An Improved Collaborative Filtering Recommendation Algorithm for Big Data

Hafed Zarzour[1(✉)], Faiz Maazouzi[2], Mohamed Soltani[1], and Chaouki Chemam[3]

[1] LIM Research, Department of Computer Science, University of Souk Ahras, 41000 Souk Ahras, Algeria
hafed.zarzour@gmail.com, msoltani@univ-soukahras.dz
[2] Department of Computer Science, University of Souk Ahras, Souk Ahras, 41000 Souk Ahras, Algeria
mazouzi@labged.net
[3] Department of Computer Science, University of El-Tarf, 36000 El-Taref, Algeria
chemam-chaouki@univ-eltarf.dz

Abstract. With the increase of volume, velocity, and variety of big data, the traditional collaborative filtering recommendation algorithm, which recommends the items based on the ratings from those like-minded users, becomes more and more inefficient. In this paper, two varieties of algorithms for collaborative filtering recommendation system are proposed. The first one uses the improved k-means clustering technique while the second one uses the improved k-means clustering technique coupled with Principal Component Analysis as a dimensionality reduction method to enhance the recommendation accuracy for big data. The experimental results show that the proposed algorithms have better recommendation performance than the traditional collaborative filtering recommendation algorithm.

Keywords: Big data · Recommender system
Collaborative filtering recommendation algorithm · K-means
Clustering · PCA

1 Introduction

With the explosive increase in available data on the web and the rapid advances of information technology, big data has become a hot research topic in the field of data mining. Generally, it is commonly used to describe the exponential growth and availability of structured and unstructured data. Nowadays, many governmental and industrial communities become interested in the high potential of this innovative technology. However, it is very difficult for such communities to find relevant contents, recommender systems appear to solve present problems. Recommender system is defined as a decision making strategy for users under complex information platforms [1] in which it can effectively recommend the required information to end-users. Various techniques for developing recommender systems have been proposed, which

A. Amine et al. (Eds.): CIIA 2018, IFIP AICT 522, pp. 660–668, 2018.
https://doi.org/10.1007/978-3-319-89743-1_56

can use either content-based filtering, collaborative filtering or hybrid methods [2–5]. In particular, the collaborative filtering recommendation algorithm (CFRA) is popular and has been used by many providers and consumers of big data such as: eBay, Amazon and Facebook.

Recently, many researches have reported that applying k-means as clustering technique in collaborative recommender systems can significantly enhance the performance of traditional CFRA [6]. Moreover, it has been proved that using Principal Component Analysis (PCA) as a dimensionality reduction method can significantly improve the clustering techniques [7], therefore, it is necessary to conduct dimensions reeducation before formally conducting clustering tasks. Hence, in this paper, we propose two varieties of algorithms for an effective collaborative filtering recommendation system. The first one uses the improved k-means clustering technique while the second one uses the improved k-means clustering technique coupled with PCA as a dimensionality reduction method to enhance the recommendation accuracy for big data. The experimental results show that the proposed algorithms have better recommendation performance than the traditional collaborative filtering recommendation algorithm.

The rest of this paper is organized as follows: Sect. 2 discusses some related works. Section 3 presents the collaborative filtering recommendation algorithm. Section 4 explains in details the proposed approach. Section 5 describes the experimental results. Finally, Sect. 6 concludes this study and proposes the plans for future work.

