International Journal of Electrical and Computer Engineering (IJECE)
Vol. 10, No. 1, February 2020, pp. 447~454
ISSN: 2088-8708, DOI: 10.11591/ijece.v10i1.pp447-454

An enhanced kernel weighted collaborative recommended system to alleviate sparsity

S. Babeetha[1], B. Muruganantham[2], S. Ganesh Kumar[3], A. Murugan[4]
[1]Department of Information and Technology, SRM Institute of Science and Technology, India
[2]Department of Computer Science and Engineering, SRM Institute of Science and Technology, India
[3,4]SRM Institute of Science and Technology, India

Article Info

Article history:

Received Feb 17, 2019
Revised Aug 16, 2019
Accepted Aug 28, 2019

Keywords:

Classical rating prediction methods
Collaborative filtering (CF)
Kernel CF
K-Weighted clustering
Rating estimation methods
Recommender system (RS)
Sparsity

ABSTRACT

User Reviews in the form of ratings giving an opportunity to judge the user interest on the available products and providing a chance to recommend new similar items to the customers. Personalized recommender techniques placing vital role in this grown ecommerce century to predict the users' interest. Collaborative Filtering (CF) system is one of the widely used democratic recommender system where it completely rely on user ratings to provide recommendations for the users. In this paper, an enhanced Collaborative Filtering system is proposed using Kernel Weighted K-means Clustering (KWKC) approach using Radial basis Functions (RBF) for eliminate the Sparsity problem where lack of rating is the challenge of providing the accurate recommendation to the user. The proposed system having two phases of state transitions: Connected and Disconnected. During Connected state the form of transition will be 'Recommended mode' where the active user be given with the Predicted-recommended items. In Disconnected State the form of transition will be 'Learning mode' where the hybrid learning approach and user clusters will be used to define the similar user models. Disconnected State activities will be performed in hidden layer of RBF and Connected Sate activities will be performed in output Layer. Input Layer of RBF using original user Ratings. The proposed KWKC used to smoothen the sparse original rating matrix and define the similar user clusters. A benchmark comparative study also made with classical learning and prediction techniques in terms of accuracy and computational time. Experiential setup is made using MovieLens dataset.

Corresponding Author:

S. Babeetha,
Department of Information and Technology,
SRM Institute of Science and Technology,
Chennai, India.
Email: babees1207@gmail.com

1. INTRODUCTION

Recommendation is one of the majoressentialproduction activities to attract users. RS (Recommendation System) is the approach that will lend a hand in Recommendation systems are mechanism that lend a hand in making selection in an area even with lack ofacquaintance in it. The survey [1] provides the classification of recommender systems.Based on core methods recommender system is broadly classified as heuristic based /content based system and /user preference or rating logic/collaborative filtering system.

Collaborative Filtering recommender systems are based on the resemblance between customer favorite ratings for work out recommendations. Since this method does not rely on semantics of the products, it is complimentary from the two main challenges of the content-based approach:shallow-analysis problem or

the over-specialization problem [2] for which a bad choices of attribute is the root cause.Most frequently used techniques for collaborative recommendation are;correlation based methods [3, 4], Latent Semantic Indexing (LSI) [1, 2], Bayesian Learning [5, 6], etc. The accuracy of the recommendation will be evaluated on the quantity of the user rating provided on the collection of products available.The association to the clusters will be monitored using the product coverage with significant overlaps in ratings. But in reality this will not be the case, where there will be lack of large customer pool and trend of introducing new products, thus causing Sparsity Problem.If the overlap user combination is very low, then the situation will lead into poor recommendation [7].

a. Content-based System: Predictions based on the items which are similar to their content or semantics.
b. Collaborative System: Calculations are based on the similar rating opted previously be the users.
c. Hybrid System: The combination of content based and collaborative filtering methods form Hybrid recommender systems.

