Paper:

Probabilistic Auto-encoder Matrix Factorization Model For Public Digital Cultural Resources Recommendation

Lingjun Meng***, Feng Jin******, Yuqing Hou*******, Wenjuan Zhang******

*School of Automation, Beijing Institute of Technology, Beijing 100081, P. R. China
E-mail: menglingjun@cloudcver.com
**School of Automation, Beijing Institute of Technology, Beijing 100081, P. R. China
E-mail: jinfeng226@163.com
***School of Automation, Beijing Institute of Technology, Beijing 100081, P. R. China
E-mail: 1197192265@qq.com
****School of Automation, Beijing Institute of Technology, Beijing 100081, P. R. China
E-mail: wjzhang056@163.com

Abstract. The results of the public digital cultural resource recommendation model will directly affect the user experience and the popularity of traditional culture. Public digital cultural resources have the problems of information overload and data sparsity. In order to achieve more accurate recommendation, we propose a Probabilistic Auto-encoder Matrix Factorization Model. Firstly, we make a data pre-filling to alleviate the missing data, then we build a new deep collaborative network based on the double hidden layer edge noise auto-encoder to train. The recommendation results integrate the Bayes prior information and the recommendation performance of deep learning. The experimental results show that our model has a good effect on the accuracy and recall rate.

Keywords: Public digital cultural resources, edge noise reduction automatic encoder, deep collaborative recommendation

1. Introduction

Public digital cultural resources are diverse and heterogeneous. With the advent of the era of explosive data growth, in order to optimize the platform and improve user experience, it is an important measure for the construction of public digital cultural resources to introduce accurate recommendation technology into the field of public digital cultural resources [1]. Although accurate recommendation is widely used in e-commerce, film and television industry, tourism and other fields, it has just been applied to the construction of public service system and the relevant research on the recommendation of public cultural service resources for the people of the country has just started [2]. The 19th National Congress also proposed that it is necessary to promote traditional culture and improve people's knowledge of traditional culture.

In recent years, with the increasing demand for deep learning recommendation algorithms, more and more research on deep learning recommendation algorithms has been conducted[3,4,5]. The matrix decomposition-based recommendation algorithm is one of the frontier areas of recommendation algorithm research [6]. The matrix-based decomposition method has many advantages such as high accuracy of recommendation results, good scalability and high flexibility [7]. Deep Collaborative Filter Matrix Decomposition Framework (DCF) [8] is a matrix decomposition recommendation algorithm based on deep learning model, which decomposes the scoring matrix into the product of two low-dimensional hidden semantic features of user and project. We can get the abstract semantic features of the user and the project by the neural network, and get the final score value after matrix multiplication. In addition, Salakhutdinov described matrix decomposition from a probabilistic perspective, and proposed a probabilistic matrix factorization model (PMF) [9]. The PMF model has obtained good predictions on the Netflix dataset. Although these algorithms have greatly improved in recommendation, there are still some problems in solving information overload and the sparsity of data.

This paper proposes a recommend framework Probabilistic Auto-encoder Matrix Factorization Model (PAMF) for effectively alleviating the overload of information and the sparsity of data, we solve the overload of information and the sparsity of data in two ways. On the one hand, we expand the original data, fill the data without value, and alleviate the sparseness of data fundamentally; on the other hand, we propose a collaborative filtering matrix decomposition model based on a double hidden layer noise reduction auto-encoder. Compared with DCF and other models, PAMF has a better ability of feature extraction, and it can deal with the information overload data to a certain extent.

