An Efficient Collaborative Recommender System for Removing Sparsity Problem

Avita Fuskele Jain, Santosh Kumar Vishwakarma and Prashant Jain

Abstract Recommender Systems is a special type of information filtering system which has become important in the information overloaded and strategic decision making environment. Recommender System is used to produce meaningful suggestions about new items for particular consumers. These recommendations may be based on the user profile or item ratings, facilitate the users to make decisions in multiple contexts, such as what items to buy, what online news to read or what music to listen. Recommender Systems helps their founders to increase profits by recommending items and attracting new consumers. Collaborative filtering technique recommends items basis of conclusion of opinions about various products by users of similar profile to the active user. This technique requires user-items-ratings matrix. Although this is the most mature and commonly implemented technique, it faces major problem of Data Sparsity problem. Sparsity Problem occurs as a result of lack of enough information when only a few of the total number of items are rated by the users. This produces a sparse user item matrix leads to weak recommendations. This paper presents a recommender system using collaborative filtering implemented with RapidMiner tool. The proposed recommendation system is designed with users' similarity calculated by Sequence and Set Similarity Measure (S^3M) with utilizing similarity upper approximation and a Singular Value Decomposition (SVD) model based technique used for recommending ratings for removing sparsity.

Keywords Collaborative filtering · Sequence and set similarity measure (S^3M) · Sparsity problem · Singular value decomposition (SVD)

A. F. Jain (✉) · S. K. Vishwakarma
Gyan Ganga Institute of Technology and Sciences, Jabalpur, India
e-mail: avitadocuments@gmail.com

S. K. Vishwakarma
e-mail: santoshscholar@gmail.com

P. Jain
Jabalpur Engineering College, Jabalpur, India
e-mail: pjain@jecjabalpur.ac.in

S. Fong et al. (eds.), *ICT Analysis and Applications*, Lecture Notes
in Networks and Systems 93, https://doi.org/10.1007/978-981-15-0630-7_14

1 Introduction

The massive growth in digital information in this era of information technology was data is being generating with speed of thought. Availability of internet and glamour of online world have made people not only to overload information about them but also attracting e-commerce companies to understand the behavior of users so that attracting more customers by giving better customized products as well as recommending them for enhancing business profits or revenues of the company. Recommendation system are information filtering system that deal with the problem information overload [8], by filtering vital information fragment out of large amount of dynamically generated information according to user's preferences, interest, or observed behavior about item [12]. Since the recommendation system has the ability to predict the preference of any item for a user by understanding and finding similar "User-Item-Ratings" pairs.

The work we are following is a kind of Recommendation system with hybrid filtering where SVD model is applied on S^3M [9] metric which is Set Sequence Similarity Measure; a linear expression of weighted parameter "p" to emphasize on sequence similarity and relative variance to calculate Jaccard similarity. By varying the value of 'p' we compare the recommendation value. The contents we are experimenting is website ratings from dataset based on the websites visited by users. The major problem in collaborative Recommendation system is sparsity. The work we are performing to achieve sparsity reduction by suggesting recommendation ratings 'R_i'. A comparison is made on recommended rating outcome by varying weighted variable 'p' and 'q' used for quantitative and/or qualitative parts of similarity linear equation used in S^3M [9]. In this paper the work is focusing on web usage mining along with user item rating. The dataset used has been created based on user rating for websites.

S^3M similarly which is used to find similarity among users and then rough set based upper approximation for clustering is applied for forming soft clusters. The technique generate overlapping clusters contains common users which are common is showing interests/rating to multiple websites. We have compared our method performance with various conventional methods. The paper follows the work done on web recommendation by Mishra et al. [10]. The recommendation system is different from sequential pattern mining algorithms. In our work we have studied the variation in 'p' where we consider content (quantitative) not the sequential information (qualitative) as by Mishra et al. [10]. The users similarity is calculated by S^3M, SVD is used for sparsity removal and the R_i is calculated on basis of prediction quotient formula.

