar{R}_v)^2}} \quad (5)$$

3.2. Improved User Characteristics Model

In real life, people living in the same area tend to have similar lifestyle and eating habits, while people in different areas may show greater differences. Similarly, if two people's characteristics are more similar, such as gender, age, and occupation, then their interests are more likely to be similar. For example, there will be more common topics between students, but students and teachers may have different interests due to their different work and social experiences. Therefore, it is reasonable to recommend a user preference item to other users similar to their characteristics when making recommendations. There are some improved collaborative filtering algorithms, which have used the user's characteristics information. However, there are still some problems in the existing method, for example, the similarity of age and occupation is calculated in a crude way, which makes the recommendation results have some limitations [29].

To deal with these problems above, an improved user characteristics similarity model is set up in this paper, which is based on the fuzzy membership method. The proposed user characteristics similarity model can alleviate the cold start problem of the recommendation

system caused by the lack of rating data for new users. The user characteristics similarities in this study are defined as follows:

(1) Age similarity

Suppose that if the age difference is less than 5 years, the similarity is regarded as 1, and if the age difference is more than 25, the similarity is regarded as 0. The fuzzy membership for the age similarity of users is defined as follows:

$$Sim_{Age}(u, v) = \begin{cases} 1, & |Age_u - Age_v| \leq 5 \\ \frac{25 - |Age_u - Age_v|}{20}, & 5 < |Age_u - Age_v| \leq 25 \\ 0, & |Age_u - Age_v| > 25 \end{cases} \quad (6)$$

(2) Occupation similarity

The traditional method of occupation similarity calculation is that: if the occupation is the same, the occupation similarity is set as 1, otherwise it is set as 0. Although it can measure the similarity of two users to a certain extent, the user's characteristics are not fully exploited. In this paper, a tree diagram for the classification of occupations is set up first based on the international standard classification of occupations [30], which is shown in Figure 4.

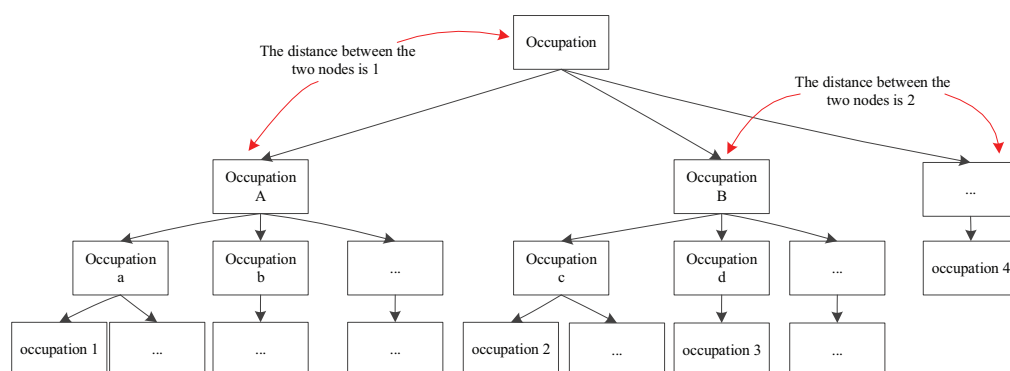


Figure 4. Occupational classification tree.

In this occupation classification tree, the distance between two nodes is defined as the number of edges between these two nodes. The distance between the parent node and child node is 1, and the distance between adjacent brother nodes is 2. The distance between the farthest two occupations in the occupation classification tree is defined as D_{max} . Then the fuzzy membership for the occupation similarity of users is defined as follows:

$$Sim_{Ocp}(u, v) = \begin{cases} 0, & d_{u,v} = D_{max} \\ \frac{1}{1 + \tau d_{u,v}^2}, & 0 < d_{u,v} < D_{max} \\ 1, & d_{u,v} = 0 \end{cases} \quad (7)$$

where $d_{u,v}$ is the occupation distance between users A and B ; and τ is the correction coefficient, which is adjusted dynamically according to the occupation.