2 Related Work

In the recent years, the philosophy of big data attracts great attention from several official organizations including governments, universities, and industries in which the recommender systems are introduced to help them to find what they need via a mechanism that can make prediction depending on different criteria. One of the recommender strategies that can provide several kinds of recommendation is the open source project Apache Mahout [8]. It is primary enables free scalable implementation of machine learning methods [9, 10]. Another free and open source scalable library of recommender system is MyMediaLite [11], which addresses both common rating and item prediction from positive-only feedback. The rating prediction can be a scale of 1 to 5 stars while the item prediction from positive-only implicit feedback can be purchase actions or from clicks. In [12], the authors propose a keyword-aware service recommendation method, named KASR, to indicate users' preferences and generate appropriate recommendations on MapReduce [13] for big data applications. In [14], Lee et al. propose an adaptive recommendation algorithm, ACFSC, that is focused on scalable clustering to solve the problem of scalability by composing neighborhood based on reducing time complexity. They also address the problem of sparsity by making items' and users' feature vectors incrementally learning. CSRS [15] is a customized service recommendation system for Big Data. It uses the MapReduce framework and focuses on service recommendation method to create proper recommendations based on users' preferences. In [16], Zarzour et al. propose a new collaborative filtering recommendation algorithm based on dimensionality reduction and

clustering techniques. They use clustering k-means algorithm and Singular Value Decomposition (SVD) to cluster similar users and reduce the dimensionality, respectively. In [17], the authors use k-means algorithm to cluster users according to their interests and then voting algorithm to generate prediction in recommender systems.

3 Collaborative Filtering Recommendation Algorithm

In the field of recommender systems, the collaborative filtering recommendation algorithm (CFRA) is the most successful recommendation method. The behind idea of CFRA is to provide for an active user recommendations or predictions by first looking for users who share the same rating patterns with him and then using the ratings from those like-minded users found to calculate a prediction for him. In other words, CFRA can suggests new similar items or predict the interest of a certain item for an active user based on their previous likings and the preferences of other similar users. More technically, it uses a user-item rating matrix that includes the preferences for items by users for matching users with relevant performances obtained by employing a similarity function between theirs profile to make recommendations or predict the ratings of selected items [18, 19].

$$sim\,(a,b) = \frac{\sum_{i \in I_{ab}} (r_{ai} - \bar{r}_a) \times (r_{bi} - \bar{r}_b)}{\sqrt{\sum_{i \in I_{ab}} (r_{ai} - \bar{r}_a)^2} \times \sqrt{\sum_{i \in I_{ab}} (r_{bi} - \bar{r}_b)^2}} \tag{1}$$

To compute the similarity between users or items, there are several similarity measure functions. One of the most popular methods is by using Pearson Correlation Coefficient (PCC), which is defined as follows:

$$p_{ti} = \bar{r}_t + \frac{\sum_{u \in U_{nei}} sim\,(t,\ u) \times (r_{ui} - \bar{r}_u)}{\sum_{u \in U_{nei}} |sim\,(t,u)|} \tag{2}$$

Once the similarity is computed, the most N nearest users are selected as a group of similar users called neighborhood and predicted ratings of unrated item can be then computed. The recommendation formula is presented as follow:

The main steps of the collaborative filtering recommendation algorithm (CFRA) are as follows:

Step 1: Input the matrix M[m, n] of user-item rating data, active user, K;
Step 2: Calculate the similarity between users by using Pearson Correlation Coefficient (PCC) and generate the similarity matrix S[m, m];
Step 3: Calculate the similarity between the active user and the clusters;
Step 4: Select the first n similar users of the active user;
Step 5: Calculate the prediction values of active user to every cluster by using the formula (2);
Step 6: Choose the top N items of users as recommendations;
Step 7: Output the recommendations.

4 K-means Based-Collaborative Filtering Algorithm

In this paper, two varieties of algorithms for collaborative filtering recommendation system are proposed. The first one uses directly the k-means clustering technique while the second one uses the k-means clustering technique after performing the PCA method. PCA aims at reducing the dimensions of the big data by extracting the most important information from the data. It can make big data mining more useful and get similar results by the reduction of dimensions [20].