The hub of this paper isprimarilyon item based rating prophecy which areCollaborative Filtering recommendations; where it provides the scheming loom to recommender systems.The aim of Collaborative Filtering (CF) system on predicting a user's mind on collective itemsexists in the area, using users' previous available opinions or rating on items.Collaborative filtering is the widely used prediction approach [1]. The Classic CF approach works using theory of predicting the user rating using the user-rating matrix [8].

In this paper, Kernel Weighted K-means Clustering (KWKC) approach is used in Radial basis Network (RBN) Layer to alleviate the sparsity problem. The proposed two phases will make the prediction process as an incremental approach, where each state will increase the accuracy of the predication. The first state 'Disconnected' will smoothen the Sparse data matrix using KWKC and the second state 'Connected' will provide the recommendation to the current user. Experimental setup is made using MovieLens Data set and the results are compared with classic systems like Single Value decomposition (SVD), Support Vector machines (SVM) and KFCM (Kernel Fuzzy C-Means). The proposed systems proved with comparatively high accuracy and quality.

The rest of this paper is organized as: discussions on existing similarity finding techniques of CF in Section 2, proposed system and architecture flow Algorithms used at every stage of proposed system in Section 3, experimental results in Section 4 and Section 5 concludes the system along with Future extensions.

2. RELATED WORK

As per [9], there are two types in Collaborative Filtering system:model(User/item) and memory based (or heuristic-based). Memorybased procedure [9-11] primarily are semanticthe prediction relies on the completeset of items ratedby customers beforehand.The memory-based CF algorithm uses the following steps:

Step 1: Finding Similarity using similarity finding techniques like Pearson correlation or Cosine similarity.
Step 2: Predict Rating for the current active user on Top N items found.

Most of the time, the quantified amount of rating will not be available to obtain the aspired recommendations.This unsatisfactory number of ratings causes thesparsity problem [12, 13]. In [1, 2], a dimensionality reduction methodology for tackling with inadequate rating matrix was discussed,Singular Value Decomposition [14] is a sound method on matrix factorization that gives lower rank estimations for original sparse matrix [1].

The two possible solutions available for addressing sparsity problem are: the first one uses smoothen [15, 16] the sparse matrix or shrinking the dimension forminimizing the sparsity of rating matrix. The Second resolution provides the improved methods to increase the efficiency without changing the sparsity of the matrix [17]. For Smoothening [15, 18] which will reduce the sparsity, Radial Basis Function Network is used by providing the approximate rating for meager Kernel Fuzzy C means clustered matrix. To predict the missing value in less dense rating matrix, a two step solution is introduced in [8, 19] using Co-clustering and Radial Basis Function. In this situation, ifa new user is not rated sufficient number of items properly then the process of predicting the recommendations to that user is difficult job to the system. The cold start problem (new user problem) [12, 20] will be raised, if there is no match between the new user's ratings with the previously present users.

This problem be present in bunch of industrial domains. Cold start is defined as, the situation where either there are not adequate data to examine or the user are meager. E.g. a new active user in an online community webpage, the user doesn't have even a single companion or likely item, and it's not easy to provide recommendations to such user. There comes the major nature of cold start problems in CF, New User or New Item.

Hybrid CF systems, like content enhanced CF technique [21], used to address the sparsity problem; where exterior content can be used to generate missing ratings for new user/item. During the RBF

recommendation approach [8, 18], if the activeuser is new to this system or application without having rating (New user cold start problem); the top rated items on very cluster is suggested for that active user.

The Model-based [3, 4, 22] CF approach addresses these problems using the set of ratings to be trained a model and it will be utilized to predict rating. The likelihoodthat userprate the particular item *i* given that previous rating of user on the rated items, is usually determined by two probabilistic models: cluster models and Bayesian networks [17]. The basic Bayesian collaborative Filteringalgorithm [17] uses a NB (naive Bayes) stratagemon predictions. With the given isolatedclasses, the class having peek probability will be categorizedas the predicted class [1]. Place the base Bayesian algorithm to diversedata for CF objectives [17], produces predictiveaccuracyin worse manner with advisable scalability than the Pearson correlation.