(3) Gender similarity

Different gender users have different preferences for items, so the gender should be taken into account when calculating the similarity of user characteristics [29]. Assuming

that the gender of user u is G_u , and the gender of user v is G_v , the gender similarity of users is calculated as:

$$Sim_{Gender}(u, v) = \begin{cases} 0, & G_u \neq G_v \\ 1, & G_u = G_v \end{cases} \quad (8)$$

(4) User characteristics similarity

Combining the mentioned characteristics similarity of users in different dimensions based on the age, gender, and occupation, the final characteristics similarity of users is calculated as:

$$Sim_{Character}(u, v) = \alpha Sim_{Age}(u, v) + \beta Sim_{Gender}(u, v) + \delta Sim_{Ocp}(u, v) \quad (9)$$

where $\alpha + \beta + \delta = 1$ and $\alpha, \beta, \delta \in (0, 1)$ are the similarity weights for the user's age, gender, and occupation. For different recommender systems, these weights can be adjusted dynamically to achieve the optimal recommender effect.

3.3. Proposed Fusion Strategy to Generate Recommendation

Based on the improved similarity calculation method above, the final user comprehensive similarity calculation method can be obtained by weighted fusion, namely:

$$Sim_p(u, v) = \zeta Sim_{User}(u, v) + \mu Sim_{Character}(u, v) \quad (10)$$

where $\zeta + \mu = 1$, and $\zeta, \mu \in (0, 1)$ represent the weights for the similarity obtained based on the TF-IDF method and the user characteristics. For different recommender systems, ζ and μ should be optimized. In this paper, a searching algorithm is proposed to obtain the optimal values of ζ and μ , which is shown in Algorithm 1:

Algorithm 1 Optimal solution search algorithm

```

1:  $\mu = 0; \zeta = 1 - \mu;$  %Initialization parameters
2: for  $i$  in range( $N$ ) :
3:   if  $\mu == 1$  :
4:     break;
5:   else:
6:      $Sim_p = \zeta Sim_a + \mu Sim_b;$  %Model fusion
7:      $MAE(\mu, \zeta); RMSE(\mu, \zeta);$ 
8:      $f(\mu, \zeta) = \min(MAE(\mu, \zeta), RMSE(\mu, \zeta));$ 
9:      $\mu = \mu + 0.5; \zeta = 1 - \mu;$  %Parameter update
10:   end
11: end
12: if  $i \geq N$  or  $\mu == 1$ ;
13:    $\mu, \zeta = f(\mu, \zeta);$ 
14: end
```

After obtaining the user's comprehensive similarity $Sim_p(u, v)$, K users which are most similar to the target user are selected as the nearest neighbors to form the similar neighbor set of the target user. Combined with the rating information of all neighbors and

the similarity with the target user, the target user's rating on the i -th unknown item is predicted. In this study, the rating prediction in (3) is changed to:

$$P_{u,i} = \bar{R}_u + \frac{\sum_{v \in S(u,K) \cap N(i)} \text{Sim}_p(u,v) \times (R_{v,i} - \bar{R}_v)}{\sum_{v \in S(u,K) \cap N(i)} |\text{Sim}_p(u,v)|} \quad (11)$$

The total work processing of the proposed collaborative recommendation algorithm is summarized as follows:

- Step 1: Preprocess the rating data and construct the user-item rating matrix $R(m \times n)$;
- Step 2: Use TF-IDF method and rating data to calculate the user similarity matrix $\text{Sim}_{\text{User}}(u,v)$;
- Step 3: Use user characteristics information to calculate the user characteristics similarity matrix $\text{Sim}_{\text{Character}}(u,v)$;
- Step 4: Fuse the similarity matrices from Step2 and Step3 to generate the final user comprehensive similarity matrix $\text{Sim}_p(u,v)$;
- Step 5: After the comprehensive similarity matrix of a user is obtained, the nearest neighbor set of the target user is selected to make rating prediction and generate recommendations.

4. Experiments

4.1. Dataset and Metrics

To verify the effectiveness of the improved algorithm, this paper uses a dataset from the MovieLens recommender system [31]. The MovieLens dataset is a public movie dataset released by the GroupLens Laboratory of the University of Minnesota. At present, there are eight versions with different sizes. The dataset mainly includes the following information: users ID, items ID, user's rating information of the items, and time stamp of the rating, etc. The MovieLens-100k (ML-100K) data set and MovieLens-1M (ML-1M) data set are used in this paper, and the basic information of the two datasets are shown in Table 1. In the experiment, the dataset is randomly divided into training set and testing set according to the ratio of 8:2 for comparative analysis.