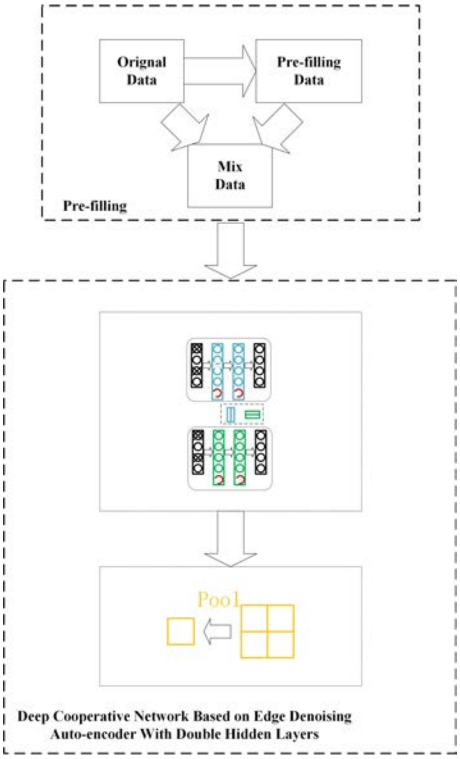
Orignal Data
Pre-filling Data
Mix Data
Pre-filling
Pool
Deep Cooperative Network Based on Edge Denoising Auto-encoder With Double Hidden Layers

Fig. 1. Probabilistic Autoencoder Matrix Factorization Model.

2. Probabilistic Auto-encoder Matrix Factorization

Public digital cultural resources have problems such as information overload and sparse data. These problems have made it difficult to accurately recommend public digital cultural resources. This paper presents a probabilistic self-encoding decomposition model that effectively alleviates this problem. The overall framework is shown in **Fig. 1**. PAMF consists of two parts: data pre-filling operation and deep collaborative network based on double hidden layer edge noise reduction auto-encoder. Because the original data is relatively sparse, it is not conducive to the training of deep learning networks, so we make an effective expansion of the original data through data pre-filling firstly. In the network model part, we build a new network based on the classic DCF framework, The feature extraction ability of network become stronger than original DCF model, and get a better accuracy of recommendation results.

2.1. Data Pre-Filling Based On Probability Matrix Decomposition

Probability matrix factorization model is a recommendation model based on matrix factorization and Bayes theory. The model uses probability and matrix factorization to fill in the missing scores in the original matrix to achieve recommendation, the model shows a good result when the data is sparse. so we first uses the probability matrix decomposition model to pre-fill the original data to obtain a pre-filled scoring matrix. To a certain extent, the original information was restored. However, public digital cultural resources also have the characteristic of information overload. There are a large amount of prior information when relying on the result of a single probability matrix decomposition model, And too much prior information lead to the problem of excessive noise interference, which makes the final recommendation result not very good. So in this paper, we build a deep learning model to make a prior correction.

2.2. Deep Collaborative Network Based On Double Hidden Layer Auto-encoder

Based on the data pre-filling operation based on probability matrix decomposition, the original data is filled with certain missing values. The combined data is denser than the original data and is more suitable for deep-learning-based recommendation algorithms. In order to better extract the hidden semantic features of the data, on the basis of the classic deep collaborative filtering matrix decomposition framework, we changed the type of auto-encoder and the number of hidden layers, and introduced a pooling processing mechanism, then we proposed a deep cooperative network based on double hidden layer edge noise reduction auto-encoder.

The deep collaborative filtering matrix decomposition framework is a hybrid matrix decomposition model framework based on deep learning models. The model framework is shown in **Fig. 2**. The deep collaborative filtering matrix decomposition framework constructs two net include encoder and decoder for user features and project features respectively. The net has a hidden layer. The model uses deep learning feature extraction capabilities to extract k-dimensional features U and V for users and projects. A hidden matrix is obtained by multiplying U and V.

The hidden layer of the DCF framework is a single hidden layer, and the expression ability of the model is limited. Considering that the public digital culture has the problem of the overload of information, we set the single hidden layer of the self-encoder to two hidden layers with the same number of neurons to improve the expression of the model. Furthermore, the encoder used in the DCF framework is an ordinary self-encoder, which has a poor ability to prevent over fitting. In this paper, we use the edge noise reduction automatic encoder [8] as the hidden layer in the model instead of the ordinary self-encoder. The double hidden layer edge noise reduction auto-encoder is shown in **Fig. 3**. The edge noise reduction auto-encoder introduces random noise to improve the generalization ability of the model and alleviate the phenomenon of over fitting.