Clustering Collaborating Filtering memory basedApproach: ACluster is defined as aset of data itemswhich are correlated towards each other within the current cluster and are unconnectedto the items amongrest of theclusters [23]. Legacy collaborative filtering having less scalabilitythan Clustering models ,because they make predictions within comparatively small clusters which is known us reduced dimension data set [7, 21, 24]. As the user base grows in volumes, then User–user collaborative Filter, while effective, suffer from scalability issues. Top-N item based recommendation methods will be used to alleviate the scalability problem of top-N user based recommendation [25]. Instead of finding user-user similarity, it proposes to calculate similarities of two sets and then sort the similarity values in decreasing order [26].

3. PROPOSED SYSTEM USING KWKC

The proposed system is shown in Figure 1. As shown, the phase 'Disconnected' will perform the learning activity from the existing rating matrix where three major steps will be carried out:
Step1: UCC (Unsupervised Correlation Clustering)
Step2: Smoothening using Euclidean Norm
Step3: KWKNN clustering (Kernel Weighted K-means Nearest Neighbour)
The phase 'Connected' will perform the online recommendation to the current active user.
Step4: Predict the recommendations for current user

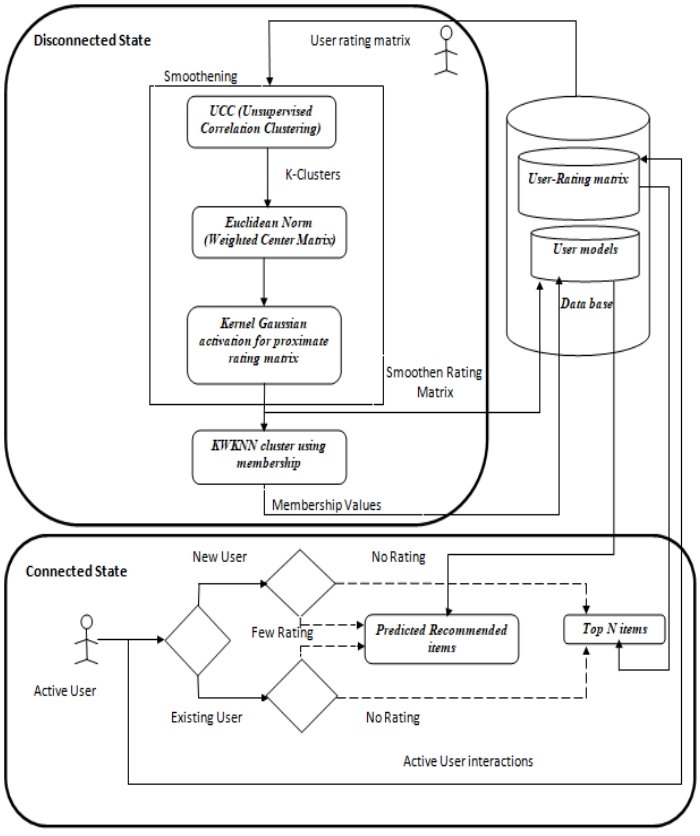

Figure 1. KWKC approach

3.1. Disconnected state

3.1.1. Algorithm1: UCC (unsupervised correlation clustering) phase

Input: $R_{c,i}$ is the Sparse Customer-Item rating Matrix. Average Rating of users on item $'i'$ is $\bar{R}_i$

Output: k- number of user groups- Called Clusters

Steps:

1. For x=1 to N // Similarity is going to be calculated among N users //

1.1. Similarity_Index (x) =0 //Initialize by Zero which will determine the similar users

1.2. For y=1 to N

//Find Similarity between users (u_i, u_j)using Pearson correlation function//

$$sim(p,q) = \frac{\sum_{u\epsilon U}(R_{u,p}-\bar{R}p)(R_{u,q}-\bar{R}q)}{\sqrt{\sum_{u\epsilon U}(R_{u,p}-\bar{R}p)^2}\sqrt{\sum_{u\epsilon U}((R_{u,q}-\bar{R}q)^2}} \quad (1)$$

where p and q are users, U is the item collection rated by users p & q, $\bar{R}$ is average rating of user (p or q), $R_{u,q}$ user u 's rating on q^{th} item.

