Intelligent Systems and Computer Technology
D.J. Hemanth et al. (Eds.)

doi:10.3233/APC200176

A Distributed Item Based Similarity Approach for Collaborative Filtering on Hadoop Framework

S.NGunjal[a,1], Yadav S K[b], Kshirsagar D B[b]
[a,b]*UG Scholar,Dept of CSE, Shri Jagdish Prasad JhabarmalTibrewala University, Chudela, Jhunjhunu (Rajasthan), India*

Abstract. Now a day's multiple website provides millions of product to their on-line users. Due to the interaction of millions customer with the e-commerce websites creates massive volume of data. Recommendation system is a dynamic data capturing system filters massive volume of information generated through the interaction of users to web-portals & generate suggestion that fits the user expectations. Recommendation framework is data filtering tools that make use of algorithms and user rating data to recommend the most relevant items to a particular user. Collaborative filtering is one of the successful techniques in the recommendation system to recommend the top N-item. In the majority of the Recommendation framework information sparsity, high dimensionality, adaptability which are normal issue in the RS domain have adversely influence the exhibition of CF. The proposed system is developed to resolve the mentioned problems in most of the recommendation system using item based similarities approach in CF with the help of user sub space clustering approach on hadoop framework. In the proposed system user subspaces formed by considering the interest of the users in the items like Interested, Neither Interested nor Uninterested (NIU), and Uninterested. After the user subspace clustering the neighbor item tree is constructed. To find out the similarities between the items the similarities measures is developed from the neighbor item tree. It observed that in traditional item-based collaborative filtering method requires more computational cost but the computation of item similarities is performed off line & computational cost required for on line prediction is less. In proposed work to improve the computation speed of off line item-item similarity the computation is performed on various nodes in cluster in the hadoop distributed system. The similarity between the two items is used to predict the rating provided by user on the target items. The proposed method tested on the Movie lens 100K, Movie lens 1M in order to make comparisons with the existing techniques. The proposed method improves the performance of the recommendation systems by resolving the issues like scalability, high dimensionality, data sparsity etc.

Keywords.item-based, collaborative filtering, hadoop framework, recommendation system.

1. Introduction

Now a day the e-commerce websites provides the huge number of items to their millions of customer. The customers of e-commerce websites use the on-line shopping

[1] Gunjal S N, UG Scholar, deptof CSE, Shri Jagdish Prasad JhabarmalTibrewala University, India; E-mail: gunjalsanjay1982@gmail.com

facility to purchase the various products. Due to the availability of huge number of product & adding the new product on e-commerce websites it make difficult to the customer so select the product in which customer is interested [4].

So that the recommendation system emerged as information filtering system to understand the interest of the customer through the information collected from the customer like previous purchased product, feedback regarding the product interms of rating, user-user relation, item-item relation to generate recommendations that fits the expectations of the users [3]. Recommendation system is a data filtering tool to predict the future preference of the active user for the Top N-product. Recommendation system provides assistant to the active user on e-commerce website to select the product from the huge number of available product. Role of the Recommendation system is excellent in the online services provided the online users. Depending on the recommendation mechanism RS is classified into to two important classes 1) Content-based filtering (CBF) and 2) Collaborative filtering (CF) model. In Content based filtering mechanism use the features extracted from user profile & characteristics of the item that gives the description of item. In CBF define the model for user -item interaction where user and item representation given in terms of Explicit features. Collaborative filtering is one of the powerful & popular recommender frameworks utilized by a few on-line business organizations like Netflix, eBay and Amazon. Collaborative filtering based on principle that customer like items similar to other items they like, and items that are liked by other people with similar taste. In Collaborative filtering approach predict the interest of the user based on similarity of user or the items. Collaborative filtering based RSs can be classified into memory based and model-based techniques. Memory based collaborative technique is based on the user-item matrix to find out the nearest neighbor user or item for the target users or items. In model based collaborative filtering technique define the model for user-item interaction where the user & item representations have to be learned from interaction matrix [08]. The Collaborative filtering having some challenges as follows:

Data Sparsity: The user can visit only the limited number of items on the websites so the millions of items available on the websites .It also very rare that different user prefers the same item. Due to this high dimensionality of data and sparse database where the most of rating values in user-item matrix are unknown affect the performance of recommendation system [6].Computation time: As the number of users & items increases time required for the computation is steeply rises.