After introducing the double hidden layer edge noise reduction auto-encoder, a deep collaborative network based on the double hidden layer edge noise reduction auto-encoder is obtained, as shown in **Fig. 4**. In the matrix decomposition network based on the double hidden layer edge noise reduction auto-encoder, the two hidden

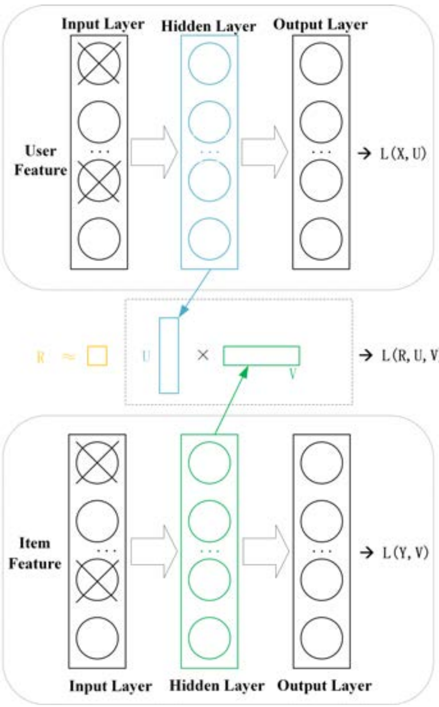

Fig. 2. Deep Collaborative Filtering Matrix Decomposition Framework.

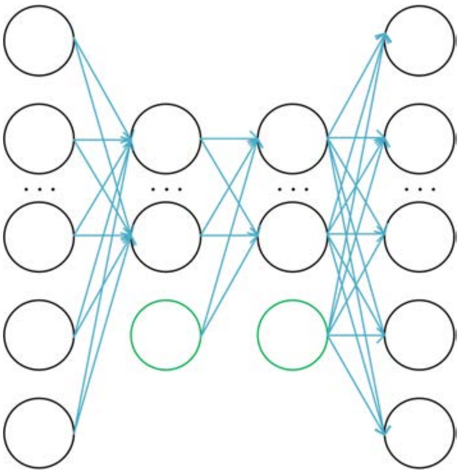

Fig. 3. Double Hidden Layers Edge Denoising Auto-encoder Network.

layer features are combined to form a 2 × n user feature and a n × 2 project feature, and we can get a 2 × 2 matrix. In the recommendation system network, one score is generally obtained to represent the user's scoring situation for the project. Then, we use a global pooling method in the convolutional neural network [9] to calculate the final score. Average pooling and maximum pooling are two classic forms of the convolutional network pooling, we use global average pooling to comprehensively consider the four characteristic results, which represent the final scoring results, and we will prove the effectiveness of average pooling in the ablation study.

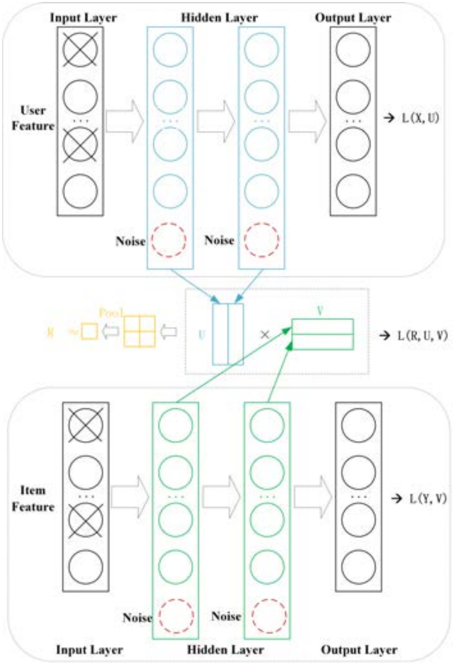

Fig. 4. Deep Cooperative Network Based on Edge Denoising Auto-encoder With Double Hidden Layers.