1.3. If $sim(p,q)$=1 then Similarity_Index (x)= Similarity_Index (x)+1

2. Fetch k users' whose Similarity_Index value is the highest among N. Assign those k users as cluster Centers.

3. Club the rest of the users (N-k)to corresponding cluster, where the $sim(p,q)$ =1. (if p –is cluster center then $q \in (N-k)$)

3.1.2. Algorithm 2: Smoothening using euclidean norm

Input: Sparse Customer-Item rating Matrix.($WhereratingR_{c,i}|1 \le c \le N, 1 \le i \le N'$);N - Total number of users, N'- Total Number of Items.$\varphi_i^{Min}$and φ_i^{Max} are Minimum and Maximum Values of the activation or Objective function of a particular parameter i (where i can be either item or user).

Output: Smoothened rating Matrix $SR_{c,i}$($WhereratingSR_{c,i}|1 \le c \le N, 1 \le i \le N'$)

Steps:

1. Collection=(Rating_Max-Rating-min)+1
2. Derive k clusters such that $\lfloor Collection/k \rfloor \le 3$
3. For i=1 to N'

3.1. Separate the users into k clusters using algorithm 1. ($WhereratingR_{c,i}|1 \le c \le N, 1 \le i \le N'$)

3.2. Calculate Euclidean centers

$$C_k = \frac{\sum_{c'=1}^{k_u} R_{c',p}}{k_u} \quad (2)$$

where $k_u$is representing number of users within a cluster.

3.3. Find the Euclidean distance matrix $D_{c',p} = \|R_{c',p} - C_k\|$ where $1 \le p \le k_u$

3.4. Compute the objective function $\varphi_{c',p}(D)$ using Gaussian positive definite function

$$\varphi(x) = \exp(-\frac{x^2}{2\sigma^2}) \text{ where } \sigma > 0 \quad (3)$$

3.5. Determine the weights using pseudo random weight function

$$w_{c,i} = \frac{(\varphi_i^{Max} - \varphi_i(R_{c,i}))/(\varphi_i^{Max} - \varphi_i^{Min})}{\sum_{x=1}^{n}(\varphi_{x,i}^{Max} - \varphi_{x,i}(R_{x,i}))/(\varphi_x^{Max} - \varphi_x^{Min})} \quad (4)$$

3.6. Calculate $\widetilde{R_{c,i}} = F(R_{c,i})$ where

$$F(R_{c,i}) = \sum w_{c,i}\, \varphi(\|R_{c,i} - C_k\|) \quad (5)$$

is the activation function and$WhereratingR_{c,i}|1 \le c \le N, 1 \le i \le N'$,$C_k$ is the cluster center.

3.7. Derive the Smoothened rating Matrix $SR_{c,i}$

$$SR_{c,i} = \begin{cases} R_{c,i} if user crate the item i \\ \widetilde{R_{c,i}} else \end{cases} \quad (6)$$

3.1.3. Algorithm 3: KWKNN clustering (kernel weighted k-means nearest neighbour)

Input: Smoothened rating Matrix $SR_{c,i}$

Output: KWC_n Kernel weighted Cluster centers, membership degree $m_{c,cl}$ for user c in cluster cl. The membership range will be 0-1 which indicates the interest of user on the cluster.

