Movie Recommender System using Single Value Decomposition and K-means Clustering

Mayur Rahul¹, Vinod Kumar², Vikash Yadav³ and Rishabh⁴

¹Department of Computer Application, UIET CSJM Kanpur, Uttar Pradesh, India
²,³,⁴ABES Engineering College, Ghaziabad, Uttar Pradesh, India
¹mayurrahul209@gmail.com, ²vinod.kumae@abes.ac.in, ³vikash.yadav@abes.ac.in,
⁴jain2986@gmail.com

Abstract: The recommender system is technique and tool to filter the massive overloaded information for suggesting most useful information to the user in a personalized manner. In the period of “Big Data”, the researcher experiences many problems to process big data accurately and efficiently. In this work, we introduce an efficient model using Single Value Decomposition (SVD) as a method for dimensions reduction and K-means clustering as classification method. Our proposed method and its corresponding results have been evaluated and compared with other existing methods using metrics like standard deviation (SD), mean absolute error (MAE) root mean square error (RMSE), t-value, Dunn index, average similarity and computational time using two publicly available datasets using Flixter dataset and MovieLens dataset. The result demonstrates that our proposed method is able to outperform other existing methods.

Keywords: Recommender systems, single value decomposition, standard deviation, mean absolute error, Flixter dataset.

1. Introduction

With the huge demand of websites and development in computer technology, accessing information from the website is very widespread and people need more filtered information [9,10]. To fulfill the requirement of people around us, the recommender systems are used to filter huge information. The recommender systems mostly use collaborative filtering to suggest personalized information to the user [11,12]. The technique behind this filtering is to suggest the item for the user according to likings of similar users. The assumption after this filtering technique is that if the users have similar interest then they also share same information [13,14]. The advantages of collaborative filtering which makes it popular among researchers are: (1) it doesn’t depend on the content of the items. (2) It can work with the social network. (3) Comparatively good accuracy.

The common problem of recommender system is to manage big data to produce accurate recommendations. Mostly few problems arise while developing appropriate recommender systems, they are: (1) the ratings given by the user is very scanty, collate with lots of information and users in the RS. (2) The more data requires more efficient technique to reply quickly. (3) The frequently changing nature of data requires efficient technique to update correctly. The RS continuously requires data, and the likings of user and interest continuously changing. The training data changes according to the surfing of the user. So, we have to consider these changes in the recommender systems [15,16].
To overcome the problem arises in recommendation systems, many clustering and dimensionality reduction technique have been introduced. They used partition around medoids (PAM) [1] and minimum spanning tree [2] for clustering and non-negative matrix factorization, single value decomposition for important feature extraction in recommender systems[17,18]. These techniques can efficiently reduce the data sparsity and reduce the computational cost. The disadvantages are that their accuracy is not good and unable to handle dynamic problems effectively.

To know the concept of dynamic and online computations for collaborative recommender systems, we provide an example for it. The users, movies and ratings are not changing, when user enters to the movie RS. The continuous addition of users, movies and ratings create problem for the movie RS to suggest movies and ratings [19,20]. It is also not easy to predict ratings of movie added every day. In this case it takes the ratings of similar users in terms of preferences of the users for the ratings of the movies.

In this research, we applied single value decomposition (SVD) for dimensions reduction and k-means clustering for the prediction of the ratings of the movie using two publicly available movie dataset called MovieLens and Flixter. Our main outcomes of this work are mentioned below:

(1) We introduced a novel movie recommender system using SVD and k-means clustering.
(2) The results of our model is assessed using metrics like standard deviation (SD), root mean square error (RMSE), mean absolute error (MAE), t-value, dunn matrix and average similarity and computational time.
(3) The results are obtained using two publicly available datasets called MovieLens and Flixter.

The remaining paper is organised as follows: Related works are explained in section 2, proposed method is described in section 3, experiment and results are in section 4 and finally concluded in section 5.