2.3. Probabilistic Auto-encoder Matrix Factorization Model

The probabilistic Auto-encoder decomposition model relieves the data sparsity problem by pre-filling the data, and improves the accuracy of the scoring matrix by a deep collaborative network based on double hidden layer edge noise reduction autoencoders. The algorithm flow is as follows:

(1) Use the probability matrix decomposition algorithm to restore the original data and mix it with the original data;

(2) Use pre-populated mixed data to train the deep collaborative network based on the double hidden layer edge noise reduction autoencoder;

(3) Obtain the result of the scoring matrix, and select the top N items in the prediction score of each user for recommendation.

3. Experiment and Ablation Study

3.1. Datasets and Evaluation Indicators

In order to verify the effectiveness of the PAMF, we choose two most common datasets MovieLens 20M and Netflix datasets.

The MovieLens dataset is an open source recommendation algorithm research dataset from the GroupLens laboratory of the University of Minnesota. The MovieLens dataset is divided into 100K, 1M, 10M and other datasets of different sizes according to the amount of data. Each dataset contains user score files, movie data files, and user data files. The largest data set in the MovieLens dataset includes about 140,000 users' scoring data for 27,000 movies, with a total of about 200,000 ratings. The MovieLens dataset is widely used in the recommendation field, because it is open source and it has the ability to meet different tasks. The Netflix dataset comes from the recommendation algorithm contest organized by Netflix in 2005, which includes 1 billion pieces of score information of nearly 20,000 movies by more than 400,000 users.

The evaluation indicators of recommendation algorithm [10] include accuracy and recall rate. We choose the root mean square error (RMSE) between the model predicted score and the actual score of the user as the evaluation index of accuracy, RMSE defined as

$$RMSE = \sqrt{\frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n} 1(\widetilde{R_{ij}} - R_{ij})^{2}} \quad . \; . \; . \; . \quad (1)$$

where m denotes the number of users, n denotes the number of projects,$\widetilde{R_{ij}}$denotes the model's prediction of user i's scoring results for item j,R_{ij} denotes the score result of actual user i for item j.

For the recall rate, we select the top N result as the recommendation result, and use the top N result in the actual scoring result as the user's favorite item to calculate the recall rate of the model. $Recall@N$ is defined as

$$Recall@N = \sum_{i=1}^{m} \frac{F_{iN}}{N} \quad . \; . \; . \; . \; . \; . \; . \; . \; . \; . \quad (2)$$

where F_{iN} denotes the number of items belonging to the user's favorite among the top N results in the user's prediction.

3.2. Comparative Experiment

In order to verify the recommended effect of the PAMF model, We make a comparative experiment with DCF and DCF+PMF. Then, the number of neurons in the hidden layer needs to be determined before conducting verification experiments, so we first determine the dimension K of the hidden layer through a set of experiments.

3.2.1. Determination of the dimension K of the hidden vector

K value is the number of hidden layer neurons of matrix decomposition and auto-encoder, which represents the number of features in score matrix decomposition. The selection of K value has a greater influence on the final effect of the model than other hyper parameters, so we make two experiments on the two data sets to determine the K value, and the remaining hyper parameters were selected by K-fold cross-validation through the verification set. The results of RMSE experiment with different K values on Movielens dataset and Netflix dataset are shown in **Fig. 5**. The experimental results of the recall rate under different K values are shown in **Fig. 6**.

As can be seen from the RMSE results, when the K value is in the range of 50 to 60, the RMSE value is smaller on the MovieLens data set and the Netflix data set; from the recall result, when the K value is around 55, which can have a good result each datasets. Combining the experimental results of RMSE and recall@N, the K value is set to 55.