Table 1. The information of the dataset.

Dataset	User	Item	Number of Ratings	Rating Range
ML-100K	943	1682	100,000	1–5
ML-1M	6040	3900	1,000,209	1–5

There are many evaluation indexes of recommender systems [32]. Because the ultimate goal of the improved collaborative filtering recommendation algorithm is to improve the accuracy of the recommendation results, this paper mainly considers the accuracy of the algorithm. To evaluate the recommendation accuracy of the improved recommendation algorithm, the root mean square error (RMSE) and the mean absolute error (MAE) are used to measure the effect of recommended systems [33]. MAE and RMSE are the measurement of the deviation of recommendations from their true user-specified. MAE and RMSE values can be obtained by calculating the rating deviation between the actual rating and the predicted rating between users. The lower the values of RMSE and MAE, the higher the accuracy of the algorithm recommended. The calculation methods for MAE and RMSE are defined as:

$$MAE = \frac{\sum_{i=1}^N |p_{u,i} - r_{u,i}|}{N} \quad (12)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^N (p_{u,i} - r_{u,i})^2}{N}} \quad (13)$$

where N is the total number of rating forecast items in the testing set, $r_{u,i}$ represents the actual rating of user u for the i -th item, and $p_{u,i}$ is the prediction rating of user u for the i -th item.

4.2. Comparison Experiment

To evaluate the performance of the proposed algorithm (ICFTU), the traditional user-based collaborative filtering algorithm (UCF), the collaborative filtering algorithm with user characteristics (CFUC), the collaborative filtering algorithm based on clustering (K-MCF), and the algorithm based on optimizing similarity calculation (ICFOS) [34] are selected for comparison. These four algorithms used for comparison are classical and often used in the recommender systems. The above five algorithms are trained in the training set of two data sets, respectively, and the ratings prediction is carried out in the test set to compare the MAE and RMSE value of different algorithms. The nearest neighbor K is set as 35 for all of the algorithms used in this experiment, the comparison of recommendation accuracy is show in Table 2 and Figure 5.

Table 2. The results of the comparison experiments.

Dataset	MAE					RMSE				
	UCF	CFUC	ICFOS	K-MCF	ICFTU	UCF	CFUC	ICFOS	K-MCF	ICFTU
ML-100k	0.976	0.839	0.806	0.792	0.771	1.156	1.011	1.05	0.982	0.941
ML-1M	0.992	0.852	0.836	0.822	0.801	1.219	1.101	1.102	1.042	0.976

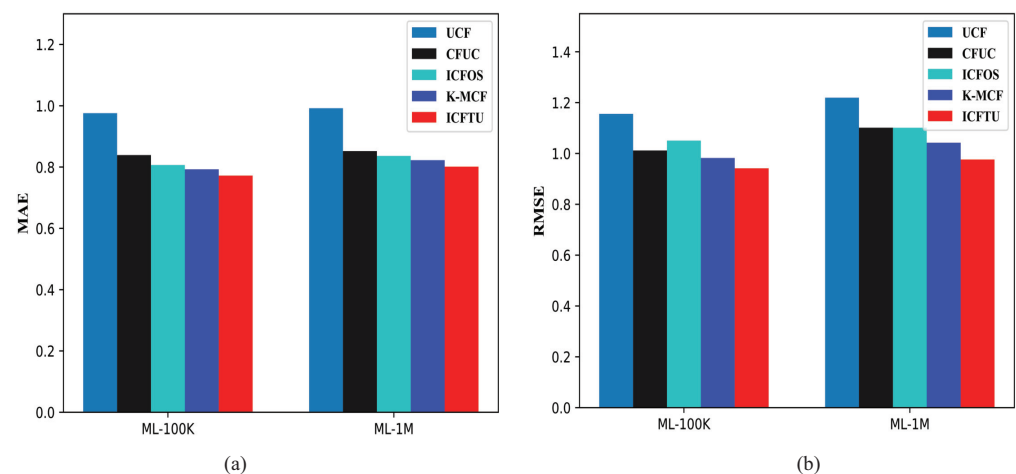


Figure 5. The results of the comparison experiment: (a) of MAE; (b) of RMSE.