2. Related Works

The recommender system is used to take decision in very complex situations. It is also used as a tool to filter valuable information in the complicated e-commerce applications according to the user’s likings and preferences [3]. Recommender system is defined by accommodating and supplementing for prediction of social procedure for others to make some relevant preferences when there is lack of information. Recommender system is also used as an information filtering process in the situation where there is lots of information exists.

Leiberman et al. proposed a system which focuses on user ratings to create training sets called Fab. It was an excellent example of content-based filtering. Other users can find the appropriate information in the internet with the help of Letizia [4]. The given system helps the user to find the information using GUI, guided using user’s internet history. They basically make the pattern of browsing history to predict the appropriate web page. Pazzani et al. introduced a Naïve Bayesian classification tool to identify the browsing configuration of the user to examine the web pages in which user interested in [5]. Some uses neural networks to create framework to model user’s interests in Usenet news.

A recommender system is used to achieve useful information of the products, information and services to the user by combining preference from the other users, from different authorities and from user attributes [6]. The recommender systems are of two types: collaborative filtering and content-
based filtering. The items are selected in collaborative filtering are generally based on the relation between current users and past users of the product and services. The items are suggested in content-based filtering based on relation between user and its preferences of the product and services [7]. The main difference between both the techniques is that the similarity of preferences between users has been considered in collaborative filtering while in content-based, the similarity of preferences between user and item is considered. Besides these two filtering technique, there is another filtering method which is the arrangement of collaborative and content-based filtering called hybrid filtering.

Adomavicius et al. proposed a framework called GroupLens based on collaborative filtering technique used to help users to find articles in huge database [21]. Chen et al. introduced an efficient collaborative based technique called Ringo which is used to recommend profile based ratings to suggest music albums [22]. Amazon uses title diversification technique to enhance its recommendations [23]. The collaborative filtering technique is used to overcome the problem of scalability arises by generating tables using items-items matrix. Then system predicts user’s preference based on the history of the user. In content based filtering, compare the items with the user’s preferences. This technique is based on the user’s information.

Even with the development of recommendation systems using these two techniques, some restrictions are also there to decrease the performance. Data anatomy and sparsity including overspecialisation is the most common problems in content based filtering [21]. However, in collaborative based filtering, the most common problems are sparsity, cold-start and scalability. These are generally most common problems which affects the performance of the recommendation systems. To overcome these problems and to increase the productivity of the recommendation systems, a new filtering technique is introduced called hybrid filtering. It is basically the combination of collaborative and content based filtering methods. This technique is used to increase the capability of both techniques while neutralizing their weaknesses. The content and collaborative based techniques are widely used technologies so far in recommendation system by both collaborative and content based techniques separately produce results, further combining both the results yields a recommendation system having both techniques. This becomes unified recommender system [21].

Ghazantar et al. proposed a cascaded hybrid recommender system by combining features, ratings, demographic approach. They also address the cold-start and data sparsity while developing their recommender system [22]. Zeiglar et al. proposed a collaborative based approach for classify product or item using huge taxonomic information. They address data sparsity problem in their system. They focus on topic diversification and super-topic score to generate profiles of users [23]. Sarwar et al introduced a system which is combination of a filtering agent and collaborative filtering [24]. They also combine both collaborative and content based filtering agent. They also able to overcome the problem of average user in collaborative and new user in content based filtering techniques [25]. Cunningham et al. introduced a recommender system which is basically the combination of content and collaborative based technique [26]. Konstas et al. proposed a method which combines play counts, social relations and tagging information to produce accurate results [27]. Lee et al. proposed a system which incorporated the social information in collaborative filtering to get the information of users connected in social networks [28]. Condiff et al. creates a model based on Bayesian effects on ratings and user information and ratings to enhance the performance of the recommender systems [29].
Table 1. Comparison of Related Works

| Filtering Methods     | Techniques used                          | Advantages                                      | Disadvantages                                |
|-----------------------|-----------------------------------------|-------------------------------------------------|----------------------------------------------|
| Collaborative Filtering | Linear Regressions Clustering K-Nearest Neighbour | No domain knowledge required Quality is directly proportional to time | New user boost-up problem                    |
| Content-based         | Clustering Decision Trees                | No domain knowledge required                     | New user boost-up problem Quality depends on big data |
| Hybrid                | Different voting methods Creating single unified model | Able to avoid following issues: Content description | User boost-up problem                        |