4.1 K-means Algorithm

In data mining, K-means is considered as one of the most widely used method of clustering [21] in which it generates automatically a set of clusters based on a collection of datasets in easiest way. The main aim of k-means is to make the similarity inter-points of the same cluster be high, while the similarity inter-clusters be low. The steps of the algorithm are as follows:

Step 1: Input dataset, clusters number and K;
Step 2: Select randomly initial clustering centers which is the initial value of K;
Step 3: Calculate the distances between centers and objects then assign objects to the most nearest cluster;
Step 4: For each cluster, calculate the average as new partition centers;
Step 5: Use the new partition centers to redistribute points into new clusters;
Step 6: Repeat Steps 4 and 5 until the algorithm converge to a stable partition;
Step 7: Output K clusters.

4.2 CFRA-Km: A Collaborative Filtering Recommendation Algorithm Based on K-means Clustering

The general k-means algorithm is now personalized in order to take into consideration the recommendation requirements as well as the perdition of unknown ratings for a given active user. The specific steps are as follows:

Step 1: Input the matrix M[m, n] of user-item rating data, active user, K;
Step 2: Calculate the similarity between users by using Pearson Correlation Coefficient (PCC) and generate the similarity matrix S[m, m];
Step 3: Use the matrix S[m, m] as dataset and select randomly initial clustering centers which is the initial value of K;
Step 4: Calculate the distances between centers and objects then assign objects to the most nearest cluster;
Step 5: For each cluster, calculate the average as new partition centers;
Step 6: Use the new partition centers to redistribute points into new clusters;
Step 7: Repeat Steps 5 and 6 until the algorithm converge to a stable partition;
Step 8: Calculate the similarity between the active user and the clusters;
Step 9: Select the first n similar clusters of the active user;
Step 10: Calculate the prediction values of active user to every cluster by using the formula (2);

Step 11: Choose the top N items of users as recommendations;
Step 12: Output the recommendations.

4.3 Reducing the Dimension by PCA

One of the purposes of a PCA is the analysis of big data for eliminating noises and finding patterns to reduce the dimensions of the data without loss of relevant information. To do this, it converts a collection of observations of possibly correlated variables into a collection of values of principal components by using a linear transformation called orthogonal transformation. In general, the quantity of the obtained principal components is less than or equal to the quantity of original variables. Therefore, PCA is used as a statistical method to reduce not only the dimension of the user–user ratings matrix but also to reduce the loss of information by employing eigenvalue decomposition of data covariance matrix to obtain principal components of dataset with their weights. The general steps of PCA are as follows:

Step 1: Input the dataset;
Step 2: Normalize the data in the dataset;
Step 3: Calculate the covariance of the corresponding matrix;
Step 4: Calculate the eigenvectors of the covariance matrix;
Step 5: From matrix multiplication, translate the data to be in terms of the principal components.
Step 6: Output principal components.

4.4 CFRA-Km-PCA: A Collaborative Filtering Recommendation Algorithm Based on K-means Clustering and PCA

The first version of our k-means clustering- based collaborative filtering recommendation algorithm does not consider the effect of the dimensions reduction which may significantly influence the prediction results and make them inaccurate. Thus, PCA is applied before conducting the k-means clustering and performing the prediction step to reduce the dimension of the dataset and improve the performance of the prediction results. In other words, the collaborative filtering recommendation algorithm based on K-means clustering and PCA called CFRA-Km-PCA combines the advantages of PCA method with those of k-means clustering technique. The specific steps of CFA-Km-PCA are as follows:

Step 1: Input The matrix M[m, n] of user-item rating data, active user, K;
Step 2: Calculate the similarity between users by using Pearson Correlation Coefficient (PCC) and generate the similarity matrix S[m, m];
Step 3: Normalize the data in the obtained S[m, m];
Step 4: Calculate the covariance of the corresponding matrix;
Step 5: Calculate the eigenvectors of the covariance matrix;
Step 6: From matrix multiplication, translate the data to be in terms of the principal components.
Step 7: Use the obtained principal components matrix as dataset and select randomly initial clustering centers which is the initial value of K;

Step 8: Calculate the distances between centers and objects then assign objects to the most nearest cluster;
Step 9: For each cluster, calculate the average as new partition centers;
Step 10: Use the new partition centers to redistribute points into new clusters;
Step 11: Repeat Steps 5 and 6 until the algorithm converge to a stable partition;
Step 12: Calculate the similarity between the active user and the clusters;
Step 13: Select the first n similar clusters of the active user;
Step 14: Calculate the prediction values of active user to every cluster by using the formula (2);
Step 15: Choose the top N items of users as recommendations;
Step 16: Output the recommendations.