It can be seen from the figure that the ICFTU proposed in this paper has a better recommendation effect on both datasets. Among them, the traditional user-based collaborative filtering UCF algorithm has the largest error and the lowest prediction accuracy. The CFUC method (a collaborative filtering algorithm with user characteristics) combines the user's characteristic information, which makes up for some defects of the traditional algorithms and improves the accuracy of recommendation. The K-MCF algorithm based on clustering and ICFOS algorithm based on optimizing similarity calculation both improve the accuracy of recommendation to a certain extent. The ICFTU algorithm combines the TF-IDF method and user characteristics, reduces the impact of popular items on user similarity, improves the calculation of user characteristics similarity, improves the accuracy of recommendation, and still has some advantages in large-scale data sets.

5. Discussions

5.1. Parameter Discussion

In this section, the influence of the parameters involved in the algorithm mentioned in this paper will be discussed. The experiments are carried out on ML-100k data set, where about 20,000 rating data are used to test the influence of the parameters in the proposed algorithm.

(1) About the nearest neighbor

First, the reasonable number of the nearest neighbor K is discussed, which is one of the key factors for the recommendation algorithm to achieve good results. The MAE and RMSE of the proposed method under different K are shown in Figure 6. The results in Figure 6 show that with the increase of the number of nearest neighbors of the target user, the MAE and RMSE of the algorithm show a downward trend and gradually tend to be saturated. Therefore, the nearest neighbor K is set as 35 in this study, to keep a relatively high accuracy and low computation.

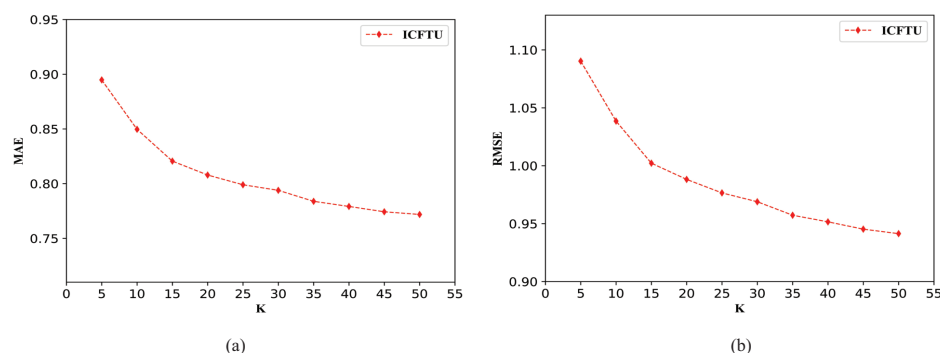


Figure 6. The influence of K on the proposed method: (a) of MAE; (b) of RMSE.

(2) About the user characteristic parameters

Secondly, the user characteristic parameters α , β , and δ are discussed. In this experiment, the parameter adjustment step is set as 0.1. Because $\alpha + \beta + \delta = 1$, to keep all the three parameters bigger than 0, all the three parameters are set between $[0.1, 0.8]$. The MAE and RMSE of the proposed method under different α , β , and δ are shown in Figure 7. It can be seen that the MAE and RMSE are the minimum when $\alpha = 0.5$, $\beta = 0.2$, and $\delta = 0.3$.

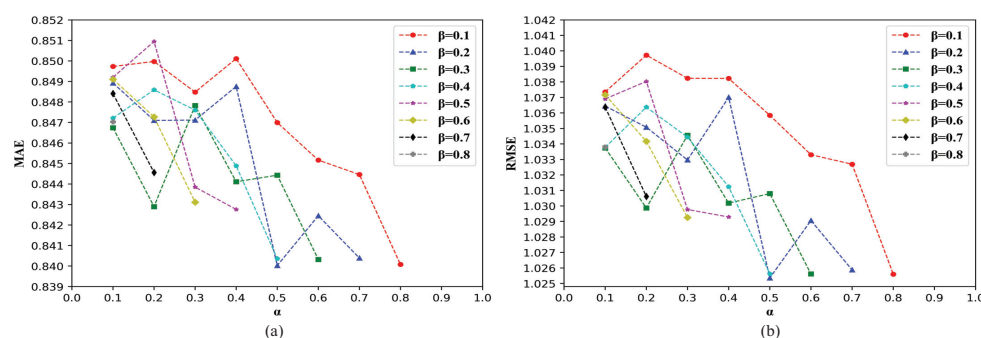


Figure 7. The influence of the user characteristic parameters α , β , and δ on the proposed method. (a) of MAE; (b) of RMSE.