Recommendation accuracy: Accuracy of Recommendation engine depends upon the accuracy of prediction of user preference & rating of user to the items. If predicted preference & rating is too much different than actual then system is accuracy is very less.Scalability: In E-Commerce website the new product continuously get added & number of users are continuously increases. Such websites requires the scalable Recommendation system support for increasing & decreasing number of users and product on websites.

Data volume: Now a days the data generation over the on line e-commerce website is tremendously increases. So that to handle the large volume of data need to develop the parallel and distributed algorithms in the recommendation system. So that the system is become scalable to handle the massive amount of data and tremendously increasing the number of users [5].

A new approach to CF filtering is proposed in this paper use the user subspace clustering technique to cluster user in different clusters to address the sparsity, high dimensionality, scalability issues on hadoop platform.

With the help of subspace clustering of users it is possible to find out the users group which are interested, neither interested nor uninterested, uninterested in the certain items. When users are in the same group means they interested in certain items.

These user subspaces are used to construct the neighbor item tree to find out the similarities between the items using proposed item similarities measure. The amount of similarity of every item with target item is determined on the basis of the position of item in the tree. When the target item is directly matched with the first level item in the tree then the Pearson correlation coefficient (PCC) similarity is used. To measure the indirect similarities between the items the new similarity method is developed which uses the tree structure of the item to measures similarities between the items. In proposed work the clustering process and subspace construction is done in off-line & prediction of item to the user is done in online only. So that in proposed wok most of the computation is done off line. Only the prediction of item to the user is done on-line. The proposed work handles the scalability, sparsity& high-dimensionality problem of the recommendation system. The performance of recommendation system is measure through the precision, recall, accuracy parameter. The result show that proposed system gives the excellent performance in recommendation process than the existing recommendation process.

2. Related Work

Hybrid user-item based CF method provides the more sophisticated personalized recommendation for user to address the data sparsity and scalability issue in the collaborative filtering. In the Case Based Reasoning (CBR) combined with average filling to handle the information sparsity problem. Self Organizing map (SOM) optimized with Genetic Algorithm (GA) to perform clustering on large dataset to improve the performance of the CF[9].

The new user similarity model is proposed that uses the local context and global context of the user. In local context of the user the rating provided by user is considered where as in global context the user behavior is considered. In similarity measure the common rating between the two users is considered. In this paper mean and variance of the rating is considered [3].Thing based calculations keep away from this bottleneck by investigating the connections between things first, instead of the connections between clients. Suggestions for clients are figured by discovering things that are like different things the client has loved. Since the connections between things are moderately static, thing based calculations might have the option to give a similar quality as the client based calculations with less online calculation [10] .

A client is spoken to by a client commonality vector that can demonstrate the client's inclination on every sort of things. A particular component of TyCo is that it chooses "neighbors" of clients by estimating clients' closeness dependent on their commonality degrees rather than corated things by users[11,16].The Movie focal point dataset is recorded by perusing the document and dataset is separated into groups utilizing k-implies bunching into k bunches with the goal that each group has a centroid. The separation between the client and the centroid is determined, and the client is set in the bunch whose centroid is the least good ways from him. At the point when every single such client have been migrated, the centroids are moved and the new positions determined. Subsequently, the evaluated rating that the client will give is determined, and structure is advanced utilizing cuckoo search algorithm[7].

