# Movie Recommendation using Unrated Data

Dong Nie, Lingzi Hong, Tingshao Zhu*
Institute of Psychology, University of Chinese Academy of Sciences
Chinese Academy of Sciences, Beijing China 100190
Email: tszhu@psych.ac.cn

***Abstract*—Model based movie recommender systems have been thoroughly investigated in the past few years, and they rely on rating data. In this paper, we take into account unrated data of genre information to improve the performance of movie recommendation. We propose a novel method to measure users' preference on movie genres, and use Pearson Correlation Coefficient (PCC) to compute the user similarity. A matrix factorization framework is introduced for genre preference regularization. Experimental results on MovieLens data set demonstrate that the approach performs well. Our method can also be used to increase the genre diversity of recommendations to some extent.**

## I. Introduction

Nowadays, with the explosion of information on the internet, it becomes more and more difficult to locate information needed. Recommendation is one of the most widely-used technique to help web users, and many computer scientists have been devoted to the technique in the past few years. The development of recommender systems have brought great convenience for internet users and considerable commercial value.

Collaborative filtering (CF) [6] has been thoroughly investigated these years, and becomes one of the most commonly used recommendation approaches. There are mainly two types of CF systems: memory-based and model-based [2]. Memory-based approaches rely on the rating database to find similar users or items [8]. While model-based approaches use the observed user-item rating matrix to train a model to explain the observed data. There are many model-based collaborative filtering algorithms, including Bayesian networks, clustering models, aspect models, latent factor models, ranking models and markov decision process based models [14].Among the model-based approaches, matrix approximation technique [3], [13], [7] has shown great promise, since it can find latent variables through matrix factorization. [3] involves a matrix factorization which constructs a feature matrix for both users and objects. [13] presents probability matrix factorization with Gaussian observation noise. These methods focus on factorizing the user-item rating matrix to produce latent variables, and employ the inner product of latent variables to give further predictions. SVD++ [7] makes use of implicit feedback information, and implicit feedback can refer to any kinds of users' history information that can help indicate users' preference.

A lots of algorithms are developed to improve the accuracy of recommendations, while the quality of recommendations can be evaluated through a myriad of dimensions, but obviously relying merely on the accuracy of recommendations may not be enough to find the most relevant movies for each user. For example, the systems usually recommend popular movies while individual's preferred movies are submerged in the crowd popular movies (often known as "long tail" phenomenon). In this case, the importance of diverse recommendations should be emphasized. Therefore, apart from accuracy, we wish our system performs as diverse as possible. Recently some researchers pay high attention to diverse recommendations [4]. These studies proposed new recommendation methods that can increase the diversity of recommendation sets for a given individual user, often measured by an average dissimilarity between all pairs of recommended items, while maintaining an acceptable level of accuracy. These techniques aim to avoid providing too similar recommendations for the same user, thus the loss of accuracy is inevitable though many other approaches are taken to compensate it [9]. In our experiment, we try to reconcile the problem. With extra genre preference information, we tend to improve improve the predictive accuracy and the genre diversity at the same time.

In this paper, we present a novel method to enhance movie recommendation's accuracy as well as diversity, with supposing that movie genre is a key factor of the user's preference, and users with high similarity in genre preference are more likely to watch similar movies. More specifically, we propose a Movie Frequency-Inverted Genre Frequency (MF-IGF) method to measure a user's movie genre preference on watched movies, using Pearson Correlation Coefficient (PCC) [12] method, we properly compute the similarity between users. A matrix factorization framework with genre preference regularization has been developed to incorporate watched movies' information. Experimental results on large data set (Movielens) demonstrate our approach outperforms singular value decomposition (SVD) [11] method.

The rest of this paper is organized as follows. Section II describes how to compute genre preference and matrix factorization with genre preference regularization framework. The experimental results and discussion are presented in Section III, followed by the conclusion and future work in Section IV.

## II. Genre Preference Regularization Model

Given a user set $U = \{u_1, u_2...u_n\}$, a movie set $V = \{v_1, v_2...v_m\}$, and the watched movie list for a user $u_i$ is $\{v_{i_1}, v_{i_2}...v_{i_l}, v_{i_{l+1}}...v_{i_p}\} \subset V$, where the first $l$ movies are rated with a scale from 1 to 5, while the remainder is only watched without rating process by $u_i$. The rating matrix is represented by $R_{n\times m}$, and the movie genres can be classified as $L = \{1, ...., g\}$, and each movie can be categorized to three genres at most.