3. Proposed Methodology

The block diagram of proposed movie recommendation system is given in figure 1. Our system uses the SVD for reduction if dimensions and classification based on k-means clustering method using two publicly available dataset called MovieLens and Flixter [8, 9]. The k-means clustering is the most popular method among the researchers and SVD is consistent to eliminate the unwanted data from the population. The combination of these methods produces good results.

The population is taken from the publicly available datasets, MovieLens and Flixter. The SVD is used to extract some important features from the population. The dimensionality reduction is done because handling of huge data is very complicated and expensive. Also, it is used to remove some meaningless data from the dataset.

Figure 1. Block diagram of proposed methodology

Clustering is an unsupervised learning method. The objective of this technique is to allocate an item to the group so that the items in the group are more identical than that of other groups. The aim of this method is to produce more meaningful groups in the datasets. In this work, similarity is calculated using Euclidean distance given by:

\[ d(x,y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \]  (1)
The main aim of clustering technique is to minimizing intra-cluster distance and maximizing inter-cluster distance. K-means clustering is the most popular partitioned technique. The function partition the given N dataset into k disjoint subsets Sj which consist Nj data points as close as possible to the other points on the basis of similarity measures. Each partition contains Nj data points for the given Cj centroids. In order to make good partition the distance between all the data points with the centroids should be minimized. The objective of our research is to minimized E from eq. 2 given below.

$$E = \sum_{1}^{k} \sum_{n \in S_j} d(X_n, C_j)$$

(2)

Where

- $X_n = n^{th}$ Vector
- $C_j = Centroid$
- $d = distance$ measure

The interchange of items takes place between clusters until the value of E can’t decreases further.

4. Experiments and Results

We conducted experiments using two datasets: Flexiter and MovieLens. The first dataset is Flexiter, is publicly available free data created for research work. The Flexiter dataset has been made from Flexiter website having movie ratings [8]. This dataset contains user rating of movies taken from year 2005 to 2009. The dataset contains 48,794 items, 7,86,936 users and 8196077 ratings. We take small data from the Flexiter original data using following conditions: allow user more than 240 ratings and allow item more than 25 ratings only. The original dataset can easily converted into matrix which have 8890 users, 10106 movies and 57,38,920 ratings. The ratings value ranges from 0.5 to 5. Maximum number represents strong recommendations and lower number represents dislike.

| Evaluation metrics using Flexiter and MovieLens datasets |
|---------------------------------------------------------|
| **Flexiter**                                           | **MovieLens**  |
| SD                                                      | 0.13743       | 0.11453       |
| MAE                                                     | 0.73372       | 0.62896       |
| RMSE                                                    | 0.94876       | 0.92934       |
| T-value                                                 | 2.66674       | 3.74562       |
| Dunn Index (4 Clusters)                                 | 0.34873       | 0.31945       |
| Average Similarity (4 Clusters)                          | 0.96          | 0.95          |
| Computational Time in sec                                | 71.89         | 57.43         |

The second dataset is MovieLens. The MovieLens database is a freely available online dataset for the analysis of result of our proposed recommender system, developed by Minnesota University under
the project of Group Lens research. This database contains 1 lakhs ratings (1-5) and having 1682 movies rated by 943 users. Each user rated at least 20 movies. We n-cross validation rule for the dataset in our recommendation systems. In this rule, both datasets have been divided into n parts. The first n-1 is used to train and remaining is used to test. The final result is the average of all accuracy from n parts.

All codes have been written in python 3.1.8 version in the computer, which has a configuration of i5 processor, 6 GB RAM.