(3) About the model fusion parameters

Thirdly, the similarity weighted fusion parameters ζ and μ are discussed. In this experiment, the parameter adjustment step is set as 0.05. The MAE and RMSE of the proposed method under different ζ and μ are shown in Figure 8. It can be seen that the MAE and RMSE are the minimum when $\zeta = 0.8$ and $\mu = 0.2$. This proves that the user

similarity calculated by the TF-IDF method has the main influence on the recommendation algorithm, and the recommendation effect can be improved by properly fusing the user characteristics similarity.

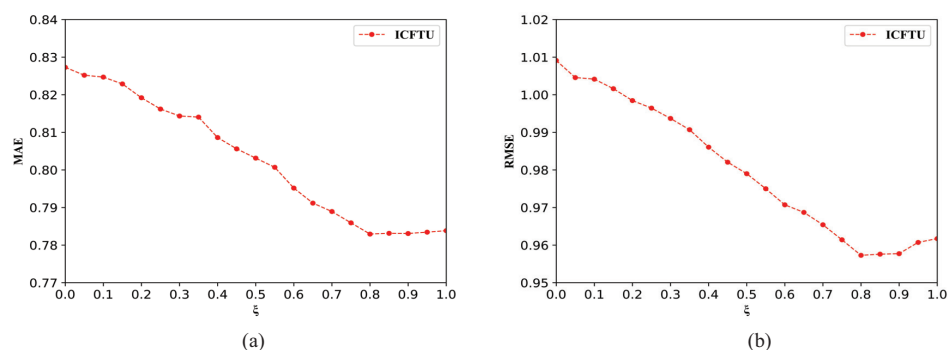


Figure 8. The influence of the model weights ζ and μ on the proposed method: (a) of MAE; (b) of RMSE.

5.2. Ablation Experiment

In this paper, an improved collaborative filtering recommendation algorithm based on the TF-IDF model and user characteristics model is proposed. To discuss the influence of the two main improvements of the proposed method, two ablation experiments are carried out. The experiment is carried out on ML-100k data set, and the results of the proposed method (ICFTU) in Section 4.2 are used as reference. The method which is only based on the TF-IDF model is called ICFTU-TI, and the method which is only based on the proposed user characteristics model is called ICFTU-UC. The results of this ablation experiment are shown in Tables 3 and 4 and Figure 9.

Table 3. The results of the ablation experiment of MAE.

Method	MAE									
	5	10	15	20	25	30	35	40	45	50
ICFTU-UC	0.975	0.916	0.897	0.873	0.859	0.847	0.838	0.827	0.822	0.819
ICFTU-TI	0.894	0.855	0.832	0.817	0.804	0.798	0.792	0.785	0.784	0.779
ICFTU	0.894	0.849	0.82	0.807	0.798	0.793	0.783	0.779	0.774	0.771

Table 4. The results of the ablation experiment of RMSE.

Method	RMSE									
	5	10	15	20	25	30	35	40	45	50
ICFTU-UC	1.174	1.118	1.097	1.069	1.052	1.023	1.01	1.004	1.001	0.999
ICFTU-TI	1.087	1.045	1.015	0.998	0.985	0.976	0.969	0.959	0.957	0.952
ICFTU	1.090	1.038	1.002	0.988	0.976	0.968	0.957	0.951	0.945	0.941

The results show that: (1) The ICFTU algorithm has the best performance and the smallest error, which shows that the method using both the TF-IDF and the improved user characteristics models are effective; (2) The error of ICFTU-TI is smaller than that of ICFTU-UC, and it is close to that of ICFTU. This shows that the TF-IDF method is the main factor to improve the accuracy of the model, and the improved user characteristics are the secondary factor, which is consistent with the discussion results on similarity weighted fusion parameters ζ and μ in Section 5.1.