5 Experimentation Results and Evaluation

To evaluate the performance of the k-means clustering-based collaborative filtering recommendation algorithm with and without using PCA compared to traditional collaborative filtering recommendation algorithm, experimentations were conducted on real big data. The experimental dataset was obtained from Netflix [22] which contains over 17,770 movies rated by approximately 480 000 users. In this dataset, there are over 100 million ratings ranging from 1 to 5 stars. A random sample was chosen and 80% of these data were also randomly used for training, and the remaining data were selected to test the performance of the considered algorithms.

In the performance evolution of recommender systems, Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) are the most widely used. Therefore, we used those metrics to evaluate the performance of recommendations in CFRA, CFRA-Km, and CFRA-Km-PCA algorithms.

The formulas of RMSE and MAE are shown as follows, respectively.

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N} (p(i) - q(i))^2}{N}} \tag{3}$$

$$MAE = \frac{\sum_{i=1}^{N} |p(i) - q(i)|}{N} \tag{4}$$

Figure 1 shows the experimental results in terms of RMSE metric for the proposed algorithms. As we can see from the graph, the RMSE results of the proposed CFRA-Km and CFRA-Km-PCA is low in the whole neighbors range compared to that for the CFRA algorithm. More precisely, the CFRA-Km-PCA achieves better results than both other algorithms.

Figure 2 shows the experimental results in terms of MAE metric for the three algorithms. In the same way, we can observe from the graph that the MAE results of the proposed CFRA-Km and CFRA-Km-PCA is low in the whole neighbors range

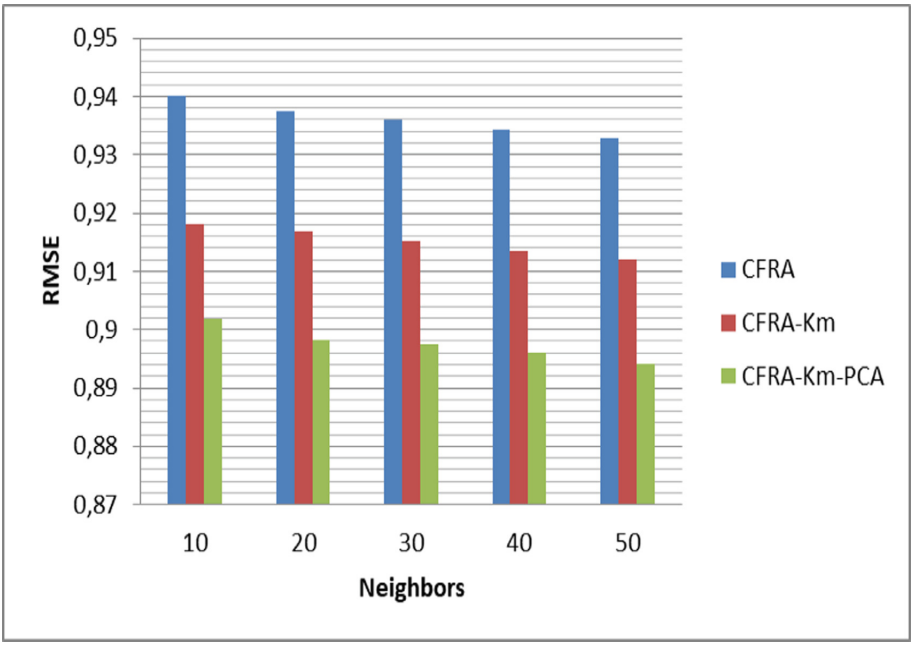

Fig. 1. RMSE results for CFRA, CFRA-Km, and CFRA-Km-PCA.

