Discretization Based Framework to Improve the Recommendation Quality

Bilal Ahmed and Wang Li
Department of Information and Computer, Taiyuan University of Technology, China

Abstract: *Recommendation systems are information filtering software that delivers suggestions about relevant stuff from a massive collection of data. Collaborative filtering approaches are the most popular in recommendations. The primary concern of any recommender system is to provide favorable recommendations based on the rating prediction of user preferences. In this article, we propose a novel discretization based framework for collaborative filtering to improve rating prediction. Our framework includes discretization-based preprocessing, chi-square based attribution selection, and K-Nearest Neighbors (KNN) based similarity computation. Rating prediction affords some basis for the judgment to decide whether recommendations are generated or not, subject to the ratio of performance of any recommendation system. Experiments on two datasets MovieLens and BookCrossing, demonstrate the effectiveness of our method.*

Keywords: *Recommender systems, collaborative filtering, prediction, discretization, chi-square.*

Received October 21, 2019; accepted July 20, 2020
https://doi.org/10.34028/iajit/18/3/13

1. Introduction

The accuracy of any recommendation system is chiefly determined by two main criterions, i.e., rating prediction and ranking [30]. Rating prediction has been extensively used in recent years in the domain of recommender systems. In any recommendation problem, the user preferences for an item represent some numerical values, also called ratings from different users for any specific items, and the goal is to predict unfamiliar ratings based on familiar ratings. It provides some basis for the judgment to decide whether recommendations are to be generated or not with respect to the ratio of performance [3]. In such a case, the accuracy of any recommendation system is estimated by measuring the error value between known and predicted ratings.

Recommendation systems are designed to help users retrieve and access relevant information about different items by suggesting relevant information automatically out of a massive volume of data. In everyday life, we face several contrasting situations where we have to pick and choose out of so many preferences, e.g., which product should be bought, which clothes we should wear, which movie to watch, and what kind of stocks to buy Bobadilla *et al.* [5] which blog or post we should read, which place to go for recreation, and which hotel we should choose Lu *et al.* [18]. To decide this entire single-handedly poses a challenging task. People these days depend on recommendations from their followers or expert guidance to make choices in any of the mentioned domains [1]. Some systems have limited functionality in terms of the defined threshold for the services such that they are not able to make the best recommendations for their consumers. If the system is not able to predict according to the taste and preference of the consumer, then probably he will stop using it after some time. This situation has guided the attention of companies towards improving their recommendation systems [21].

The rating prediction aspect of the recommendation system holds a great deal of importance and value because of its application usage [4]. Different practices are used in varied contexts to improve rating performance; each of these practices has their strengths and weaknesses [6]. Researchers use different collaborative filtering techniques such as matrix factorization methods, and deep learning methods to minimize the ratio of these errors. With an emphasis on the mentioned issue, the critical contribution of this article is to propose a new framework to improve the rating prediction accuracy. Our proposed framework includes a discretization mechanism and a chi-square testing that offers several notable advantages, such as minimizing the error significance to enhance the rating prediction along with an improvement in the quality of recommendation systems.

The article is structured as follows. In section two, we covered some related work about rating prediction and a brief background of recommendation techniques. In section three, the proposed method for rating prediction is described thoroughly. Section four is about experiments. The last part of the article provides concluding remarks.

2. Related Work

Several methods have been reported in the literature to address the rating prediction problem in recommender systems. First, we reviewed some rating prediction models related to collaborative filtering. After that, we consider some rating prediction models based on matrix factorization with collaborative filtering.

2.1. Collaborative Filtering Based Models

The method used in collaborative Filtering is to predict the preference of users for all unrated items [8]. Many collaborative filtering algorithms have been proposed to improve the recommendation performance [9, 10]. The most popular algorithm is user-based CF, and it states that consumers with the same preference in the past will also the same in the future. Sarwar *et al.* [27] propose a new algorithm for collaborative Filtering that computes the cosine similarities between item vectors and item-item correlation in a large dataset. In [16] the author proposes an improved similarity computation method that combines the item ratings and attributes for better prediction accuracy. Sarwar *et al.* [28] proposed incremental Singular Value Decomposition (SVD) algorithm based on folding-in and achieved high scalability and also provides better prediction accuracy. Periyasamy *et al.* [23] proposed a new method for rating prediction and the performance is measured under different evaluators.