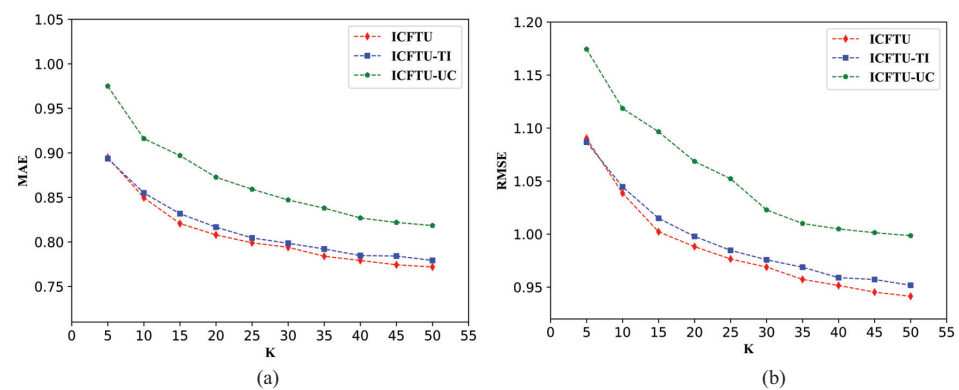


Figure 9. The results of the ablation experiment: (a) of MAE; (b) of RMSE.

5.3. Compared with Other Methods

To further show the performance of the proposed collaborative recommendation algorithm (ICFTU), it is compared with two other state-of-the-art recommendation approaches. The first one is a new collaborative filtering framework based on a gauss core and an extension classification method (known as GCEDA) [26]. The second one is an advanced approach, which is based on Deep Feed-Forward Neural Networks (known as DFFN) [35]. This comparison experiment is carried out on ML-100k data set, and the results are shown in Table 5 and Figure 10.

Table 5. The results of the compared experiment with other two recommendation approaches.

Metric	GCEDA	DFFN	ICFTU
MAE	0.799	0.779	0.771
Improvements	3.50%	1.03%	-
RMSE	0.979	0.959	0.941
Improvements	3.88%	1.88%	-

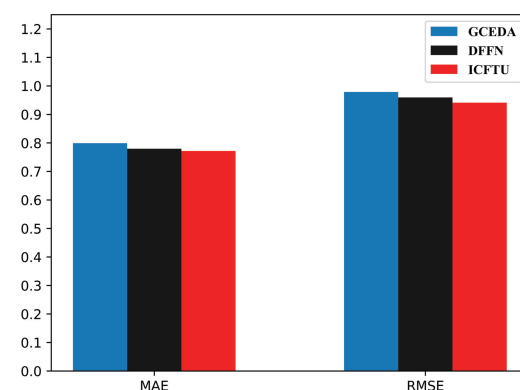


Figure 10. The results of the compared experiment with other two recommendation approaches.

The results show that the proposed ICFTU in this paper has a better recommendation effect compared to GCEDA and DFFN. The results in Table 5 show that the performance of the GCEDA and DFFN methods are close to the K-MCF method (see Tables 2 and 5), the main reason is that all the three methods use the information of the input data by different strategy. However, the comprehensive performance of the proposed model is the best. The MAE value of the proposed model is 3.50% and 1.03% lower than that of GCEDA and DFFN. Meanwhile, the RMSE value of our model is 3.88% and 1.88% lower than that of GCEDA and DFFN.

6. Conclusions

To solve the impact of popular items on user similarity, the TF-IDF statistical method is used in this paper, and optimizes the formula to adapt to the recommendation model. At the same time, an improved user characteristics similarity calculation method is proposed, which makes use of the user characteristics information and alleviates the cold start problem. Finally, this paper conducts off-line experiments on Movielens data sets. Experimental results show that the proposed algorithm is more accurate than the comparison algorithm. There are still some problems that should be further studied in future, such as some new user similarity models by fusing the item tag and user characteristic, and the deep learning technology to mine the potential information of user and item.

Author Contributions: Funding acquisition, J.N.; Project administration, J.N. and Y.X.; Writing—original draft, Y.C.; Writing—review and editing, Y.C. and G.T. All authors have read and agreed to the published version of the manuscript.

Funding: This work has been supported by the National Natural Science Foundation of China (61873086).