Steps:

1. Initialize $m_{c,cl} = randomprimenum(j) \Big/ \sum_q randomprimenum(j)$

   set q=1, which determine the number of error cycles.
2. Calculate Kernel weighted Cluster centers

$$KWC_n = \frac{\sum_{i=1}^{N'} m_{c,n}^{q} * SR_{c,i}}{\sum_{i=1}^{N'} m_{c,n}^{q}} \quad (7)$$

3. where $1 \leq n \leq k$
4. Compute membership value

$$m_{c,n} = \frac{1}{\sum_{cl=1}^{k} \left(\frac{SR_{c,i} - KWC_n}{SR_{c,i} - KWC_{cl}}\right)^{\frac{2}{cl-1}}} \quad (8)$$

5. Calculate Error index=max(old $m_{c,n}$, new $m_{c,n}$)
6. If Error index >0.5, then q=q+1; go to step 2

3.2. Connected state

Algorithm 4: Predict the recommendations for current user (online activity)

Input: Current active user Rating array R_{ca} and P_{ri} denotes the predicted recommended items, Kernel weighted Cluster centers KWC_n and $\varphi(x)$ the Gaussian activation function.

Output: Recommended Items for the current user

Steps:

1. If the current user is a new user and the Rating array is NULL (New user cold start), then provide Top N rated items from each cluster to the recommender engine.
2. If the current user is a new user and have rated few items (and also Rating array is not NULL) then perform below steps followed by sharing of Top N rated items from each cluster to the recommender engine.

2.1. Calculate degree of membership $m_{a,cl}$ with all clusters $1 \leq cl \leq k$

2.2. Predict the unknown rating of the current active user

$$\boldsymbol{P_{a,i}} = \frac{\sum_{\boldsymbol{cl=1}}^{\boldsymbol{k}} m_{a,cl} \varphi_i(R_{ca}, C_{cl}) C_{cl,i}}{\sum_{\boldsymbol{cl=1}}^{\boldsymbol{k}} m_{a,cl} \varphi_i(R_{ca}, C_{cl})} \quad (9)$$

3. If the current user is an existing user and with no added rating then provide Top N trained phase predicted rating to the recommender engine.
4. If the current user is an existing user and with added new ratings the go to step 2.

4. EXPERIMENTAL SETUP AND RESULTS

4.1. Test set

The experimental setup was made using MovieLens data set with, 100,000 user ratings from 1000 odd users for 1783 movies. For comparative purpose, the data is split with different level of sparsity (20% to 90%). To analyze the results the data set divided into two sets: training dataset and testing dataset.

4.2. Performance measures

4.2.1. Mean absolute error (MAE)

MAE is the traditional classification accuracy to measure how close the recommender systems predicted ratings are to the true user ratings (for all the ratings in the test set). In our experiments, we calculate the accuracy for existing CF algorithms and KWKC using Mean Absolute Error (MAE) as shown in Figure 2 using (10).

$$MAE = \sum_{i=1}^{N}\sum_{j=1}^{M}\frac{|\,p_i - r_i\,|}{N*M} \quad (10)$$

where, N is the number of Items, M is the number of Users, p_i is the Predicted rating and r_i is the actual or original rating given by the user initially.

4.2.2. Precision and recall

The second measure is the relevant recommendations measure using Precision and Recall. Precision and recall are the most popular metrics for evaluating information retrieval systems. Precision is the ratio of relevant items selected by the recommender to the number of items selected using (11). Recall is the ratio of relevant items selected to the number of relevant using (12).

$$\mathrm{Pr}ecision = \frac{\sum_{i=1}^{N}\frac{N_s}{N_s + N_{is}}}{N} \quad (11)$$

$$\mathrm{Re}call = \frac{\sum_{i=1}^{N}\frac{N_s}{N_s + N_{rn}}}{N} \quad (12)$$

where the values are taken as shown Table1.

Table 1. Item categories

	Selected	Not Selected	Total
Relevant	N_{rs}	N_{rn}	N_r
Irrelevant	N_{is}	N_{in}	N_i
Total	N_s	N_n	N

5. RESULTS AND DISCUSSION

5.1. Checkpoint for computational error

To investigate the Mean Absolute Error, the primary techniques like SVD (Singular Value Decomposition), SVM (Singular Vector Machine), RBFN (Radial Basis Functional Network) with Pearson correlation and Kernel Fuzzy C-Means (KFCM) of CF are compared with the proposed algorithm KWKC. The error computation made for both in Disconnected phase (Modeling time) and connected phase (Prediction time) which are shown in Figure 2.