2.2. Matrix Factorization Based Models

The scalability and accuracy of matrix factorization based models are very high in many perspectives [14]. These methods construct a low-rank matrix from the original rating matrix. The prediction accuracy of these models can be achieved by designing the loss function Ma *et al.* [20]. Several matrix factorization based methods for rating prediction has been applied to collaborative Filtering. In [13] author proposed a statistical latent class model with collaborative Filtering for better recommendation and rating prediction accuracy. These results show substantial improvements compared to traditional memory-based and model-based methods. Luo *et al.* [19] propose the Regularized Single-Element-Based Non-negative Matrix Factorization (RSNMF) model for computational efficiency and rating prediction accuracy for large commercial datasets. As far as we know, there is no previous research work for collaborative Filtering that uses the discretization mechanism for rating prediction.

2.3. Recommendation Methods

Different methods have been offered to make recommendations like Collaborative Filtering, Content-Based Filtering, and Hybrid Filtering. Figure 1 categorizes different baseline recommendation techniques.

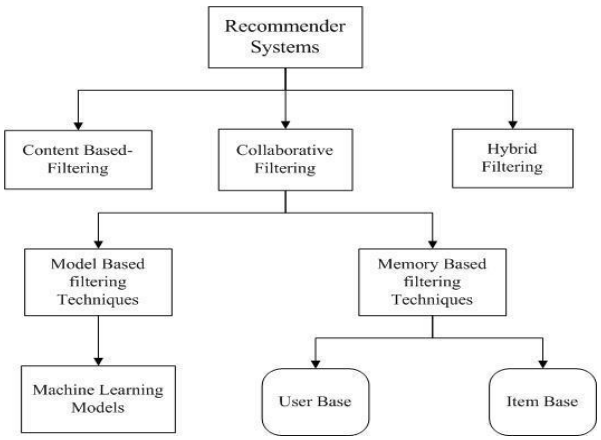

Figure 1. Different methods for recommendation.

Collaborative Filtering is one of the most commonly used assumptions in recommender systems that peoples who have a certain taste now and in the past; they would continue with their same taste also in the future. It contains some m users as U = {U1, U2, U3 …Um} and some n-type items P = {P1, P2, P3….Pn}. Then the method constructs an m x n users and items matrix, which contain the user's ratings for that specific item. The workflow of collaborative Filtering is shown in Figure 2.

Figure 2. Working of collaborative filtering.

These collaborative filtering models have two main approaches to deal with the recommendation generation problem.

Memory-Based Methods recommendations are constructed on the similarity values. Ratings are used to compute the similarity among consumers and items. The most popular and accepted standard for memory-based collaborative filtering methods is neighbor-based. They predict ratings based on similar user and related stuff. The technique implies that if two consumers have the same ratings on specific items, they may also have detailed ratings on the residual stuff. Similarly, item-based collaborative Filtering identifies items that are the same as the required stuff [24]. Model-Based Methods build an offline mode by applying data mining and machine learning techniques with the training data that can be used later for prediction. SVD factorizes the rating matrix into minimum-rank matrices to compute the missing entries. Some alternative methods are Maximum Margin Matrix Factorization (MMMF) [26] Bayesian Probabilistic matrix factorization (PMF) [29] Non-linear PMF, Non-Negative Matrix Factorization

(NMF) and Nonlinear Principal Component Analysis (NPCA) [15].

In Content-Based Filtering, items attributes are used to build recommendations for this type of filtering methods. Content-Based methods combine the ratings and purchasing activities of consumers with content information of available items. Content-based recommender methods are applicable in a broad range of areas, including recommending web sites, restaurant recommendations, article recommendations, items for sale, and different television programs [22]. The workflow of content-based Filtering is shown in Figure 3.

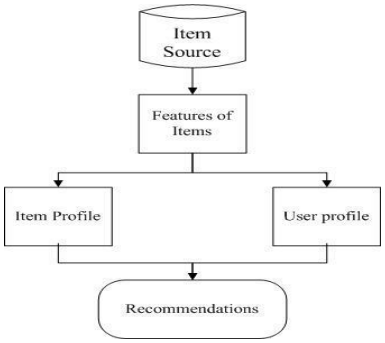

Figure 3. Working of content-based filtering.

