International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-4, November 2019

Community-Driven Collaborative Recommendation System

Laxmi Chaudhary, Buddha Singh

Abstract*: Recommendation systems (RSs) are an application of community detection, becoming more significant in our daily lives. They play a significant role in suggesting information to users such as products, services, friends and so on. A novel community driven collaborative recommendation system (CDCRS) has been proposed by the authors, in this particular paper. Furthermore, K means approach has been utilized to detect communities and extract the relationship among the users. The singular value decomposition method (SVD) is also applied. Issues of sparsity and scalability of the collaborative method are considered. Experiments were conducted on MovieLens datasets. Movie ratings were predicted and top-k recommendations for the user produced. The comparative study that was performed between the proposed as well as the collaborative filtering method dependent on SVD (CFSVD) as well as the results of experiments shows that CFSVD is outperformed by the proposed CDCRS method.

Keywords: collaborative filtering, community detection, recommendation system, scalability, sparsity

I. INTRODUCTION

One of the latest phenomena is Recommendation System (RSs) because of the advanced development in the online social networks as well as electronic commerce all over the network. The RS's primary objective is to help users find the products and services of their interest and saving them time when searching [1]. Recommendation systems can efficiently address the overload of information, a problem initiated by the increasingly overwhelming amount of data. It does so by filtering relevant information, predicting the likely preferences of users and then recommending items closer to users' preferred pastimes to facilitate further decision-making [2]. Belonging to communities also points to preferences and non-preferences of a given user.

Nowadays, a new type of collaborative recommendation method called clustering/community-based recommendation [3],[4],[5], has emerged. Using clustering/community-based approaches, the collective behavior of users is predictable. Such RSs mostly utilizes the community users, extracted from the social network data of many users, so that items can be recommended to such users. Among the community-based approaches, CF methods have been widely used for RSs. Although, they managed recommending famous items that majority of users as well as their friends have some interest in but there may not be any equivalent user preferences individually that leads to the problem of sparse data [6].

Scalability as well as sparsity are the recommendation systems' most critical issues. When the system is being scaled up the item number increases to millions or billions, resulting in a lower likelihood that similar items are being focused by two different users. The clustering-based recommendation is used as one of the common approaches for alleviating the issue of data sparsity [7]. It is found that the scalability of RSs are not good because the CF (Collaborative Filtering) methods complexity precisely based on the users and item number. The increase in the items as well as users number leads in requiring more resources or slowing down the resources. One way to handle both the sparsity and the scalability issue in CF is by employing the SVD approach [8]. The SVD model helps in extracting latent features. Essentially, it can map each user and each item in a latent space. Therefore, it helps gain understanding of the relationship of users and items as they become directly correlated.

In this paper, the community driven collaborative recommendation system (CDCRS) is proposed as a new method; it explores the relationship between users employing the user-rating matrix by the k means clustering method [9]. It detects groups of users based on the ratings it assigned.

This paper mainly contributes as:

(a) a user level community driven collaborative recommendation system using K means;

(b) incorporating SVD to address issues of sparsity and scalability;

(c) Proposed method evaluation using the MovieLens dataset. Furthermore, the movie rating is predicted as well as then top-k recommendations are produced for the user; and

(d) The comparison of the results from the new method with the CFSVD method.

Further, the remaining paper is arranged as: Section 2 represents the related work's brief review and a discussion of the proposed community driven collaborative recommendation system (CDCRS) in Section 3, furthermore, section 4 presented proposed method's experimental results. In the end, conclusion and future work are discussed in section 5.

II. RELATED WORK

Several studies on user recommendations have emerged during the past decades.

Manuscript published on November 30, 2019.
* Correspondence Author
Laxmi Chaudhary, SC & SS, Jawaharlal Nehru University, New Delhi, India.
Buddha Singh, SC & SS, Jawaharlal Nehru University, New Delhi, India.