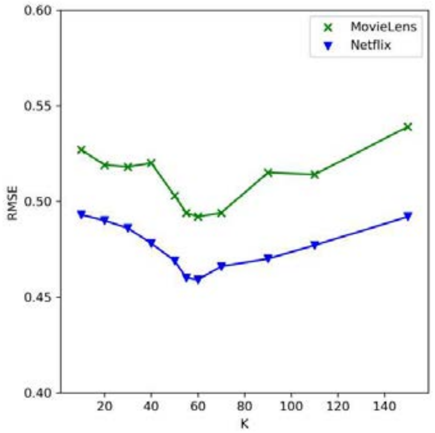

Fig. 5. RMSE results at different K values.

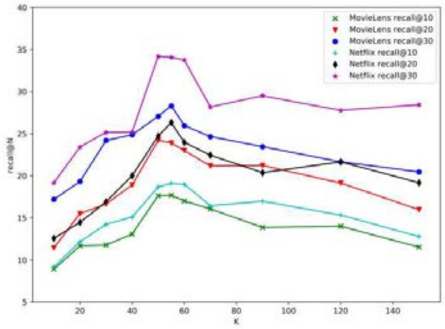

Fig. 6. Recall result at different K values.

3.2.2. Contrast model experiment

In order to evaluate the effectiveness of the probabilistic self-encoding decomposition model (PAMF) proposed in this paper, we use the deep collaborative decomposition model (DCF) and the fusion of the probabilistic decomposition model (PMF) and the deep collaborative matrix decomposition model (DCF) as a comparison model. The fusion model of PMF and DCF is uesd to prove that the probability matrix factorization pre-filling is better than the model fusion. The accuracy experiment results on the MovieLens dataset are shown in **Fig. 7**, and the accuracy experiment results on the Netflix dataset are shown in **Fig. 8**, where the abscissa is the number of iterations for model training. The recall rate experiment results of MovieLens and Netflix recall rate experiment results are shown in **Fig. 9**.

The following conclusions can be drawn from the experimental results:

(1) It can be seen from the results of **Fig. 7** and **Fig. 8** that the probabilistic self-encoding decomposition model proposed in this paper has a RMSE lower 0.046 than

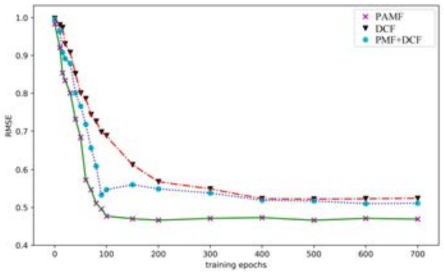

Fig. 7. MovieLens comparison experiment RMSE results.

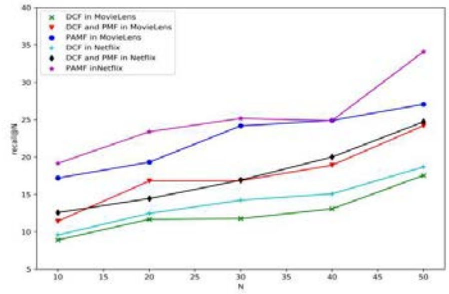

Fig. 9. Recall results on both data sets.

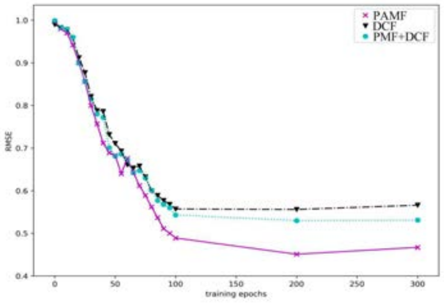

Fig. 8. Netflix comparison experiment RMSE results.

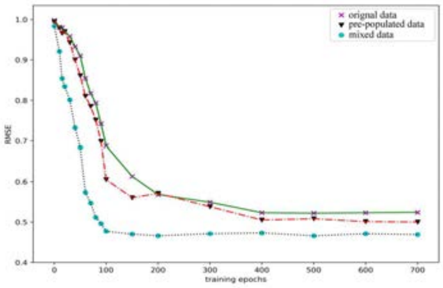

Fig. 10. Impact of PMF pre-reduction on accuracy.