In hybrid filtering, the system combines both features of collaborative filtering methods and content-based filtering methods in a way to bundle their complementary advantages. These systems overcome the limitations of both systems and consolidate their power to make a system with better prediction results. One example of a hybrid recommender system is Google news recommender system Das *et al.* [7].

3. Methodology

In this paper, we put forward a new framework to improve the rating prediction accuracy in the recommendation. It can be divided into three steps: discretization-based preprocessing, chi-square based attribution selection, and K-Nearest Neighbors (KNN) based similarity computation. The workflow of over proposed model is shown in Figure 4.

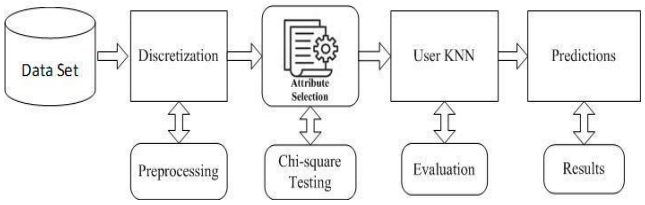

Figure 4. Workflow architecture of the proposed methodology.

First, we retrieve the dataset in memory and apply some preprocessing with the help of discretization. It converts numeric attributes into discrete attributes. The goal of discretization is to reduce the number of values and group them into some intervals or bins of equal range. This range of numerical values is grouped into different segments of equivalent sizes Ahmed *et al.* [2]. Every sector is known as a bin that represents the range covering the numerical value [12].

3.1. Discretization of MovieLens Dataset

MovieLens dataset has three tables Users, Movies, and ratings. The attributes of users are age, gender, and occupation. We are interested in analyzing the different characteristics for videos such as its release year, its genre (i.e., action, adventure, animation, or western), and rating scores are from 0 to 5.

We discretized the attribute age of users into different intervals as shown in the Table 1. For movies, we discretized the movie release decade instead of release year in the interval between the 1920s to 1990s. The decade is split into different intervals as shown in the Table 2. We removed the 0 ratings and arranged them into intervals [1, 3] and [4, 5] as shown in the Table 3.

Table1. Discretization of age for movielens dataset.

AGE	INTERVAL
17	Under 18
18	18-24
25	25-34
35	35-44
45	45-49
50	50-55
56	Over 56

Table 2. Discretization of decade for movielens dataset.

DECADE	INTERVAL
1960	Under 1970
1970	1970-1979
1980	1980-1989
1990	1990-1999
2000	Over 2000

Table 3. Discretization of ratings for movielens dataset.

RATING	INTERVAL
5	4-5
1	1-3

3.2. Discretization of BookCrossing Dataset

BookCrossing dataset contains demographic information of users such as the age of users, geographic location of users, and user approval rating on those books. The dataset keeps the book's information based on various attributes such as ISBN, Title, Author, and its Publication year. We created a subset with the help of random sampling. It included a total record of 17,150 ratings over a range of 1 to 10 for a whole number of 783 books and 10557 users. Ratings are used to decide whether the book is going to be recommended to the user or not. We discretized it into two intervals, not recommended or recommended. The discretization of ratings for BookCrossing dataset is shown in Table 4.

Table 4. Discretization of ratings for bookcrossing dataset.

RATING	INTERVAL
Not-Recommended	1-5
Recommended	6-10

3.3. Feature Selection

After discretization, we use the chi-square algorithm for the detection of features relevance. We take those features that were highly relevant to each other. With the help of relevant features or attributes, the algorithm can improve prediction accuracy and reduce the overall period of learning. Many feature selection algorithms work with discrete data rather than numerical data [17]. We periodically add all those features that are relevant to each other. After that, we calculate the relevance of each feature with the help of the chi-square statistic. Based on the high relevancy, we take it to the next level for further evaluation. We remove some features from the feature set that contains noise. For each feature, we calculate the prediction error and choose those features that have very little prediction error.

3.4. User-User Collaborative Filtering

This technique is also renowned as the user (user-KNN) collaborative Filtering. GroupLens was the first to introduce the method. The key idea in this technique is to discover consumers whose rating behavior in the past is similar to the rating behavior of users in the present. Afterward, the algorithm predicts the rating about the liking and disliking of current user Resnick *et al.* [25].