Most of the recommendation approaches experience scalability and sparsity issues. Recommendation methods that depends on the Community/Clustering have been developed to address these limitations by grouping users (items), identifying similar users. Generally, rating data dependent similarity is computed by the majority of clustering-based recommendation methods as well as after that a basic clustering algorithm, like *K*-means method, is employed so as to generate the groups of items (users).

Ji *et al.* [10] discovered the items and user's implicit similarity. For this, initially item (or user) latent factor vectors' are clustered into item (or user) cluster-level factor vectors. After that, original approximation is compressed into cluster-level factor vectors dependent cluster-level rating-patterns. Wang *et al.* [11] used the *K*-means algorithm for user clustering as well as after that absent rating is estimated in the user–item matrix for predicting the target user preference. Jain and Rana [5] suggested a DRS (Dynamic Recommendation System) which helps in clustering the users through a new algorithm.

Tsuji and Puntheeranurak [12] recommended a hybrid recommendation system in which a fuzzy *K*-means clustering algorithm is used for clustering the users. Furthermore, the traditional CF (Collaborative Filtering) algorithms is improved by combining both the clustered and original data that is resulted by recommendations. Sarwar *et al.* [13] clustered users through their suggested bisecting *K*-means clustering algorithm into multiple clusters. In such approach, the target users nearest neighbours are elected depending on the particular segment to which user belongs.

III. PRPOSED MODEL

This proposed model is a community driven collaborative recommendation system (CDCRS); it is divided into two parts.

- Application of the SVD Model
- Detection of Community or Cluster

The authors have utilized k-means clustering approach [14] to find the communities of users based on ratings. The Singular Valued Decomposition (SVD) [15] model is employed to address sparsity and scalability issues of the dataset. Ratings are predicted for users and, top 5 recommendations are included for new users.

A. Employing the SVD Model

A matrix factorization-based method, namely, SVD, [15] that can be used to divide any given single utility matrix, A, into 3 matrices product given as:

$$A = U \times \sum \times V^T \quad (1)$$

where, *U* as well as *V* represents left as well as right singular vectors as well as the ∑'s diagonal values are known as singular values. *U* is a diagonal matrix, it represents the relationship between users and latent (hidden) features. ∑ represents each latent factor within those users and items matrix. V^T is the transpose of the item matrix, it indicates the similarity between items and the latent factors.

The SVD model is employed on the user item matrix to handle sparsity and scalability issues in the collaborative filtering method. The authors analyze how close the recommendations are for a given user, thereby solving an optimization problem. The most common metric for identifying accuracy in recommendations is the Root Mean Square Method (RMSE). The lower the RMSE, the better the performance of the recommendation system or prediction. Unseen items have unknown ratings, and are for now ignored, and only known items included, for which we minimize the RMSE entered in the user item Euclidean matrix. SVD is applied to reach minimal RMSE.

B. Community Detection

The rating matrix helps determine the user communities by using unsupervised learning, that is, the k means clustering approach [14]. The k means approach is a simple non-hierarchical clustering approach to find the groups. Distance is used as the metric within the K communities data set. Each community is described by the centroid that is calculated as the distance mean. Furthermore, a dataset X is assumed that is having n multidimensional data points along with K category that needs to be divided, as well as for the similarity measure Euclidean distance is selected. Also, approach of clustering mainly aims in minimizing the SSE (Sum of the Squares Errors) of several kinds [15].

$$SSE = \sum_{k=1}^{k} \sum_{i=1}^{n} \| x_i - \mu_k \|^2 \quad (2)$$

where K cluster centres is represented by k, k^{th} centre is represented by μ_k, and data set's i^{th} point is represented by xi. T is the mean of the points in the clusters $c(x_i)$. The SSE, or inertia, is the variance within the communities' data from the centroid. K means determines centroids that minimize the inertia across all clusters.

The key concept in implementation of algorithm is that K sample points are randomly extracted from the sample set as the initial cluster's center: Initially, every sample point is divided as the cluster that is signified by the closest center point; after that every sample point's center point in every cluster is considered as the cluster's center point. The above steps are repeated till the cluster's center point does not change or attains a fixed iteration number. As the center point is changed, it results in the changes in the outcomes of the algorithms, which then results in outcome's instability. The center point determination is based on the k-value selection that is the algorithms main focus; this point precisely influences the outcomes of clustering, like global or local optimality [16].