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Publicly available datasets were analyzed in this study. This data can be found here: <https://grouplens.org/datasets/movielens>, accessed on 15 April 2021.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Sun, Z.; Guo, Q.; Yang, J.; Fang, H.; Guo, G.; Zhang, J.; Burke, R. Research commentary on recommendations with side information: A survey and research directions. *Electron. Commer. Res. Appl.* **2019**, *37*, 100879. [CrossRef]
2. Quijano-Sanchez, L.; Cantador, I.; Cortes-Cediel, M.E.; Gil, O. Recommender systems for smart cities. *Inf. Syst.* **2020**, *92*, 101545. [CrossRef]
3. Conceicao, F.L.A.; Padua, F.L.C.; Lacerda, A.; Machado, A.C.; Dalip, D.H. Multimodal data fusion framework based on autoencoders for top-N recommender systems. *Appl. Intell.* **2019**, *49*, 3267–3282. [CrossRef]
4. Chen, Y.C.; Hui, L.; Thaipisutikul, T.; Chen, H.L. A Collaborative Filtering Recommendation System with Dynamic Time Decay. *J. Supercomput.* **2021**, *77*, 244–262. [CrossRef]
5. Jiang, M.; Zhang, Z.; Jiang, J.; Wang, Q.; Pei, Z. A collaborative filtering recommendation algorithm based on information theory and bi-clustering. *Neural Comput. Appl.* **2019**, *31*, 8279–8287. [CrossRef]
6. Nakagawa, A.; Ito, T. An implementation of a knowledge recommendation system based on similarity among users' profiles. In Proceedings of the 41st SICE Annual Conference, SICE 2002, Osaka, Japan, 5–7 August 2002; Volume 1, pp. 326–327.
7. Yu, L.; Liu, L.; Li, X. A hybrid collaborative filtering method for multiple-interests and multiple-content recommendation in E-Commerce. *Expert Syst. Appl.* **2005**, *28*, 67–77.
8. Park, Y.; Park, S.; Lee, S.G.; Jung, W. Fast Collaborative Filtering with a k-nearest neighbor graph. In Proceedings of the 2014 International Conference on Big Data and Smart Computing, BIGCOMP 2014, Bangkok, Thailand, 15–17 January 2014; pp. 92–95.
9. Wu, Y.K.; Yao, J.R.; Tang, Z.H.; Meng, J.Q. Collaborative filtering based on multi-level item category system. *J. Conver. Inf. Technol.* **2012**, *7*, 64–71.
10. Bartolini, I.; Zhang, Z.; Papadias, D. Collaborative filtering with personalized skylines. *IEEE Trans. Knowl. Data Eng.* **2011**, *23*, 190–203. [CrossRef]
11. Sajedi-Badashian, A.; Stroulia, E. Vocabulary and time based bug-assignment: A recommender system for open-source projects. *Softw.-Pract. Exp.* **2020**, *50*, 1539–1564. [CrossRef]
12. Pirasteh, P.; Jung, J.J.; Hwang, D. Item-based collaborative filtering with attribute correlation: A case study on movie recommendation. In Proceedings of the 6th Asian Conference on Intelligent Information and Database Systems, Bangkok, Thailand, 7–9 April 2014; Volume 8398, pp. 245–252.
13. Kumar, R.; Verma, B.K.; Rastogi, S.S. Social Popularity based SVD++ Recommender System. *Int. J. Comput. Appl.* **2014**, *87*, 33–37. [CrossRef]
14. Sun, B.; Dong, L. Dynamic Model Adaptive to User Interest Drift Based on Cluster and Nearest Neighbors. *IEEE Access* **2017**, *5*, 1682–1691. [CrossRef]
15. Wang, J.; Lan, Y.-X.; Wu, C.-Y. Survey of Recommendation Based on Collaborative Filtering. In Proceedings of the 2019 3rd International Conference on Electrical, Mechanical and Computer Engineering, ICEMCE 2019, Guiyang, China, 9–11 August 2019; Volume 1314.