3.4.1. Computing Predictions

For the prediction of user *u*, the algorithm uses *S* to compute the neighbors $N \subset U$ of user *U*. When *N* is calculated, then the algorithm combines the rating of the user to generate predictions for items that a user prefers. It calculates the weighted average, as expressed in Equation (1).

$$\boldsymbol{P_{u,i}} = \frac{\sum_{u\prime \in N} \boldsymbol{s(u,u\prime)}(r_{u\prime,i} - \bar{r}_{u\prime})}{\sum_{u\prime \in N} |\boldsymbol{s(u,u\prime)}|} \quad (1)$$

After that, it will be normalized to a *z* score by dividing the mean rating with standard deviation **σ**u, as expressed in Equation (2).

$$P_{u,i} = \bar{r}_u + \sigma_u \frac{\sum_{u\prime \in N} s(u,u\prime)(r_{u\prime,i} - \bar{r}_{u\prime})/\sigma_{u\prime}}{\sum_{u\prime \in N} |s(u,u\prime)|} \quad (2)$$

3.4.2. Computing User Similarity

User-user collaborative Filtering applies different similarity functions to calculate the similarity. For the computation of similarity among items and users, Pearson Correlation and Vector Cosine based analogy are used. Pearson Correlation between two consumers, u and v are computed in Equation (3).

$$V_{i,j} = \frac{\sum_{u \in U} (r_{u,j} - \bar{r}_i)(r_{u,j} - \bar{r}_j)}{\sqrt{\sum_{u \in U} (r_{u,j} - \bar{r}_i)^2} \sqrt{\sum_{u \in U} (r_{u,j} - \bar{r}_j)^2}} \quad (3)$$

We use Cosine Similarity to compute the similarity between users. The calculation among *X* and *Y* can be calculated in Equation (4).

$$W_{x,y} = \cos\ (\vec{x}, y) = \frac{\vec{x}.\vec{y}}{||\vec{x}|| * ||\vec{y}||} \quad (4)$$

4. Experimental Results

This section deliberates the empirical framework for our proposed methodology, the description of our dataset, the evaluation metrics, and the results of our proposed techniques.

4.1. Dataset Description

We take two data sets for our experiments. The first dataset is well known famous dataset called MovieLens dataset. This dataset is downloaded from the MovieLens website. Different versions are available for that dataset. We use the MovieLens M1 dataset. The dataset contains 1000000 ratings, 9000 movies, 3600 tag applications, and 600 users. All data is stored in respective.csv files, namely rating.csv, movies.csv, tags.csv, and links.csv. The other dataset known as BookCrossing dataset is collected based on four weeks crawling in August and September 2004, from the BookCrossing webpage, a free online book club. BookCrossing dataset has 278,858 users, 1,149,780 ratings and 271,379 books.

4.2. Comparative Algorithms

We perform a series of experiments for comparing our proposed rating prediction model with some existing models.

4.2.1. Non Negative Matrix Factorization (Nmf)

RSNMF model for computational efficiency and rating prediction accuracy in large recommendation datasets.

4.2.2. Item Based KNN

A new algorithm for collaborative filtering that computes the cosine similarities between item vectors and item-item correlation in a large dataset.

4.2.3. Sparse SVD

An incremental singular value decomposition-based algorithm that is based on folding-in achieved high sscalability and also provided better prediction accuracy.

4.3. Evaluation Metrics

To test how accurately the recommendation system predicts the preference of consumers, researchers use different accuracy measurements. [11]. In the literature of recommendation systems, Root Mean Square Error (RMSE) and Mean Absolute Error (MAE) is the most commonly used standard measurement for measuring the rating prediction accuracy.