C. Selecting the K Value

There are several different methods for selecting the number of communities/clusters, K, such as by rule of thumb, Elbow method, Information Criterion, etc. [17]. This work uses the elbow method; its basic idea is to calculate a series of K values with the squares of the distance among sample points in every cluster along with cluster's centroid. The SSE is utilized as performance indicator through iterating K-value as well as measure the SSE. The small K-values show high cluster convergence. If clusters move towards real clusters then SSE demonstrates a quick decline. While the clusters exceed to real clusters, then SSE will also move slowly towards decline. The elbow technique by plotting a graph between K-values and total calculated errors by utilizing K values. Fig. 1 shows the SSE vs. number of clusters. It shows that k=6 is the optimal value.

The algorithm of Elbow method is given as:

Algorithm 1: Elbow method
Input: Dataset
Output: SSE, k

1. SSE= []
2. for k=1, k in range do
3. $SSE = \sum_{k=1}^{k} \sum_{i=1}^{n} \| x_i - \mu_k \|^2$
4. Return SSE, k

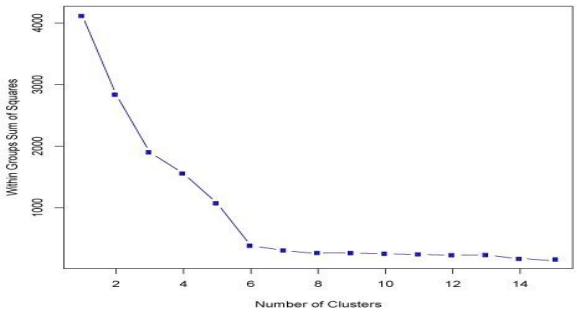

Fig. 1. SSE vs Number of Clusters (K)

IV. SIMULATION RESULTS

An experiment is conducted by us on MovieLens dataset (http://grouplens.org/datasets/movielens/). Such particular dataset contain 100,000 ratings on near about 9,125 movies by 943 users. Items as well as users are consecutively numbered from 1. A *userid* is given to a single user whereas *movieid* is given to a single movie.

The file format is given as: information of movies (item) is given by *movies.csv*. A demographic information of user is given by *ratings.csv*, it is basically a tab separated list of "*user id | movieId | rating | timestamp"*. Every data line shows the rating list of every movie by a single user.
Python 3.4.4 software is used for this particular proposed method, a computer having an Intel Core i7, 12 GB RAM and 3.20 GHz. A subset of the dataset 671 users and 9064 movies was taken to perform the experiments. Fig. 2 and Fig. 3 show the result obtained before and after applying the SVD. It addresses sparsity issue of the data very well.

Then the users' communities are determined depends upon the ratings, utilizing K means approach. The k value is found by the help of Elbow technique. Fig. 2 shows that the perfect K value is 5. The prediction of the rating and the top 5 recommendations for the user 7 based on the communities is given in Fig. 5.

title	"Great Performances" Cats (1998)	$9.99 (2008)	'Hellboy': The Seeds of Creation (2004)	'Neath the Arizona Skies (1934)	'Round Midnight (1986)	'Salem's Lot (2004)	'Til There Was You (1997)	'burbs, The (1989)	'night Mother (1986)	(500) Days of Summer (2009)
userId										
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
6	NaN	NaN	NaN	NaN	NaN	NaN	NaN	4.0	NaN	NaN