The results obtained from our proposed method using Flexiter and MovieLens dataset has been depicted in table 2. The value of standard deviation using MovieLens is 0.11453, which is better than the value of standard deviation using Flexiter. Similarly, MAE, RMSE, dunn index, average similarity and computational time using MovieLens dataset is far better than the values obtained using Flexiter except in T-value. The plot of all the values found in table is represented in pictorial form (see fig. 2). The conclusion made by the results is that the overall performance of our proposed method is better in MovieLens dataset.

5. Conclusion and Future Work

In this paper, we introduce a method using SVD and K-means clustering using two publicly available dataset i.e. Flexiter and MovieLens. We also able to evaluate our proposed method using various metrics like standard deviation (SD), root mean square error (RMSE), mean absolute error (MAE), t-value, dunn index and average similarity and computational time. The analysis of each results have been rigorously done to check which dataset is better for our proposed method for movie RS. In future, we incorporate some more supervised learning techniques to improve the results for our proposed method. Also, incorporate more evaluation metrics to check the performance of recommender systems.

References

[1] P.S. Chakarborty,”A Scalable collaborative filtering based recommender system using incremental clustering”, In Advance Computing Conference, IACC2009, IEEE International, IEEE, 2009, pp. 1526-1529.
[2] G.N. Demir, A.S. Udyar, S.G.Udugucu, Graph-based sequence clustering through multi objective evolutionary algorithms for web recommender systems, In Proceedings of the 9th annual conference on genetic and evolutionary computation, ACM, 2007, pp. 1943-1950.

[3] Schaefer JB, Konstan J, Riedl J. Recommender system in e-commerce, In Proceedings of the 1st ACM conference on electronic commerce, 1999, pp. 158-66.

[4] Leberman H. letizia: an agent that assists web browsing. In Proceeding of the 1995 international joint conference on artificial intelligence. Montreal, Canada, pp. 924-929.

[5] Pazzani MJ. A framework for collaborative, content-based and demographic filtering, Artificial intelligence rev, 1999; 13, pp. 393-408, No. 5(6).

[6] Frias-Martinez, E., Magoulas, G., Chen, S. Y., & Macredie, R. (2006). Automated user modeling for personalized digital libraries. International Journal of Information Management, 26(3), pp. 234–248.

[7] Su, X. & Khoshgoftaar, T. M. (2009). A survey of collaborative filtering techniques, Advances in artificial intelligence, 2009, 4.

[8] Flixtersdataset, http://www.cs.sfu.ca/~sja25/personal/datasets.

[9] Y. Deldjoo and M. Schedl, "Retrieving Relevant and Diverse Movie Clips Using the MFVCD-7K Multifaceted Video Clip Dataset," 2019 International Conference on Content-Based Multimedia Indexing (CBMI), Dublin, Ireland, 2019, pp. 1-4, doi: 10.1109/CBMI.2019.8877420.

[10] Hug, N., (2020). Surprise: A Python library for recommender systems. Journal of Open Source Software, 5(52), 2174. https://doi.org/10.21105/joss.02174.

[11] Emrah Inan, Fatih Tekbacak, Cemalettin Ozturk, Moreopt: A goal programming based movie recommender system, Journal of Computational Science, 28, 2018, Pages 43-50, ISSN 1877-7503, https://doi.org/10.1016/j.jocs.2018.08.004.

[12] Bogdan Walek, Vladimir Fojtik, A hybrid recommender system for recommending relevant movies using an expert system, Expert Systems with Applications, 158, 2020, 113452, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2020.113452.

[13] Enthilselvan Natarajan, Subramaniyaswamy Vairavasundaram, Sivaramakrishnan Natarajan, Amir H. Gandomi, Resolving data sparsity and cold start problem in collaborative filtering recommender system using Linked Open Data, Expert Systems with Applications, 149, 2020, 113248, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2020.113248.

[14] Mohammed Fadhel Aljunid, Manjaiah Dh, An Efficient Deep Learning Approach for Collaborative Filtering Recommender System, Procedia Computer Science, 171, 2020, Pages 829-836, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2020.04.090.