Root Mean Square Error is a standard technique for scoring algorithms. If $P_{a,b}$ is the expected rating for user A over the item b and also $V_{a,b}$ is the proper rating and K={(A, B)}is the set of unknown user-item ratings. We can calculate RMSE, as expressed in Equation (5).

$$\sqrt{\frac{\sum_{(i,j)\in K}(p_{i,j}-v_{i,j})^2}{n}} \quad (5)$$

Mean Absolute Error is the degree of deviation of predictions from their consumer stated values. Each rating prediction pair represented in the form of the metric <Xi, Yi> calculates mean absolute error between them. It is computed by adding this absolute error of N rating predictions and finally computing the average, as expressed in Equation (6).

$$MAE = \frac{\sum_{I=1}^{N}|X_I - Y_I|}{N} \quad (6)$$

The lower error means the prediction engine can predict user ratings accurately.

4.4. Result Comparisons for MovieLens Dataset

Figure 5 summarizes the comparison of the results with Item-based KNN, Sparse SVD, and Matrix Factorization based methods with above proposed method. It can be seen that our framework performs well and minimize the RMSE value.

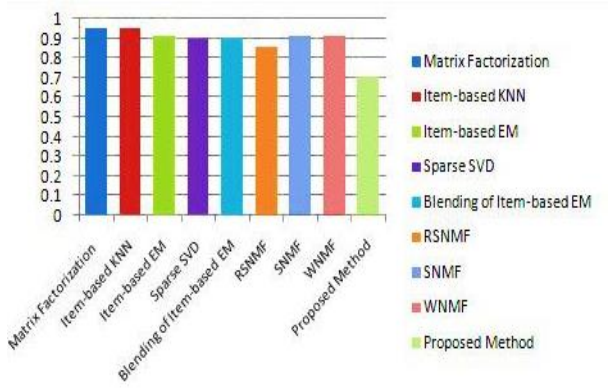

Figure 5. RMSE of different methods.

Figure 6 summarizes the comparison of the results with item-based collaborative filtering and matrix factorization based methods with above proposed method. It can be seen that our method performs well and minimize the MAE value.

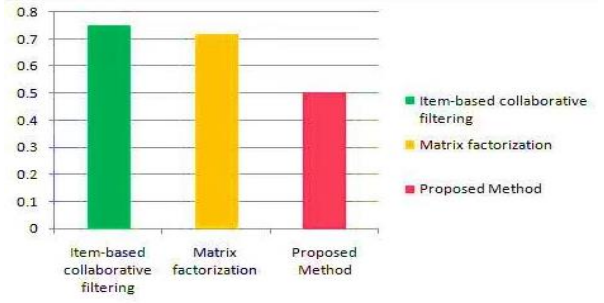

Figure 6. MAE of different methods.

Figure 7 summarizes the comparison of the results with maximum margin matrix factorization methods with above proposed method. It can be seen that our method performs well and minimize the NAME value.

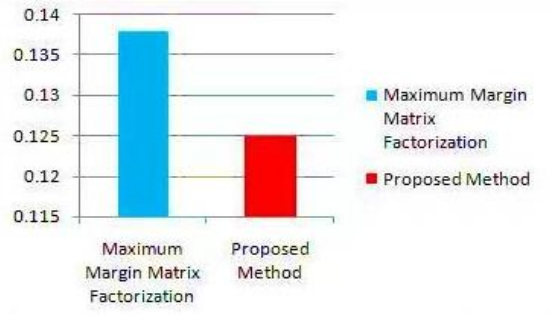

Figure 7. NMAE of different methods.

4.5. Result Comparison for BookCrossing Dataset

Figure 8 summarizes the comparison of the results of KNN without discretization. It can be seen that our method performs well and minimizes the RMSE while performing experiments on the BookCrossing dataset.

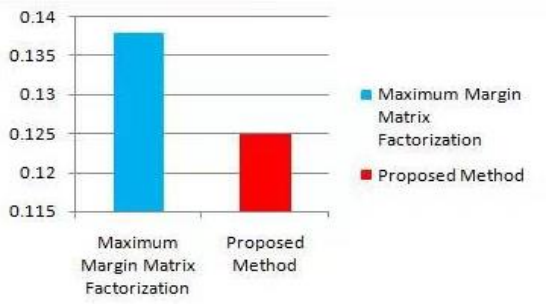

Figure 8. RMSE of different methods.

Figure 9 summarizes the results of normalized mean absolute error that was achieved by using above proposed method.