Our work has been initiated with the major issues of recommender system. One of the premier work is to remove sparsity problem in Recommender System with collaborative filtering. Mostly web users have a sequential approach for webpage accessing. The algorithm is motivational outcome of work done on web recommendation by Mishra et al. [10], which is actually web content mining and sequence mining based algorithm where the data is a sequential data. The prime work i.e.

the similarity calculation and prediction weight calculation both are adhere to the sequential data and the sequential behavior of web sessions.

The method proposed here is novel with respect to the user-item rating matrix as data. The similarity is calculated with partial consideration on sequential nature and more on the content by changing the value of '*p*'. Also the prediction is made not only on the basis of cluster of similar users but also the calculation lies on the similar user with a specific rating given by new user for a particular website. The major contribution is that in order to remove sparsity the prediction vector is proportional to the rating value by similar user in the cluster. The selection of user for computing prediction quotient is also proportional to the rating. As the collaborative filtering primarily work on user-item ratings, above mentioned two levels of rating based prediction makes the work more promising to for accuracy.

The paper organization begins with introduction of the work with basic idea of proposed method, the motivation and contribution of our work, followed by the literature review with respect to understand the sparsity problem and the work done to alleviate this problem. Third section explains the methodology adopted and dataset. The paper concludes the performance of method in the last section.

2 Related Work

The evolution of Recommendation system begun with research paper on Recommendation system where it is defined as a decision making strategy for users under complex information environments [14]. Recommendation system also used as an E-commerce tool helps users filter through records of knowledge which is related to user's interests and preferences [16]. The phases of Recommendation system are information collection, learning or filtering phase, Recommendation phase. In the filtering phase approaches may be content based technique which predict on basis of user's information ignoring contribution of others users, collaborative filtering when user item rating are used from other similar users. Hybrid filtering approaches by harnessing benefits of both techniques [18]. The matrix is generated by user preferences or likes for items, finds similar users based on relevant interests. This approach is of two types memory based and model based [1, 5, 6].

Memory based collaborative filtering computers similarity between user item ratings. The algorithm of memory based systems is heuristics that make recommendations based on an entire collection of item already rated by users [4, 11, 15]. Model based collaborative filtering generates the descriptive model of the system, based on the user's preferences using various DM and ML techniques like Bayesian Model, Clustering Model etc.

Recommendation system faces some problem with respect to efficiency of recommendation system; recommendations quality is affected by cold start problem, data sparsity problem, scalability, synonymy and Matthew effect [2, 3, 13, 17]. These problems are reducing commercial benefits to an extent. The sparsity problem is lack

of enough information when only a few items ratings by users in the matrix. Data sparsity problem directly affects the coverage of recommendation result [17]. This makes the matrix sparse which in turn disables to locate proper neighbours which finally leads to weak recommendations. Model based techniques solve the sparsity problem. This problem also exists in a user product based Product Attribute Model which is due to the subjectivity of product reviews since these reviews are not covering all aspects of product. The problem is resolved by Multiplication Convergence Rule and Constraint Condition equations to find the replacement of sparse values [19]. A web recommendation system which works on sequential mining and web mining also applies a weight calculation which adequately leads to substitute the next web page visit vector which is a sparse vector, has been proposed by Mishra et al. [10]. S^3M [9] is the similarity measure applied to set sequence similarity proposed by Kumar et al. [15] which is a linear equation based on the weighted parameter 'p' Quality is determined by SetSeq(A,B) measure and quantity is calculated by SetSim(A,B) for content matching. S^3M is used to find similar users. Singular value decomposition is a model based technique. This deploys the previous ratings (user-item) to improve the performance of Collaborative Filtering Technique. SVD is used in Opt.space algorithm by Keshavan et al. [7] to deal with matrix completion problem.

In this paper we follow the work done on web recommendation by Mishra et al. [10]. The recommendation system is different from sequential pattern mining algorithms. The users similarity by S^3M metric and the SVD model is used for removing sparsity significantly and the R_i is calculated on basis of formula proposed by Mishra et al. [10].

3 Methodology

The work in the paper focuses on results which are recommendation generated using SVD model on soft clusters which are made on similarity of users by calculating S^3M similarity Mishra et al. [10]. Different values of 'p' are taken and compared the result of "R_i" by the model proposed by Mishra et al. [10].