The traditional matrix approximation approaches only deal with the rated movies from users, and our problem is to

make use of watched but unrated movies. We first apply a movie frequency-inverted genre frequency method to compute the genre preference for a user with their available watched movies, PCC is employed to compute similarity between users based their genre preferences, and a regularization term with genre preference will be added to traditional matrix factorization framework.

### A. Movie Frequency-Inverted Genre Frequency (MF-IGF)

Currently, quite a few recommender systems are built based on TF-IDF [5]. We present a similar style to compute a user's genre preference for movies. Now we take the calculation of $u_i's$ preference on movie genre $c$ as an example, and here we suppose the number of $u_i's$ watched movies is $p$.

$$MF - IGF = MF * IGF \tag{1}$$

where

$$MF = \frac{\left|\{v_{i_j} | genre\left(v_{i_j}\right) \in c, j <= p\}\right|}{\left|\{v_{i_j} | j = 1, ..., p\}\right|} \tag{2}$$

and

$$IGF = \log\left(\frac{|V|}{|\{v_j | genre(v_j) \in c, j = 1, ..., m\}|}\right) \tag{3}$$

The above MF-IGF formula is the basic form, we can derive other forms like TF-IDF. The preference for a genre increases direct proportionally to the number of movies of a specific genre appears in the watched list of a user, but inverse proportionally with the frequency of the specific movie genre in the whole movie set. Hence, if a user's favorite movie genre is sparse in the whole movie market, i.e., Werewolf Movie, our MF-IGF metric can well identify the user's real preference.

### B. Genre Preference Regularization

Though there are many metric functions to measure similarity, PCC considers different users may have different genre preference styles. Hence, PCC is proposed to solve our problem:

$$W_{ij} = \frac{\sum_{c=1}^{g}\left(mfigf_{(u_i,c)} - \overline{mfigf_{u_i}}\right)\left(mfigf_{(u_j,c)} - \overline{mfigf_{u_j}}\right)}{\sqrt{\sum_{c=1}^{g}\left(mfigf_{(u_i,c)} - \overline{mfigf_{u_i}}\right)}\sqrt{\sum_{c=1}^{g}\left(mfigf_{(u_j,c)} - \overline{mfigf_{u_j}}\right)}} \tag{4}$$

where $mfigf$ is short for MF-IDF, $\overline{mfigf_{u_i}}$ denotes $u_i's$ average genre preference for the whole movie genres. From the above definition, user similarity $W_{ij}$ is ranging from $[-1, 1]$, and a larger value means users i and j have more genre preference in common.

Traditionally, the low-rank matrix factorization approach seeks to approximate the rating matrix $R$ by a multiplication of l-rank factors :

$$R \approx U^T V \tag{5}$$

And the Singular Value Decomposition (SVD) method is employed to realize the process, to avoid overfitting, regularization on user and item matrix is made:

$$\begin{aligned} E = \min_{U,V} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} I_{ij}\left(R_{ij} - {U_i}^T V_j\right)^2 \\ + \frac{k_n}{2}\sum_{i=1}^{n}\|U_i\|_F^2 + \frac{k_m}{2}\sum_{j=1}^{m}\|V_j\|_F^2 \end{aligned} \tag{6}$$

where $I_{ij}$ is the indicator function that equals to 1 if user $u_i$ rated item $v_j$ and equals to 0 otherwise, and the regularization coefficients $k_n, k_m > 0$.

As shown in Section I, genre preference may be helpful to improve movie recommendation, since genre preference is likely to be the key factor of movie taste in real world, for example, people who love Werewolf Movie probably take great interest in $< Underworld >$ series. Based on the intuition, we propose our genre preference regularization model based on matrix factorization technique:

$$\begin{aligned} E = \min_{U,V} \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} I_{ij}\left(R_{ij} - {U_i}^T V_j\right)^2 + \frac{k_n}{2}\sum_{i=1}^{n}\|U_i\|_F^2 + \\ \frac{k_m}{2}\sum_{j=1}^{m}\|V_j\|_F^2 + \frac{\alpha}{2}\sum_{i=1}^{n}\sum_{s\in knn(i)} W_{is}\|U_i - U_s\|^2 \end{aligned} \tag{7}$$

where $\alpha > 0$, and $knn(i)$ is the set of k nearest neighbors to user $u_i$ on genre preference. Here, we only take the most k-neighbor users into account.