Fig. 2. Result obtained before applying SVD

title	Forrest Gump (1994)	Pulp Fiction (1994)	Shawshank Redemption, The (1994)	Silence of the Lambs, The (1991)	Star Wars: Episode IV - A New Hope (1977)	Jurassic Park (1993)	Matrix, The (1999)	Toy Story (1995)	Schindler's List (1993)	Terminator 2: Judgment Day (1991)	...	Insomnia (2002)	What Lies Beneath (2000)	Roman Holiday (1953)
0	-0.321063	-0.176517	0.255352	0.002401	-0.065513	-0.113064	-0.203266	-0.221517	0.104014	0.088175	...	0.055563	0.047214	-0.019134
1	3.644738	3.424150	1.848717	3.541152	0.196704	3.777428	-0.228887	0.139443	4.142952	3.208389	...	0.150983	0.033033	0.166895
2	3.592689	3.196432	4.012971	2.964975	1.032569	1.539034	1.937418	1.216869	2.583940	0.465589	...	-0.140508	0.163534	-0.273651
3	2.358074	3.276486	-0.197565	0.854323	5.009534	3.436446	2.224318	0.665552	-0.579391	4.789692	...	-0.506439	-0.239231	-0.008172
4	3.694249	-0.985334	0.775256	1.024560	-1.502407	1.411088	-0.449035	0.468822	0.469582	-0.596986	...	0.329656	0.374572	-0.194395

Fig. 3. Result obtained after applying SVD

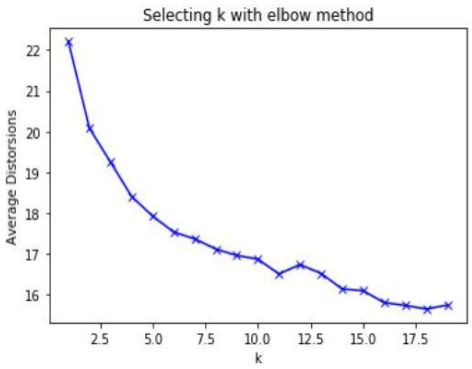

Fig. 4. Average distortions vs K

Pulp Fiction (1994)	3.346667
Forrest Gump (1994)	3.240000
Specialist, The (1994)	1.120000
Nell (1994)	1.040000
First Knight (1995)	0.960000

Fig. 5. Top 5 recommendations for user 7 using CDCRS

The result obtained by simple user collaborating filtering using SVD (CFSVD) is shown in Fig. 6. It gives the prediction of the rating and for the user 7 top 5 recommendations are mentioned.

Recommended Movies	user_ratings	user_predictions
Fugitive, The (1993)	0.0	2.224754
Aliens (1986)	0.0	1.859570
Men in Black (a.k.a. MIB) (1997)	0.0	1.849957
Apollo 13 (1995)	0.0	1.729457
Alien (1979)	0.0	1.704993

Fig. 6. Top 5 recommendations for user 7 using user collaborating filtering using SVD

A. Evaluation Methodology

The RMSE (Root Mean Square Error) was utilized to estimate the performance of CDCRS approach as well as collaborative filtering using SVD method. The RMSE value is the difference among actual along with predicted ratings. The rmse values for K [4, 5, 6, 7] and CFSVD were compared to see which is the better predictor. Table I. demonstrated the rmse score attained by using proposed method CDCRS at K=5 performs well.

Table I. Comparison of Rmse scores using different methods

Methods	Rmse Score
CDCRS K=4	0.0058
CDCRS K=5	0.0010
CDCRS K=6	0.0033
CDCRS K=7	0.0045
CFSVD	0.0065

V. CONCLUSION AND FUTURE WORK

A community driven collaborative recommendation system was proposed. The K means clustering method found the user communities based on the users' given rating. SVD helped tackle the issues of sparsity and scalability. The proposed method predicts the rating and give top-k recommendation for the user. Experimental outcomes demonstrate the proposed method as giving well results as compared to CFSVD model without using k means. The proposed method yields better rmse scores and predicts more suitable movies to the users. For future work, item level recommendations will be compared to these results.