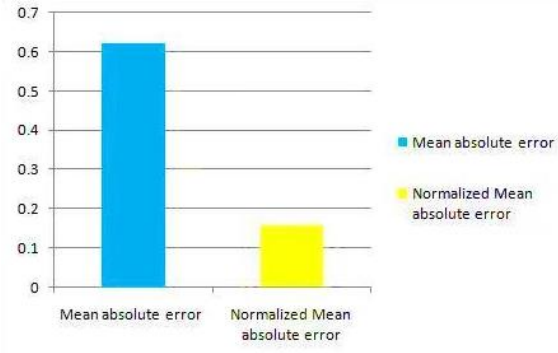

Figure 9. MAE and NMAE of the proposed method.

5. Conclusion and Future Work

Recommendation systems are powerful information filtering Software providing relevant products to consumers. Rating prediction has been extensively used in recent years in the domain of recommender

systems. The primary concern is to provide favorable recommendations based on the prediction of user preferences. In this paper, we propose a novel framework for rating prediction accuracy. Our method includes a discretization based preprocessing, chi-square based attribute selection, and KNN based similarity computation. The experiment results demonstrate that using discretization makes an excellent contribution to rating prediction. While comparison with existing approaches our method shows significant improvements on different benchmarks. Our proposed method can be extendable with the integration of different attributes information of users or items contextual information. This method can also be extendable to a group of users in group recommendation system.

References

[1] Adomavicius G. and Tuzhilin A., "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," *IEEE Transactions on Knowledge and Data Engineering*, vol. 17, no. 6, pp. 734-749, 2005.

[2] Ahmed B., Wang L., Hussain W., Qadoos M., Tingyi Z., Amjad M., Badar-ud-Duja S., Hussain A., and Raheel M., "Optimal Rating Prediction in Recommender Systems," *in Proceedings of International Conference on Data Service*, Ningbo, pp. 331-339, 2019.

[3] Bellogín A., "Performance Prediction in Recommender Systems," *in Proceedings of International Conference on User Modeling, Adaptation, and Personalization*, Girona, pp. 401-404, 2011.

[4] Bellogín A., Castells P., and Cantador I., "Predicting the Performance of Recommender Systems: An Information Theoretic Approach," *in Proceedings of Conference on the Theory of Information Retrieval*, Bertinoro, pp. 27-39, 2011.

[5] Bobadilla J., Ortega F., Hernando A., and Gutiérrez A., "Recommender Systems Survey," *Knowledge-Based Systems*, vol. 46, pp. 109-132, 2013.

[6] Cremonesi P., Koren Y., and Turrin R., "Performance of Recommender Algorithms on Top-N Recommendation Tasks," *in Proceedings of the 4th ACM Conference on Recommender Systems*, New York, pp. 39-46, 2010.

[7] Das A., Datar M., Garg A., and Rajaram S., "Google News Personalization: Scalable Online Collaborative Filtering," *in Proceedings of the 16th international Conference on World Wide Web*, New York, pp. 271-280, 2007.

[8] Ekstrand M., Riedl J., and Konstan J., *Collaborative Filtering Recommender Systems*, Now the Essence of Knowledge, 2011.

[9] Fletcher K. and Liu X., "A Collaborative Filtering Method for Personalized Preference-Based Service Recommendation," *in Proceedings of IEEE International Conference on Web Services*, New York, pp. 400-407, 2015.

[10] Gao S., Yu Z., Shi L., Yan X., and Song H., "Review Expert Collaborative Recommendation Algorithm Based on Topic Relationship," *Journal of Automatica Sinica*, vol. 2, no. 4, pp. 403-411, 2015.

[11] Gunawardana A. and Shani G., "A Survey of Accuracy Evaluation Metrics of Recommendation Tasks," *Journal of Machine Learning Research*, vol. 10, pp. 2935-2962, 2009.

[12] He X., Min F., and Zhu W., "A Comparative Study of Discretization Approaches for Granular Association Rule Mining," *in Proceedings of 26th IEEE Canadian Conference on Electrical and Computer Engineering*, Regina, pp. 1-5, 2013.

[13] Hofmann T., "Latent Semantic Models for Collaborative Filtering," *ACM Transactions on Information Systems*, vol. 22, no. 1, pp. 89-115, 2004.