A local minimum of the above objective function can be found by performing gradient descent in feature vectors $U_i$ and $V_j$ :

$$\begin{aligned} -\frac{\partial E}{\partial U_i} = \sum_{j=1}^{m} I_{ij}\left(R_{ij} - {U_i}^T V_j\right)V_j \\ -k_u U_i - \alpha\sum_{s\in knn(i)} W_{is}\left(U_i - U_s\right) \end{aligned} \tag{8}$$

and

$$-\frac{\partial E}{\partial V_j} = \sum_{i=1}^{n} I_{ij}\left(R_{ij} - {U_i}^T V_j\right)U_i - k_m V_j \tag{9}$$

The genre preference regularization term in the above objective function is

$$\frac{\alpha}{2}\sum_{i=1}^{n}\sum_{s\in knn(i)} W_{is}\|U_i - U_s\|^2 \tag{10}$$

It is imposed to minimize the movie tastes between user $u_i$ and $u_i's$ k nearest neighbors with considering genre preference. Specifically, $u_i's$ movie taste $U_i$ (feature vector) should be close to the average tastes of all the k-nearest neighbors.

## III. Experimental Results

Our experiments are conducted on MovieLens data set [10]. The data set contains 6040 anonymous users, 3952 movies, and a total of 1,000,209 ratings. Each user rates at least 20 movies, each movie belongs to at least one genre and at most three genres. There are 18 movie genres in this data set, including

Action, Adventure, Animation, Children's, Comedy, Crime, Documentary, Drama, Fantasy, Film-Noir, Horror, Musical, Mystery, Romance, Sci-Fi, Thriller, War, Western respectively.

As we want to test our model on different situations, we randomly sample data set according to the training set proportion. The data set is divided into two disjointed training and test sets. The test set is seen as watched but unrated movies. The performances are measured by RMSE:

$$RMSE = \sqrt{\frac{1}{|T|}\sum_{i,j}\left(R_{ij} - \widehat{R_{ij}}\right)^2} \tag{11}$$

where $\widehat{R_{ij}}$ represents the rating user $u_i$ gave to movie $v_j$ as predicted by our method, and $T$ represents the test sets.

### A. Results

We implement genre preference regularization model in python, and use MATLAB for data preprocessing and post-processing as it is convenient for matrix manipulation. The baseline algorithms we have employed are as follows:

1) AVGB: $\widehat{R_{ij}}$ for user i and movie j as $\overline{R_j} + b_i$, where $\overline{R_j}$ is the mean score of object $j$ in training data, and $b_i$ is the average bias of user i computed as:

$$b_i = \frac{\sum_{j=1}^{m} I_{ij}\left(R_{ij} - \overline{R_j}\right)}{\sum_{j=1}^{m} I_{ij}} \tag{12}$$

2) SVD: Basic SVD with regularization on user and movie feature factors, which has been stated in Equation 6.
3) CSVD: A compound SVD algorithm which incorporates per-user and per-object biases and the constraints on feature vectors [11].

Every experiment is repeated by 5 times, and the results are listed in Table I,

TABLE I. RMSE ACHIEVED BY EACH ALGORITHM

| Training Set Proportion | AVGB | SVD | CSVD | GPR |
|---|---|---|---|---|
| 95% | 0.9094 | 0.8497 | 0.8491 | 0.8485 |
| 80% | 0.9101 | 0.8524 | 0.8517 | 0.8498 |
| 60% | 0.9105 | 0.8623 | 0.8617 | 0.8563 |
| 40% | 0.9147 | 0.8795 | 0.8790 | 0.8702 |
| 20% | 0.9261 | 0.9096 | 0.9091 | 0.8976 |

"GPR" refers to Genre Preference Regularization method, with $k = 5$ and $\alpha$=0.01 respectively, the detail about parameter decision will be discussed in Section III-B. The number of factors is set at 10 by grid search for all SVD related methods in this paper. As shown in Table I, matrix factorization incorporating genre preference information performs better than three baseline algorithms. Moreover, with the decrease of the training set proportion, the baseline methods perform worse, while GPR performs very smoothly and quite stable. Obviously, with the increase of proportion of test set, more unrated movie information is incorporated into movie taste decision, thus the decrease of rating information is partly counteracted, and that's why GPR exceeds other approaches more on small training set.