PMPM+DCF (when epoch is 700 stable) and 0.060 lower than DCF (epoch 700 stable). PAMF's RMSE on the Netflix dataset is 0.107 lower than DCF (300 for epoch) and 0.067 lower than PMF+DCF (300 for epoch); We can learned that the accuracy of the PAMF model is significantly better than the other two models.

(2) From **Fig. 7** and **Fig. 8** We can see that PAMF is also superior to the single DCF model and the fusion model in the convergence speed.

(3) **Fig. 9** shows us the recall rate results, And we can clearly know the probabilistic self-encoding decomposition model proposed in this paper has better results on the two data sets than the other two models.

3.3. Ablation Study

In order to illustrate the role of each module, we designed some ablation experiments. Firstly, Using the original data, pre-filled data, and mixed data to train our model, then to show the effect of the pre-filled data; And we also make an experiment to change the structure of the pooling layer to verify the effect of the average pooling layer.

3.3.1. Data Pre-filled

The probabilistic self-encoding decomposition model uses the combination of the data restored by PMF and the original data as the input data of the model. In order to evaluate the impact of the data pre-filling on the deep collaborative network based on the double hidden layer auto-encoder, We conduct a set of experiments and set the model input to raw data, PMF data and combined data, and the dataset is the Movielens dataset The results of the RMSE accuracy results are shown in **Fig. 10**. The recall experiment results are shown in **Fig. 11**.

It can be seen from the experimental results:

(1) The experimental results of the combined data are 0.063 lower than the original data, and 0.037 lower than the PMF reduction data. It can be seen that the combined data has a better effect on the model than using any arbitrary data. PMF pre-filled data effectively supplements the original data and improves the fitting effect of the model.

(2) From the recall results, the recall rate of the combined data is also significantly higher than that of the original data and pre-populated data, further verifying that the combined data does play a role in improving the sparsity of the data, which can alleviate the learning of the current model to a certain extent.

3.3.2. Average Pooling

PAMF uses an average pooling mechanism to calculate the scoring results. In order to verify that the average pooling process has a better effect on model optimization than the maximum pooling process, we make a experiments by the maximum class pooling process. The same model is used as a comparison model, and the dataset is the Netflix dataset. the accuracy result of different pooling processing mechanisms are shown in **Fig. 12**. The results of the recall experiment are shown in **Fig. 13**.

The experimental results show that the model with the average pooling layer has a lower RMSE value of 0.04 than the maximum pooling model, and the result of the recall rate is also significantly improved. It can be seen

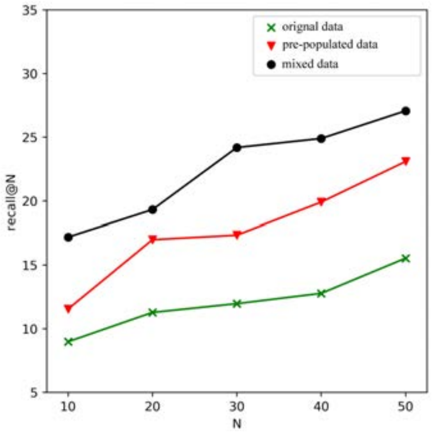

Fig. 11. Impact of PMF pre-reduction on recall rate.

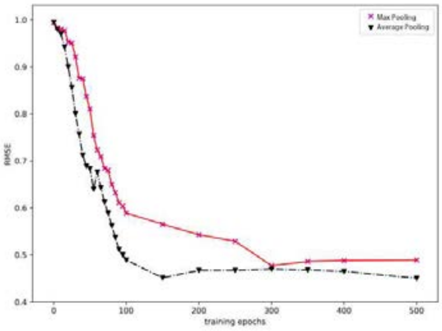

Fig. 12. The impact of like-pooling on accuracy.

that the average pooling effect is better than the maximum pooling effect.