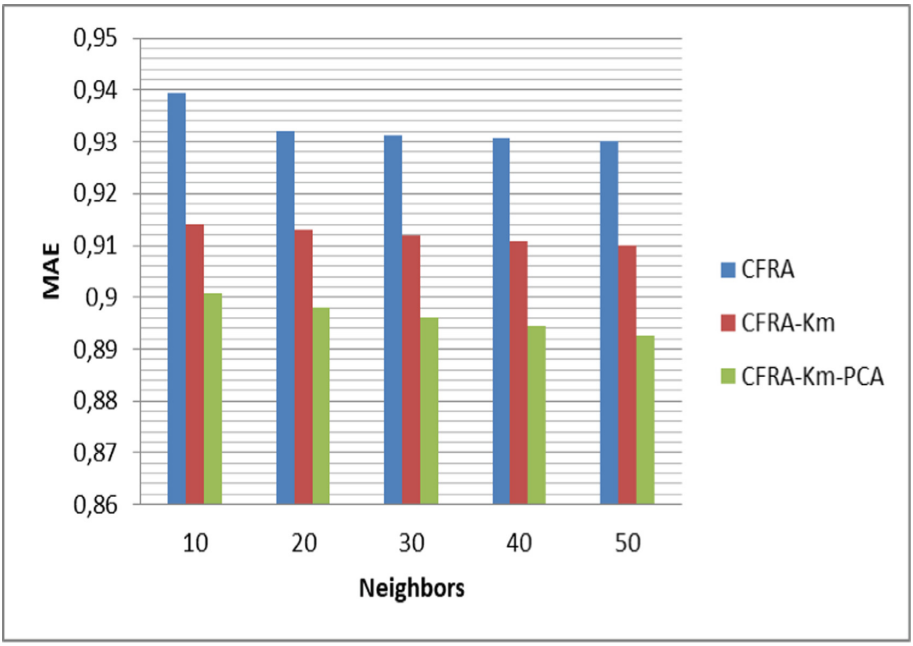

Fig. 2. MAE results for CFRA, CFRA-Km, and CFRA-Km-PCA.

compared to that for the CFRA algorithm and the CFRA-Km-PCA achieves better accuracy than both other algorithms.

From Figs. 1 and 2, we can conclude that the proposed algorithms, CFRA-Km and CFRA-Km-PCA, have better performance than the traditional algorithm CFRA in terms of RMSE and MAE. We can also conclude that the combination of PCA method with K-means clustering technique improved significantly the recommendation performance, which indicates that CFRA-Km-PCA is better algorithm for using in recommendation system for big data.

6 Conclusion and Future Work

In this paper, we have presented two kinds of improved collaborative filtering algorithms intended to enhance the prediction accuracy in the big data context. The first algorithm uses only the k-means clustering technique, while the second algorithm combines the advantages of both k-means clustering technique and PCA method. PCA was adapted to conduct dimensions reduction before formally conducting clustering tasks, which improved significantly the performance of k-means clustering-based collaborative filtering recommendation algorithm. The recommendation algorithms were evaluated in terms of RMSE and MAE metrics. The experimental results showed that the CFRA-Km-PCA achieved better results than both other algorithms, CFRA and CFRA-Km.

In the future, we will apply our algorithms to other datasets, and study the mechanism of the dimensions reduction coupled with other clustering techniques for improving recommendation precisions.