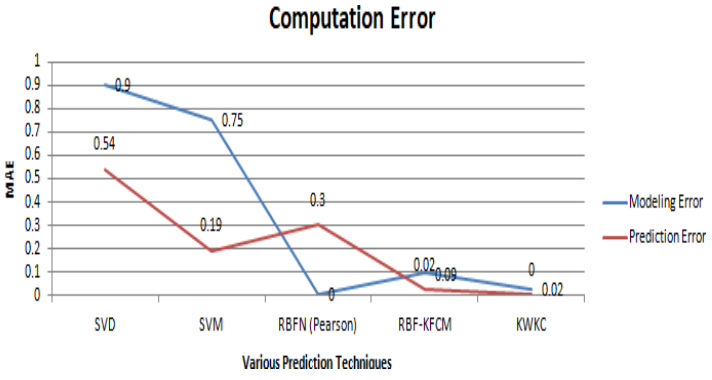

Figure 2. Computational error with MAE

a. SVD shows high error value both of the computational phases.
b. Albeit RBFN shows Nil error on Modeling timing, the proposed KWKC algorithm also shows negligible error which is very closure to Nil.
c. The proposed algorithm shows the improvised results compared to rest of RBF techniques.

5.2. Check point for Relevance in terms of efficiency

The precision and Recall percentage is near more than 90 percentage, which is comparatively high when compared to rest of benchmarked primary CF Techniques. Figure 3 shown relevance measure using precision and recall.

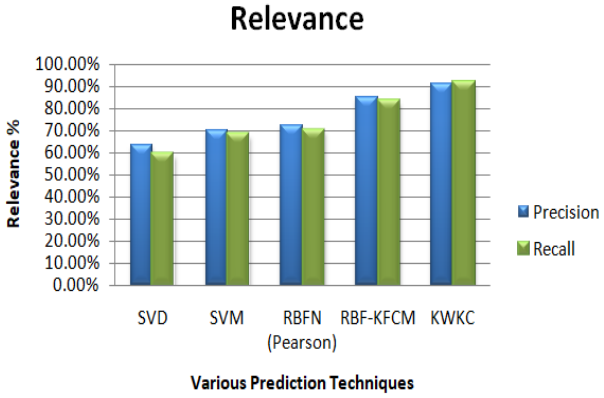

Figure 3. Relevance measure using Precision and Recall

6. FUTURE WORK AND CONCLUSION

One of the best way to mining the required products in the ecommerce word is Recommender system. If the recommendations based on reality of the user tastes then the accuracy will be unbelievable. Albeit the primary techniques are benchmarked, there are jeopardy where the volume increases but the acquired user rating decreases. In the proposed system KWKC, both accuracy and efficiency is proven on the comparison made with experimental setup with user data. Comparatively better results shown on the proposed system. A future enhancement is planned to extend the research of grouping of Items clusters along with User clusters to provide better accuracy.