### B. Impact of parameters $\alpha$ and $k$

In GPR model, there are two important parameters, $\alpha$ and $k$. $\alpha$ directly controls the degree of incorporating the unrated movies. Figure 1 shows the impacts of $\alpha$ on RMSE.

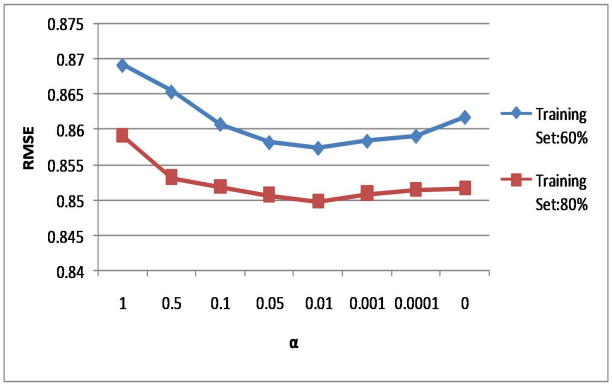


Fig. 1. The impact of $\alpha$

We find that $\alpha$ impacts the results significantly, which demonstrates that it can improve the accuracy of recommendations by incorporating user's genre preference information.

Another parameter, $k$, decides how many nearest neighbors are involved. The extreme situation is that all the other users or no users participate in the regularization term. Figure 2 depicts how $k$ impacts RMSE.

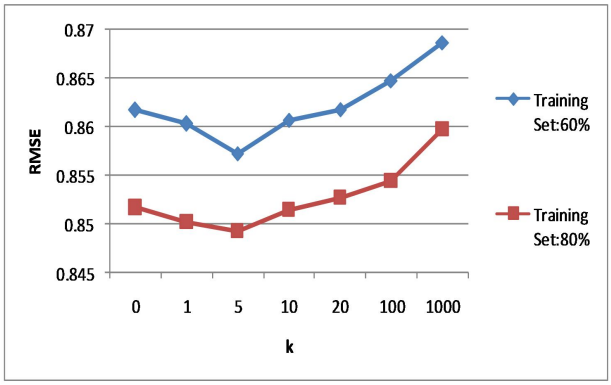


Fig. 2. The impact of parameter k

We can see from Fig. 2 that $k$ influences the results indeed, and $k = 5$ is the appropriate setting.

### C. Discussion about Recommendation Diversity

As one of the core issues of recommendation, diversity has been largely investigated [1]. The traditional SVD methods for movie recommendation have little about diversity. In this paper, a Movie Frequency-Inverse Genre Frequency method is proposed to measure a user's preference over a movie. This method has one more benefit, i.e., to increase genre diversity. As genre is considered as a key factor to decide a people's preference for a movie, our methods make the recommendation more balanced and efficient.

In this paper, we focus on genre diversity for SVD and GPR methods. We conduct an user study with 1000 users randomly

chosen. We compute the ratings between users and movies using both methods respectively, and then sort the predicted ratings for each user. We choose top $N$ movies which have the biggest predicted ratings, and take these movies to be the recommended movies for the user.

The approaches stated in Section **??** measure the diversity by an average dissimilarity between all pairs of recommended items. Since we aim to enhance genre diversity, we define a new way to be the genre diversity metric instead of dissimilarity.

The genres of the recommended movies are observed, and we calculate the number of distinct genres of the $N$ recommended movies. The number of distinct genres are defined as the recommendation diversity for a user:

$$diversity = |\{\text{genres for movie } m | m \in \text{the } N \text{ movies}\}|$$

In this experiment, $N = 15$, and all other settings are the same. To compare the two methods clearly, we do a subtraction of the recommendation diversity between GPR and SVD, and the distribution of the difference is described in Figure 3. Note that, the data is sorted by the difference, to make them easier to follow.