References

1. Rashid, A.M., Albert, I., Cosley, D., Lam, S.K., McNee, S.M., Konstan, J.A., Riedl, J.: Getting to know you. In: Proceedings of the 7th International Conference on Intelligent User Interfaces - IUI 2002 (2002)
2. Merve Acilar, A., Arslan, A.: A collaborative filtering method based on artificial immune network. Expert Syst. Appl. **36**(4), 8324–8332 (2009)
3. Chen, L., Hsu, F., Chen, M., Hsu, Y.: Developing recommender systems with the consideration of product profitability for sellers. Inf. Sci. **178**(4), 1032–1048 (2008)
4. Jalali, M., Mustapha, N., Sulaiman, M.N., Mamat, A.: WebPUM: a web-based recommendation system to predict user future movements. Expert Syst. Appl. **37**(9), 6201–6212 (2010)
5. Smith, B., Linden, G.: Two decades of recommender systems at amazon.com. IEEE Internet Comput. **21**(3), 12–18 (2017)
6. Koohi, H., Kiani, K.: A new method to find neighbor users that improves the performance of collaborative filtering. Expert Syst. Appl. **83**, 30–39 (2017)
7. Pourkamali-Anaraki, F., Becker, S.: Preconditioned data sparsification for big data with applications to PCA and k-means. IEEE Trans. Inf. Theory **63**(5), 1 (2017)
8. Gupta, P., Sharma, A., Jindal, R.: Scalable machine-learning algorithms for big data analytics: a comprehensive review. Wiley Interdisc. Rev.: Data Min. Knowl. Disc. **6**(6), 194–214 (2016)

9. Bagchi, S.: Performance and quality assessment of similarity measures in collaborative filtering using mahout. Proced. Comput. Sci. **50**, 229–234 (2015)
10. Verma, J.P., Patel, B., Patel, A.: Big data analysis: recommendation system with Hadoop framework. In: 2015 IEEE International Conference on Computational Intelligence and Communication Technology (2015)
11. Gantner, Z., Rendle, S., Freudenthaler, C., Schmidt-Thieme, L.: MyMediaLite. In: Proceedings Of The Fifth Acm Conference On Recommender Systems - Recsys 2011 (2011)
12. Meng, S., Dou, W., Zhang, X., Chen, J.: KASR: a keyword-aware service recommendation method on mapreduce for big data applications. IEEE Trans. Parallel Distrib. Syst. **25**(12), 3221–3231 (2014)
13. Cheng, D., Rao, J., Guo, Y., Jiang, C., Zhou, X.: Improving performance of heterogeneous mapreduce clusters with adaptive task tuning. IEEE Trans. Parallel Distrib. Syst. **28**(3), 774–786 (2017)
14. Lee, O.J., Hong, M.S., Jung, J.J., Shin, J., Kim, P.: Adaptive collaborative filtering based on scalable clustering for big recommender systems. Acta Polytech. Hung. **13**(2), 179–194 (2016)
15. Bande, VM., Pakle, K.: CSRS: Customized service recommendation system for big data analysis using map reduce. In: 2016 International Conference on Inventive Computation Technologies (ICICT) (2016)
16. Zarzour, H., Al-Sharif, Z., Al-Ayyoub, M., Jararweh, Y.: A new collaborative filtering recommendation algorithm based on dimensionality reduction and clustering techniques. In: 9th International Conference on Information and Communication Systems (ICICS) (2018)
17. Dakhel, GM., Mahdavi, M.: A new collaborative filtering algorithm using k-means clustering and neighbors' voting. In: 11th International Conference on Hybrid Intelligent Systems, HIS (2011)
18. Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM Trans. Inf. Syst. **22**(1), 5–53 (2004)
19. Aggarwal, Charu C.: Neighborhood-based collaborative filtering. In: Charu, C. (ed.) Recommender Systems. TIRS, pp. 29–70. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-29659-3_2
20. Villalba, S.D., Cunningham, P.: An evaluation of dimension reduction techniques for one-class classification. Artif. Intell. Rev. **27**(4), 273–294 (2007)
21. Adeniyi, D.A., Wei, Z., Yongquan, Y.: Automated web usage data mining and recommendation system using K-nearest neighbor (KNN) classification method. Appl. Comput. Inform. **12**(1), 90–108 (2016)
22. Hallinan, B., Striphas, T.: Recommended for you: the Netflix prize and the production of algorithmic culture. New Media Soc. **18**(1), 117–137 (2014)