The methodology as shown in Fig. 1 is devised in three parts; Response Matrix 'A' Generation, Prediction Quotient Q_{ij} Calculation and Recommendation Vector 'R_i' Generation.

In our work experiments performed on dataset which is generated manually from the survey done in the crowd of university UG program students regarding the websites they visit and asked them to rate. The generated dataset has ratings of websites,

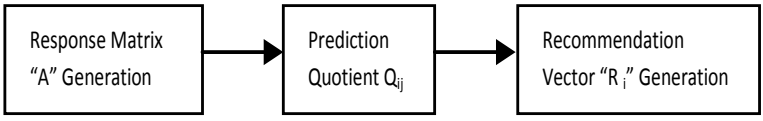

Fig. 1 Methodology

category of interested websites along with preferred website sequence partially The dataset is represented as matrix of user—website ratings. The user's ratings for websites ranges from 1 to 5 where 5 are highest and 1 is lowest rating order. The user—item-rating matrix is being developed where item a frequently visited websites is. The similarity of users from data set is found with full utilization of content similarity and partial consideration of order of Websites (a sequence) visited by user.

3.1 Response Matrix 'A' Generation

The work proceeds with formation of cluster which are soft clusters on the basis of website since a user may have multiple interest for which may belong to multiple clusters. A similarity upper approximation based clustering algorithm is used. The RS utilized a rough set based clustering approach. The similarity between users is calculated by similarity measure (metric). There exists many similarity metrics such as cosine, Jacquard etc. The metric used by Kumar et al. [15] is set sequence similarity measure. This measure enforces not only the similarity between two sets of data (vector or ordered set) but also considers the sequence (Fig. 2).

The algorithm of similarity upper approximation approach for cluster formation we are following is the same as per proposed by Mishra et al. [15]. The similarity of user is being measured here on the basis of S^3M [9]. The S^3M is Sequence Set Similarity Measure Kumar et al. [15] calculated as following:

$$S^3M[A, B] = p * \text{SeqSim}(A, B) + q * \text{Setsim}\ (A, B) \tag{1}$$

where 'p' is qualitative weight parameter of sequence similarly [9] and $q = (l - p)$ i.e. after sequence similarity content similarity is focused.

The following example explains similarity calculations. Let two users along with their preferred website rating sequence is shown as:

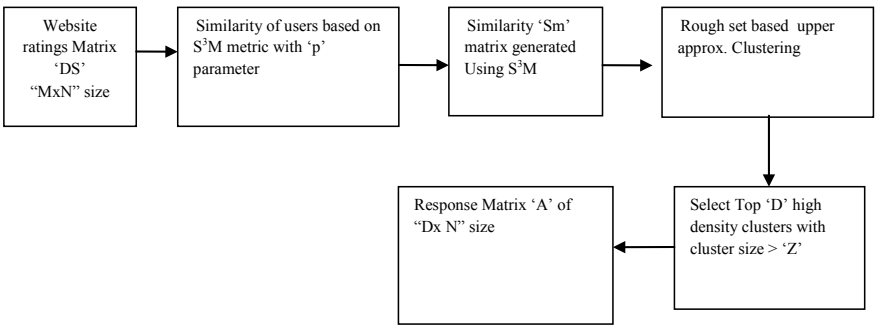

Fig. 2 Response matrix '*A*' generation

$U_A = \{1,\ 4,\ 18,\ 20,\ 11,\ 15,\ 6,\ 8,\ 5,\ 12\}$ and $U_B = \{4,\ 5,\ 11,\ 8,\ 1,\ 2,\ 3,\ 7,\ 9,\ 18,\ 15,\ 20\}$

$$\text{The length of sets}: L_A = |U_A| = 10, \quad L_B = |U_B| = 12$$

$$\text{And LLCS}(U_A, U_B) = \{4,\ 11,\ 8\} = 3$$

$$\begin{aligned}\text{So SeqSim}(U_A,\ U_B) &= \text{LLCS/Max}(|U_A|, |U_B|)\\ &= 3/12 = 0.25\end{aligned}$$