[15] Bushra Alhijawi, Yousef Kilani, A collaborative filtering recommender system using genetic algorithm, Information Processing & Management, 57(6), 2020,102310, ISSN 0306-4573, https://doi.org/10.1016/j.ipm.2020.102310.

[16] Sunita Tiwari, Anu Saini, Vaibhav Paliwal, Ajay Singh, Rajat Gupta, Ruchika Mattoo, Implicit Preferences Discovery for Biography Recommender System Using Twitter, Procedia
[17] R. Singla, S. Gupta, A. Gupta and D. K. Vishwakarma, "FLEX: A Content Based Movie Recommender," International Conference for Emerging Technology (INCET), Belgaum, India, 2020, pp. 1-4, doi: 10.1109/INCET49848.2020.9154163.

[18] M. Gupta, A. Thakkar, Aashish, V. Gupta and D. P. S. Rathore, "Movie Recommender System Using Collaborative Filtering," International Conference on Electronics and Sustainable Communication Systems (ICESC), Coimbatore, India, 2020, pp. 415-420, doi: 10.1109/ICESC48915.2020.9155879.

[19] R. E. Nakhli, H. Moradi and M. A. Sadeghi, "Movie Recommender System Based on Percentage of View," 5th Conference on Knowledge Based Engineering and Innovation (KBEd), Tehran, Iran, 2019, pp. 656-660, doi: 10.1109/KBEI.2019.8734976.

[20] Y. Deldjoo, M. Schell and M. Elahi, "Movie Genome Recommender: A Novel Recommender System Based on Multimedia Content," International Conference on Content-Based Multimedia Indexing (CBMI), Dublin, Ireland, 2019, pp. 1-4, doi: 10.1109/CBMI.2019.8877452.

[21] Adomavicius G, Tuzhilin A, Towards the next generation of recommender system, A survey of the state-of-the-art and possible extension. IEEE Trans Knowl Data Engg 2005; 17(6), pp. 734-49.

[22] Chen LS, Hsu FH, Chen MC, Hsu YC. Developing recommender systems with the consideration of product profitability for sellers, Int Jurnl of Info Sci 2008, 178(4), pp. 1032-1048.

[23] Zeigler CN, McNee SM, Konstan JA, Lausen G. Improving recommendations lists through topic diversification. In Proceedings of the 4th international conference on World Wide Web, 2005, pp. 22-32.

[24] Ghazantar MA, Pragel-Bannet A. A scalable accurate hybrid recommender system. In the 3rd international conference on knowledge discovery and data mining (WKDD 2010). IEEE computer society, Washington, DC, USA.

[25] Zeiglar CN, Lausen G, Schmidt-Theime L. Taxonomy-driven computation of product recommendations. In Proceedings of 13th international conference on information and knowledge management (CIKM’04), IEEE computer society, Washington, DC, USA; 2004, pp. 406-415.

[26] Sarwar BM, Konstan JA, Herlocker JL, Miller B, Riedl JT. Using filtering agents to improve prediction quality in the grouplens research, collaborative filtering system. In Proceedings of the ACM conference on computer supported cooperative work, NY (USA), ACM New York, 1998, pp. 345-54.

[27] Burke R. Hybrid web recommender system. In: Brusilovsky P, Kobsa A, Nejdl W, editors. The adaptive web, LNCS 4321, Berlin, Heidelberg, Germany; springer; 2007, pp. 377-408.

[28] Cunningham P, Bergmann R, Schmitt S, Trapponer R, Breen S, Smyth B,”WeSell: Intelligent sales assistants for the World Wide Web, In Proceedings CBR in Ecommerce, Vancouver BC, 2001, pp.104-109.
[29] Konstan I, Stathopoulos V, Jose JM. On social networks and collaborative recommendations. In The proceedings of the 32nd international ACM conference (SIGIR’09), ACM. New York, NY USA, 2009, pp. 195-202.