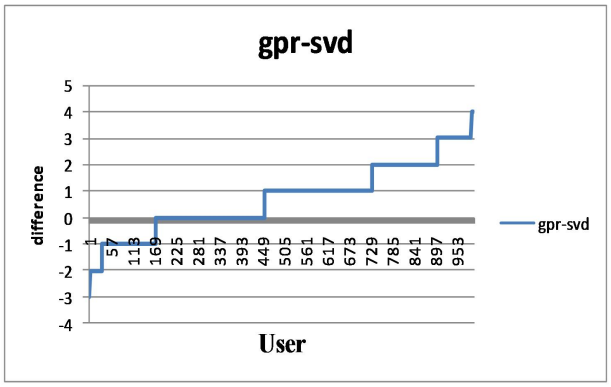


Fig. 3. The diversity difference for each user produced between GPR and SVD)

In Fig. 3, the number of recommendation genres produced by GPR is usually equal to or greater than those by SVD. The average number of genres produced by SVD is $9.88$ and the standard variation is $4.20$, and the average is $10.58$ and variation is $4.10$ for GPR. The significant analysis shows diversity produced by GPR and SVD are significantly different($p < 0.05$) indeed. In reality, we have conducted this experiment with different parameter settings several times, the results approximate the situation stated above. As GPR can recommend more diversely, it is able to improve recommendation diversity quite well.

## IV. Conclusions

In this paper, we have exploited unrated movie data to improve the performance of recommender system. Movie Frequency-Inverted Genre Frequency (MF-IGF) is proposed to measure users' genre preference, and PCC is used to compute user similarity. A matrix factorization with genre preference is introduced to produce more relevant recommendations. The result indicates that our model outperforms baseline algorithms. We have reduced the RMSE, and improved genre diversity of recommendations to a certain extent. Although the model is now used in movie recommendation, it is actually applicable for other recommendation tasks, such as book and music.

Apparently, there are still some limitations in this study, for example, computation complexity problems. In the future, we intend to conduct experiments on real data set to test whether the model can bring more diversity of recommendation. We also plan to optimize the model to recommend more relevant items.

## V. Acknowledgments

The authors gratefully acknowledges the generous support from National High-tech R&D Program of China (2013AA01A606), NSFC (61070115), Key Research Program of CAS (KJZD-EW-L04), Strategic Priority Research Program (XDA06030800) and 100-Talent Project (Y2CX093006) from Chinese Academy of Sciences. The authors would like to thank the GroupLens for collecting movie rating data.

## References

[1] G. Adomavicius and YoungOk Kwon. Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Transactions on Knowledge and Data Engineering*, 24(5):896–911, 2012.

[2] G. Adomavicius and Tuzhilin. Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. *A. IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749, 2005.

[3] R. Bell, Y. Koren, and C. Volinsky. The bellkor solution to the netflix prize. Technical report, URL http://www.research.att.com/ volinsky/netflix/ ProgressPrize2007BellKorSolution.pdf., 2007.

[4] K. Bradley and B. Smyth. Improving recommendation diversity. In *Proc. of the 12th Irish Conf. on Artificial Intelligence and Cognitive Science*, 2001.

[5] Salton G and Buckley C. Term-weighting approaches in automatic text retrieval. *Information Processing and Management*, 24(5):513C523., 1988.

[6] Z. Huang, H. Chen, and D. Zeng. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. *ACM Trans. Inf. Syst.*, 22(1):116–142, 2004.

[7] Yehuda Koren. Factorization meets the neighborhood: a multifaceted collaborative filtering model. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2008.

[8] N. N. Liu and Q. Yang. Eigenrank: a ranking-oriented approach to collaborative filtering. In *In Proc. of SIGIR' 08*, 2008.

[9] D. McSherry. Diversity-conscious retrieval. In *Proc. of the 6th European Conference on Advances in Case- Based Reasoning*, 2002.

[10] B. N. Miller, I. Albert, S. K. Lam, J. A. Konstan, and J. Riedl. Movielens unplugged: experiences with an occasionally connected recommender system. In *In IUI '03: Proceedings of the 8th international conference on Intelligent user interfaces*, 2003.

[11] A. Paterek. Improving regularized singular value decomposition for collaborative fltering. In *Proceedings of KDD Cup and Workshop*, 2007.

[12] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. Grouplens: An open architecture for collaborative filtering of netnews. In *In Proc. of CSCW' 94*, 1994.

[13] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems 20*, pages 1257–1264. MIT Press, Cambridge, MA, 2008.

[14] Xiaoyuan Su and Taghi M. Khoshgoftaar. A survey of collaborative filtering techniques. *Advances in Artificial Intelligence*, 2009.