$U_A \cap U_B = \{1,\ 4,\ 5,\ 8,\ 11,\ 18,\ 15,\ 20\}$ and $U_A \cup U_B = \{1,\ 2,\ 3,\ 4,\ 5,\ 6,\ 7,\ 9,\ 11,\ 12,\ 15,\ 18,\ 20\}$

$\text{SetSim}(U_A U_B) = |U_A \cap U_B|/|U_A \cup U_B| = 8/14 = 0.57$

Let "Sm" be a similarity matrix (Table 1) such that $\text{Sm}[i,j] = \alpha_{ij}$ where α is the similarity measure value between users U_i and U_j. The value of $\alpha = 1$ for all $i = j$. The value of 'p' is 0.7 for calculation of α in the following Table 1 from the formula of S^3M Eq. (1).

The classification of web users Response matrix A is formed by selecting Top "D" clusters where we choose the higher density cluster such a way that the cluster density of selected clusters is 'Z' where "Z" > avg_cluster_density. The response matrix has clusters 'D' as rows and items as the columns will be the total 'N' (all) websites rated by users. The row vector of matrix 'A' will the average rating of respective website by the users of clusters. It is represented as $A_i = \{A_{ij}$: Average rating of W_j by all the users of $C_j\}$. The matrix A is of size $D \times N$ where 'D' is number of the top high density cluster and 'N' is the total websites rated.

3.2 Prediction Quotient 'Q' Generation

After constructing response matrix, now it works to make a Prediction Quotient for new users who provide ratings for a few websites. A new user who provides a small pattern of ratings for some websites is the base for finding the similar users and to classify the respective cluster C_k. The prediction quotient is calculated by user ratings ratio/proportions. It is used for a new user who provides a small pattern of ratings for some websites. With the following calculations a recommended Prediction Quotient is calculated for a new user.

$$\text{The Prediction Quotient: } Q_{ij} = \frac{W_{ij}}{W_i} \tag{2}$$

where W_{ij}—total number of times ith website rated with "j" as the rating value in the cluster C_k.

W_i—total number of times ith website has been rated in the cluster C_k.

Table 1 Similarity matrix 'S_m' of users' of dataset

S_m	U_1	U_2	U_3	U_4	U_5	U_6	U_7	U_8	U_9	U_{10}
U_1	1	0	0	0	0.21	0.29	0	0	0	0
U_2	0	1	0	0.47	0.17	0.17	0.17	0	0.15	0.15
U_3	0	0	1	0	0	0.25	0.33	0.33	0	0.21
U_4	0	0.47	0	1	0.17	0	0.45	0.27	0.24	0.5
U_5	0.21	0.17	0	0.17	1	0.18	0	0	0	0
U_6	0.29	0.17	0.25	0	0.18	1	0.18	0.21	0	0.17
U_7	0	0.17	0.3	0.45	0	0.18	1	0.58	0.17	0.62
U_8	0	0	0.3	0.27	0	0.21	0.58	1	0	0.5
U_9	0	0.15	0	0.24	0	0	0.17	0	1	0.24
U_{10}	0	0.15	0.21	0.4	0	0.17	0.62	0.5	0.24	1

And if $W_i = 0$ then $Q_{ij} = 0$. A recommended Prediction Quotient vector is formed by placing Q_{ij} values in place of ith website rated by new user and the unknown values are filled with '×'. The length of the vector is 'N' the number of websites rated. Prediction Quotient vector is Ordered Set of Q_{ij} where ith website rated by new user and unknown website ratings = '× ', of 'N' elements. The vector is used with the output matrices of SVD applied on response matrix 'A'.

3.3 Recommendation Vector 'R' Generation

The response matrix 'A' is applied to Singular Value Decomposition (SVD) model which produces three matrices U, S, V^N. U is of size $D \times N$, S is of size $N \times N$, and the matrix V^N is also $N \times N$. The matrix S has diagonal elements as non-zero values other elements will be zero.

4 Result

The dataset has been created with the university students as described in Sect. 3.3 and consists of more than 5000 feedback records collected through their log files of web access. For experimentation purpose, RapidMiner the open source research data mining tool has been used. The unknown (sparse) ratings are the ratings of websites which have not been rated.