4. Conclusion

In order to solve the problem of information overload and data sparseness in the field of public digital cultural resources, we propose a deep collaborative network based on double hidden layer edge noise reduction autoencoder. Compared with DCF and other models, it has a better feature extraction capability and a good ability to resist over-fitting. For the problem of data sparsity, we propose to first extend the data through a probabilistic autoencoder and then train the model with mixed data to effectively supplement the dataset and improve the final recommendation effect; Combining these two ideas, we propose a new model probabilistic auto-encoder matrix factorization model. The PAMF model combines Bayesian prior information and feature extraction capabilities of deep learning to alleviate the problems of data sparsity and insufficient data to a certain extent.

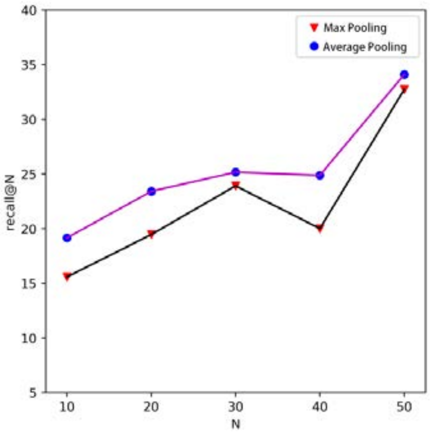

Fig. 13. The impact of like-pooling on recall rate.

Acknowledgements

I would like to thank my supervisor and my senior fellow apprentice for taking time to review this work. In addition, I would like to thank my younger brother and sister for helping me to verify the experiment and the paper.

References:

[1] Luo Yunchuan, Li Tong. Analysis on the governance sharing strategy of public cultural resources [J]. Library work and research, 2016, (04): 28-32.

[2] Xiang Liang. Practice of recommendation system [M]. Beijing: Posts and Telecommunications Press, 2012:20-40.

[3] Guo Yunfei, Fang yaoning, Hu Hongchao. Recommendation algorithm of social matrix decomposition based on logistic function [J]. Journal of Beijing University of technology, 2016, 36 (1): 70-74.

[4] Nilashi M , Bagherifard K , Ibrahim O , et al. Collaborative Filtering Recommender Systems[J]. Research Journal of Applied ences Engineering Technology, 2013, 5(16):4168-4182.

[5] Wang Peng. Research on recommendation system algorithm based on matrix decomposition[D]. Beijing Jiaotong University, 2015.

[6] Goldberg, David, Nichols, David, Oki, Brian M. Using collaborative filtering to weave an information tapestry[J]. Communications of the ACM, 1992, 35(12):61-70.

[7] Afoudi Y , Lazaar M , Achhab M A . Collaborative Filtering Recommender System[C]// International Conference on Advanced Intelligent Systems for Sustainable Development. Springer, Cham, 2018:332-345.

[8] Sheng Li, Jaya Kawale, Yun Fu. Deep Collaborative Filtering via a Marginalized Denoising Auto-encoder. Proceedings of the 24th ACM International on Conference on Information and Knowledge-Management[C].ACM,2015: 811-820.

[9] Salakhutdinov R , Mnih A . Probabilistic matrix factorization[J]. Advances in neural information processing systems, 2008:1257-1264.

[10] Minmin Chen, Zhixiang Xu, Kilan Q. Marginalized Denoising Autoencoders for Domain Adaptation. Prcoceedings of the 29th International Conference on Machine Learning[C], Edinburgh, Scotland,UK, 2012:1-8.

[11] M. Lin, Q. Chen, and S. Yan. Network in network. arXiv:1312.4400,2013

[12] Liu Jianguo, Zhou Tao, Guo Qiang, et al. Review of evaluation methods of personalized recommendation system [J]. Complex system and complexity science, 2009, 6 (3): 1-10