REFERENCES

[1] G. Adomavicius and A.Tuzhilin, "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, 2005.
[2] G. Adomavicius, *et al*., "Incorporating Contextual Information in Recommender Systems Using a Multidimensional Approach," *ACM Trans. Information Systems*, vol. 23, 2005.
[3] G. Adomavicius and A. Tuzhilin, "Expert-Driven Validation of Rule-Based User Models in Personalization Applications," *Data Mining and Knowledge Discovery*, vol. 5, pp. 33-58, 2001a.
[4] G. Adomavicius and A. Tuzhilin, "Multidimensional Recommender Systems: A Data Warehousing Approach," *Proc. Second Int'l Workshop Electronic Commerce (WELCOM '01)*, 2001b.
[5] J.L. Herlocker, *et al*., "An Algorithmic Framework for Performing Collaborative Filtering," *Proc. 22nd Ann. Int'l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR '99)*, 1999.
[6] J.L. Herlocker, *et al*., "Explaining Collaborative Filtering Recommendations," *Proc. ACM Conf. Computer Supported Cooperative Work*, 2000.
[7] J.L. Herlocker and J.A. Konstan, "Content-Independent TaskFocused Recommendation," *IEEE Internet Computing*, vol. 5, pp. 40-47, 2001.
[8] M.K.Devi and P.Venkatesh, "An Improved Collaborative Recommender System," *2009 First International Conference on Networks & Communications*, 2009.
[9] A. Ansari, *et al*., "Internet Recommendations Systems," *J. Marketing Research*, pp. 363-375, 2000.
[10] C.C. Aggarwal, *et al*., "Horting Hatches an Egg: A New Graph-Theoretic Approach to Collaborative Filtering," *Proc. Fifth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining*, 1999.
[11] Xiaoxia Sun and Nasser M. Nasrabad "Task-Driven Dictionary Learning for Hyperspectral Image Classification With Structured Sparsity Constraints", IEEE Transactions On Geoscience And Remote Sensing, vol. 53, no. 8, pp. 4457 - 4471 august 2015.
[12] B. V.Arvind, *et al*., "An improvised filtering based intelligent recommendation technique for web personalization,"*India conference (INDICON) 2012, Annual IEEE*, 2012.

[13] IlkerBayram "Sparsity Within and Across Overlapping Groups" , IEEE Signal Processing Letters, vol. 25, no. 2, pp. 288 – 292 february 2018.
[14] M. Pryor, "The effects of singular value decomposition on collaborative filtering," *Computer Science Depart., Dartmouth College, Hanover, NH, Tech. Rep. PCS-TR98-338*, 1998.
[15] M.K.Devi and P.Venkatesh, "Smoothening approach to alleviate the meager rating problem in collaborative Recommender System,"*Future generation computer systems,*vol. 29, pp. 262-270, 2013.
[16] XiangrongZeng and Mario A. T. Figueiredo "Robust Sparsity And Clustering Regularization For Regression," 22nd European Signal Processing Conference (EUSIPCO), pp1776–1780.
[17] R.O. Duda, *et al*., "Pattern Classification," John Wiley & Sons, 2001.
[18] M.K.Devi and P.Venkatesh, "IDSS: an intelligent decision support system for e-purchasing using CBR and CF,"*Int. J. Agent-Oriented Software Engineering.*
[19] Youchun J., *et al*.,"Automation Department Xiamen University 'Missing Value Prediction Using Co-clustering and RBF for Collaborative Filtering," *IEEE 2015 International Conference on Cloud Computing and Big Data*, 2015.
[20] X. Su and T. M.Khoshgoftaar,"Review Article A Survey of Collaborative Filtering Techniques,"*Advances in Artificial Intelligence*, 2009.
[21] M.D. Buhmann, "Approximation and Interpolation with Radial Functions," *Multivariate Approximation and Applications*, in N. Dyn, *et al*., Cambridge Univ. Press, 2001.
[22] D. Billsus and M. Pazzani, "User Modeling for Adaptive News Access," *User Modeling and User-Adapted Interaction*, vol. 10, 23, pp. 147-180, 2000.
[23] S.Nagalakshmi, *et al*., "Mathematical Approximation for Model based Recommender System," *Presented at Proceedings of the International Conference on Supply Chain Management and Information Systems*, 2008.
[24] J.L. Herlocker, *et al*., "Evaluating Collaborative Filtering Recommender Systems," *ACM Trans. Information Systems*, vol. 22, pp. 5-53, 2004.
[25] M. Deshpande and G. Karypis, "Item-Based Top-N Recommendation Algorithms," *ACM Trans. Information Systems*, vol. 22, pp. 143-177, 2004.
[26] N. Littlestone and M. Warmuth, "The Weighted Majority Algorithm," *Information and Computation*, vol. 108, pp. 212261, 1994.