We have also compared performance results of our method with other collaborative recommendation system methods such as BMF, KNN and Slope-one. The experiments have been carried with the RapidMiner Data Mining tools in experiments. Our Method is very similar to Factor-Wise Matrix factorization method provided by RapidMiner FWMF operator performs. The following figures are the comparative diagram of the proposed method with the traditional methods.

The Figs. 3, 4, 5, and 6 are the comparative results of the proposed method with traditional methods. The tool RapidMiner uses following operators for experimenting the performance of proposed Recommender System Matrix factorization with factor-wise learning (FWMF) operator performs modeling Relationships at Multiple Scales to Improve Accuracy of Large Recommender Systems. The Bias Matrix Factorization (BMF) operator performs Matrix factorization with explicit user and item bias. This operator uses bold driver heuristics for learning rate adaption and supports Large-Scale Matrix Factorization with Distributed Stochastic Gradient Descent method.

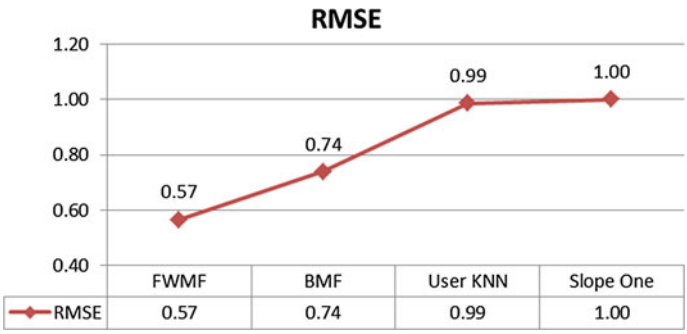

Fig. 3 Root mean square error

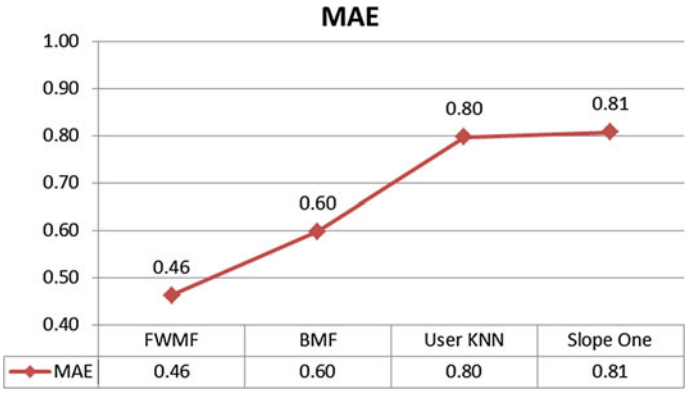

Fig. 4 Mean absolute error

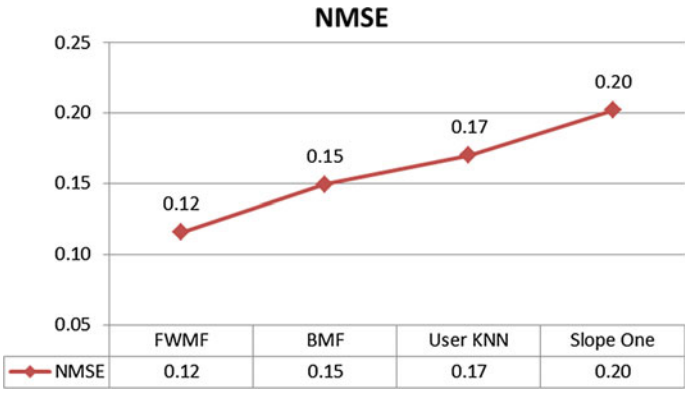

Fig. 5 Normalized mean square errors

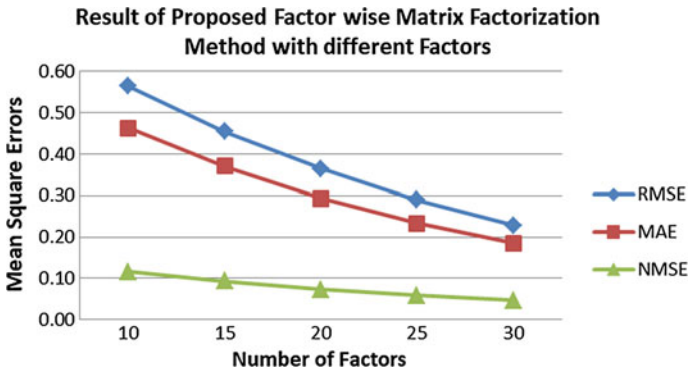

Fig. 6 Proposed method with different factors

5 Conclusion

This paper proposed a novel approach for collaborative recommender system based on S^3M approach of item and user ratings. The calculation for filling up zeros lies on the similar user with a specific rating given by new user for a particular website, is the foundation for generating the prediction vector makes it proportional to the rating at cluster level and as well at user level. The proposed method outperforms with all methods and gives the minimum errors with respect to RMSE, MAE, NMSE. The proposed method has been compared with different factors and it gives the minimum RMSE values for the ratings predictions.

References

1. Buder J, Schwind C (2012) Learning with personalized recommender systems: a psychological view. Comput Hum Behav
2. Burke R (2002) Hybrid recommender systems: survey and experiments—user modeling and user-adapted interaction. Springer
3. Burke R (2007) Hybrid web recommender systems—the adaptive web. Springer