[14] Koren Y., Bell R., and Volinsky C., "Matrix Factorization Techniques for Recommender Systems," *Computer Society*, vol. 42, no. 8, pp. 30-37, 2009.

[15] Lee J., Sun M., and Lebanon G., "A Comparative Study of Collaborative Filtering Algorithms," *arXiv preprint arXiv:1205.3193*, pp. 1-27, 2012.

[16] Li Z., Huang M., and Zhang., "A Collaborative Filtering Algorithm of Calculating Similarity Based on Item Rating and Attributes," *in Proceedings of 14th Web Information Systems and Applications Conference*, Liuzhou, pp. 215-218, 2017.

[17] Liu H. and Setiono R., "Chi2: Feature Selection and Discretization of Numeric Attributes," *in Proceedings of 7th IEEE International Conference on Tools with Artificial Intelligence*, Herndon, pp. 388-391, 1995.

[18] Lu J., Wu D., Mao M., Wang W., and Zhang G., "Recommender System Application Developments: A Survey," *Decision Support Systems*, vol. 74, pp. 12-32, 2015.

[19] Luo X., Zhou M., Xia Y., and Zhu Q., "An Efficient Non-Negative Matrix-Factorization-Based Approach to Collaborative Filtering for Recommender Systems," *IEEE Transactions on Industrial Informatics*, vol. 10, no. 2, pp. 1273-1284, 2014.

[20] Ma H., Yang H., Lyu M. R., and King I., "Sorec: Social Recommendation Using Probabilistic Matrix Factorization," *in Proceedings of the 17th*

ACM Conference on Information and Knowledge Management, New York, pp. 931-940, 2008.

[21] Mohamed M., Khafagy M., and Ibrahim M., "Recommender Systems Challenges and Solutions Survey," *in Proceedings of International Conference on Innovative Trends in Computer Engineering*, Aswan, pp. 149-155, 2019.

[22] Pazzani M. and Billsus D., *the Adaptive Web*, Springer Link, 2007.

[23] Periyasamy K., Jaiganesh J., Ponnambalam K., Rajasekar J., and Arputharaj K., "Analysis and Performance Evaluation of Cosine Neighbourhood Recommender System," *The International Arab Journal of Information Technology*, vol. 14, no. 5, pp. 747-754, 2017.

[24] Resnick P., Iacovou N., Suchak M., Bergstrom P., and Riedl J., "Grouplens: An Open Architecture for Collaborative Filtering of Netnews," *in Proceedings of the ACM Conference on Computer Supported Cooperative Work*, New York, pp. 175-186, 1994.

[25] Rennie J. and Srebro N., "Fast Maximum Margin Matrix Factorization for Collaborative Prediction," *in Proceedings of the 22nd International Conference on Machine Learning*, Germany, pp. 713-719, 2005.

[26] Rennie J. and Srebro N., "Fast Maximum Margin Matrix Factorization for Collaborative Prediction," *in Proceedings of the 22nd International Conference on Machine Learning*, Germany, pp. 713-719, 2005.

[27] Sarwar B., Karypis G., Konstan J., and Riedl J., "Item-Based Collaborative Filtering Recommendation Algorithms," *in Proceedings of the 10th international conference on World Wide Web*, New York, pp. 285-295, 2001.

[28] Sarwar B., Karypis G., Konstan J., and Riedl J., "Incremental Singular Value Decomposition Algorithms for Highly Scalable Recommender Systems," *in Proceedings of 5th International Conference on Computer and Information Science*, Aswan, pp. 8-27, 2002.

[29] Shan H. and Banerjee A., "Generalized Probabilistic Matrix Factorizations for Collaborative Filtering," *in Proceedings of IEEE International Conference on Data Mining*, Sydney, pp. 1025-1030, 2010.

[30] Steck H., "Evaluation of Recommendations: Rating Prediction and Ranking," *in Proceedings of the 7th ACM Conference on Recommender Systems*, New York, pp. 213-220, 2013.

Bilal Ahmed is a PhD candidate in Taiyuan University of Technology, China. He received his MS degree from Pakistan. His research areas include machine learning, deep learning and recommendation systems.

Wang Li is a professor and PhD supervisor in Data science college, Taiyuan University of Technology, China. Her research area is big data computing, machine learning and recommendation systems.