REFERENCES

1. X. Yang et al., "Collaborative filtering-based recommendation of online social voting," IEEE Trans. Comput. Social Syst., vol. 4, no. 1, pp. 1–13, Mar. 2017.
2. Fatemi, Maryam, and Laurissa Tokarchuk. "A Community Based Social Recommender System for Individuals & Groups." Social Computing (SocialCom), 2013 International Conference on. IEEE, 2013.
3. K. W.-T. Leung,D. L. Lee,W.-C. Lee. CLR: A Collaborative Location Recommendation Framework based on Co-Clustering. In Proceedings of SIGIR, pages 305-314, 2011.
4. I. Esslimani, A. Brun, and A. Boyer, "A collaborative filtering approach combining clustering and navigational based correlations," in Proc. 5th Int. Conf. Web Inf. Syst. Technol., Mar. 2009, pp. 364–369.
5. C. Rana and S. K. Jain, "An evolutionary clustering algorithm based on temporal features for dynamic recommender systems," Swarm Evol. Comput., vol. 14, pp. 21–30, Feb. 2014.
6. S.-J. Yen, Y.-S. Lee, C.-H. Lin and J.-C. Ying, Investigating the Effect of Sampling Methods for Imbalanced Data Distributions, Proceedings of IEEE International Conference on Systems, Man, and Cybernetics (SMC'2006), pp. 4163-1468, October 2006.
7. J. D. West, I. Wesley-Smith, and C. T. Bergstrom, "A recommendation system based on hierarchical clustering of an article-level citation network," IEEE Trans. Big Data, vol. 2, no. 2, pp. 113–123, Jun. 2016.
8. Deng, Wei, et al. "Incorporating community detection and clustering techniques into collaborative filtering model." Procedia Computer Science 31 (2014): 66-74.
9. Kanungo, Tapas, et al. "An efficient k-means clustering algorithm: Analysis and implementation." IEEE Transactions on Pattern Analysis & Machine Intelligence 7 (2002): 881-892.
10. K. Ji, R. Sun, X. Li, and W. Shu, "Improving matrix approximation for recommendation via a clustering-based reconstructive method," Neurocomputing, vol. 173, pp. 912–920, Jan. 2016.
11. Q. Wang, W. Cao, and Y. Liu, "A novel clustering based collaborative filtering recommendation system algorithm," in Advanced Technologies, Embedded and Multimedia for Human-Centric Computing. Springer, 2014, pp. 673–680.
12. S. Puntheeranurak and H. Tsuji, "A multi-clustering hybrid recommender system," in Proc. 7th IEEE Int. Conf. Comput. Inf. Technol. (CIT), Oct. 2007, pp. 223–228.
13. B. M. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Recommender systems for large-scale e-commerce: Scalable neighborhood formation using clustering," in Proc. 5th Int. Conf. Comput. Inf. Technol., vol. 1, Dec. 2002, pp. 291–324.
14. Yuan, Chunhui, and Haitao Yang. "Research on K-Value Selection Method of K-Means Clustering Algorithm." J 2.2 (2019): 226-235.
15. Wang, Q.; Wang, C.; Feng, Z.; Ye, J. Review of K-means clustering algorithm. Electron. Des. Eng. 2012, 20, 21–24.
16. Ravindra, R.; Rathod, R.D.G. Design of electricity tariff plans using gap statistic for K-means clustering based on consumers monthly electricity consumption data. Int. J. Energ. Sect. Manag. 2017, 2, 295–310.
17. Kodinariya, Trupti M., and Prashant R. Makwana. "Review on determining number of Cluster in K-Means Clustering." International Journal 1.6 (2013): 90-95.

AUTHORS PROFILE

Laxmi Chaudhary received her M. Tech degree in Computer Science from School of Computer and Systems Sciences at Jawaharlal Nehru University, New Delhi, India in 2016. Currently, she is a Ph.D. research scholar at School of Computer and Systems Sciences, Jawaharlal Nehru University. Her research interest includes Community detection, Social Networks and Artificial Intelligence.

Buddha Singh received his B.Tech degree from Madhav Institute of Technology and Science, Gwalior, India and Ph.D degree from Jawaharlal Nehru University, New Delhi, India. In 2014, he joined Jawaharlal Nehru University as an Assistant Professor. His current research interest includes Mobile Ad-hoc Network, Wireless Sensor Network, Cognitive Radio, Big Data Analytics, Complex Networks, Mobile Computing.