4. Delgado J, Ishii N (1999) Memory-based weighted majority prediction—SIGIR Workshop Recomm. Syst …, 1999—pdfs.semanticscholar.org
5. Gadanho SC, Lhuillier N (2007) Addressing uncertainty in implicit preferences. In: Proceedings of the 2007 ACM conference, 2007—macs.hw.ac.uk
6. Herlocker JL, Kostan JA, Terveen LG, Riedl JT (2004) Evaluating collaborative filtering recommender systems. ACM Trans Inf Syst 22(1):553
7. Keshavan RH, Montanari A, Sewoong O (2010) Matrix completion from a few entries. IEEE Trans Inform Theory
8. Konstan JA, Riedl J (2012) Recommender systems: from algorithms to user experience—user modeling and user-adapted interaction. Springer
9. Kumar P, Raju BS, Krishna PR (2010) A new similarity metric for sequential data—warehousing and mining (IJDWM), 2010—igi-global.com

10. Mishra R, Kumar P, Bhasker B (2015) A web recommendation system considering sequential information. Decis Support Syst
11. Nakamura A, Abe N (1998) Collaborative filtering using weighted majority prediction algorithms. In: Proceedings 15th international conference. Machine Learning 1998
12. Pan C, Li W (2010) Research paper recommendation with topic analysis. In: International conference on computer …, 2010—ieeexplore.ieee.org
13. Park DH, Kim HK, Choi IY, Kim JK (2012) A literature review and classification of recommender systems research—expert systems with applications. Elsevier
14. Rashid AM, Albert I, Cosley D, Lam SK (2002) Getting to know you: learning new user preferences in recommender systems. In: Proceedings of the 7th …, 2002—dl.acm.org
15. Resnick P, Iakovou N, Sushak M, Bergstrom P, Riedl J (1994) An open architecture for collaborative filtering of netnews/PR GroupLens. In: Proceedings of the 1994 computer supported cooperative work …, 1994
16. Schafer JB, Konstan J, Riedl J (1999) Recommender systems in e-commerce. In: Proceedings of the 1st ACM conference on …, 1999—emunix.emich.edu
17. Wang H, Wang Z, Zhang W (2010) Quantitative analysis of Matthew effect and sparsity problem of recommender systems. In: 2018 IEEE 3rd International …, 2018—ieeexplore.ieee.org
18. Isinkaye FO, Folajimi YO, Ojokoh BA: Recommendation systems: principles, methods and evaluation. Cairo University Egyptian Informatics Journal, Elsevier
19. Yang X, Zhou S, Cao M (2019) An approach to alleviate the sparsity problem of hybrid collaborative filtering based recommendations: the product-attribute perspective from user reviews-mobile networks and applications. Springer