16. Zarzour, H.; Jararweh, Y.; Al-Sharif, Z.A. An Effective Model-Based Trust Collaborative Filtering for Explainable Recommendations. In Proceedings of the 2020 11th International Conference on Information and Communication Systems, ICICS 2020, Copenhagen, Denmark, 24–26 August 2020; pp. 238–242.
17. Chen, J.; Wang, B.; Ouyang, Z.; Wang, Z. Dynamic clustering collaborative filtering recommendation algorithm based on double-layer network. *Int. J. Mach. Learn. Cybern.* **2021**, *12*, 1097–1113. [[CrossRef](#)]
18. Chen, Y.; Liu, Y.; Zhao, J.; Zhu, Q. Mobile edge cache strategy based on neural collaborative filtering. *IEEE Access* **2020**, *8*, 18475–18482. [[CrossRef](#)]
19. Deng, J.; Guo, J.; Wang, Y. A Novel K-medoids clustering recommendation algorithm based on probability distribution for collaborative filtering. *Knowl.-Based Syst.* **2019**, *175*, 96–106. [[CrossRef](#)]
20. Zhang, X. Collaborative filtering recommendation algorithm based on sparse bilinear convolution. *Eng. Intell. Syst.* **2020**, *28*, 205–214.
21. Wu, L. Collaborative Filtering Recommendation Algorithm for MOOC Resources Based on Deep Learning. *Complexity* **2021**, *2021*, 5555226.
22. Xiao, Y.; Ai, P.; Hsu, C.H.; Wang, H.; Jiao, X. Time-ordered collaborative filtering for news recommendation. *China Commun.* **2015**, *12*, 53–62. [[CrossRef](#)]
23. Ba, Q.; Li, X.; Bai, Z. A Similarity Calculating Approach Simulated from TF-IDF in Collaborative Filtering Recommendation. In Proceedings of the 2013 Fifth International Conference on Multimedia Information Networking and Security, Washington, DC, USA, 1–3 November 2013.
24. Alhijawi, B.; Kilani, Y. A collaborative filtering recommender system using genetic algorithm. *Inf. Process. Manag.* **2020**, *57*, 102310. [[CrossRef](#)]
25. Li, X.; Li, D. An Improved Collaborative Filtering Recommendation Algorithm and Recommendation Strategy. *Mob. Inf. Syst.* **2019**, *2019*, 3560968. [[CrossRef](#)]
26. Xu, L.; Li, X.; Guo, Y. Gauss-core extension dependent prediction algorithm for collaborative filtering recommendation. *Clust. Comput.* **2019**, *22*, 11501–11511. [[CrossRef](#)]
27. Wang, W.; Chen, J.; Wang, J.; Chen, J.; Liu, J.; Gong, Z. Trust-Enhanced Collaborative Filtering for Personalized Point of Interests Recommendation. *IEEE Trans. Ind. Inform.* **2020**, *16*, 6124–6132. [[CrossRef](#)]
28. Wu, S.; Kou, H.; Lv, C.; Huang, W.; Qi, L.; Wang, H. Service recommendation with high accuracy and diversity. *Wirel. Commun. Mob. Comput.* **2020**, *2020*, 8822992. [[CrossRef](#)]
29. Li, S.; Li, X. Collaborative filtering recommendation algorithm based on user characteristics and user interests. *J. Phys. Conf. Ser.* **2020**, *1616*, 012032. [[CrossRef](#)]
30. Kowal, P.; Chatterji, S.; Naidoo, N.; Biritwum, R.; Wu, F.; Ridaura, R.; Maximova, T.; Arokiasamy, P.; Phaswana-Mafuya, N.; Williams, S.; et al. Data Resource Profile: The World Health Organization Study on global AGEing and adult health (SAGE). *Int. J. Epidemiol.* **2012**, *41*, 1639–1649. [[CrossRef](#)] [[PubMed](#)]
31. Forouzandeh, S.; Berahmand, K.; Rostami, M. Presentation of a recommender system with ensemble learning and graph embedding: A case on MovieLens. *Multimed. Tools Appl.* **2021**, *80*, 7805–7832. [[CrossRef](#)]
32. Tahmasbi, H.; Jalali, M.; Shakeri, H. TSCMF: Temporal and social collective matrix factorization model for recommender systems. *J. Intell. Inf. Syst.* **2021**, *56*, 169–187. [[CrossRef](#)]
33. Velammal, B. Typicality-based collaborative filtering for book recommendation. *Expert Syst.* **2019**, *36*, e12382.
34. Chen, H.; Li, Z.; Hu, W. An improved collaborative recommendation algorithm based on optimized user similarity. *J. Supercomput.* **2016**, *72*, 2565–2578. [[CrossRef](#)]
35. Rizzo, G.L.C.; De Marco, M.; De Rosa, P.; Laura, L. Collaborative Recommendations with Deep Feed-Forward Networks: An Approach to Service Personalization. In Proceedings of the 10th International Conference on Exploring Service Science, Porto, Portugal, 5–7 February 2020; Volume 377, pp. 65–78.