Clustering approach dependent on thing metadata data's focusing on an outcome improvement. Things are bunched by the genre(s) they have a place with. As a thing can have a place with a few types, it tends to be put in a few groups. Furthermore, data between these two segments are traded through two relapse regularization things, with the goal that the area data controls the investigation of the inert space. This technique is completely founded on the client thing rating matrix [2]. Grouping based recommender frameworks progressively broad and commonsense. The proposed a system dependent on network factorization which considered multitype clustering, i.e. trust-based client bunching, comparability based client grouping and similitude based thing bunching. The strategy tends to the information sparsity and cold beginning issue. One potential confinement of the proposed structure is that just consider the circumstances including evaluations and trust, in spite of the fact that it might be direct to consolidate other data sources, for example, time-arrangement data, dynamic varieties data and the spot data, to additionally improve the nature of recommender frameworks. Another impediment is made a suspicion that clients are more comparable in a similar trust bunches than clients in various trust clusters[4].

3. A Distributed Item Based Similarity ApproachFor Collaborative Filtering On Hadoop Framework

In Recommendation system due to high dimensionality of data and sparse database where the most of rating values in user-item matrices are unknown that affect the performance of system. Now a day the amount to data generated is huge in size due to increase in number of users and items over the e-commerce website. So that the system should be scalable in order to support for the tremendously increase in users and item over the e-commerce websites. To solve the scalability problem of recommendation system in proposed system the Hadoop Distributed File System (HDFS) is used. HDFS provides high throughput access to application data and is suitable for applications that have large data sets. In Proposed Item Based similarity model for Collaborative filtering the user-rating matrices are used to compute the similarities among the items as per the user perspectives. In collaborative filtering it is assume that the customer gives the same rating to item also they give the same rating for other items also. In the proposed method we are creating the three subclutsers one is users interested in item, second is users neither interested nor uninterested, third is uninterested user in the items. In proposed work we process all interested, neither uninterested nor interested, uninterested items for users to find out the similarities between the items.

3.1. Construction of Binary Matrices

In proposed method the transpose of user-item matrices is taken to get the item-user matrices. The Item -User matrices is represented in three binary matrices form like user interested item-user matrices having item rating 4 & 5 ,user neither interested nor uninterested item -user matrices having item rating 3, user uninterested item -user matrices having item rating 1&2, In all three matrices rating values represented as 1 and unknown values are represented as 0. All these three matrices are used to construct the user list interested in the item.

3.2. Constructing the Interested User List in Items

In this step from above matrices the list of users interested in items, neither interested nor uninterested (NIU) & uninterested is prepared. This list shows the groups of users are interested in the specific items.

3.3. Constructing user subspace clusters

In this step the subspace clustering is used to find the some user subspace which can be used to find a group of similar items. By this step of proposed method, we will have three items subspaces which consist of interesting, NIU and Uninteresting items subspaces. Possible interesting user subspaces for each item will be made by comparing that items interested user with other items interested users.

3.4. Removing redundancy

The aim of this step is to reduce the redundancy occurred in the user subspace. In this step the user subspace which are the subset of other user subspace that are neglected. Subspace in global table arranged in the descending order based on the number of elements in subspace.

3.5. Neighbor items finding by using neighbor trees

After removing redundancy, upon entrance of the target item, a set of item must be defined as neighbor item. To do this, items who had rated by users on each subspace similarly, will be defined as related items to that subspace. Similarity between the targets items determine by considering the neighbor tree. For example for the target item rating given by user (u 2,u3 ,u4) interestingly, the user subspace (u2 , u 4) is the most similar, so according to the subspace users list, items (i1 , i5) are the most similar items and are located on the second level of the neighbor items tree.

3.6. Estimating rating value

In proposed system to calculate the similarity values between the target items and the items in the first level of the tree Pearson Correlation Coefficient (PCC) which is the most suitable measure to measure the similiarity. The PCC between item i and j can be calculated as follows.

$$Sim(i,j) = \frac{\sum_{u\in U}(R_{u,i}-R_i)(R_{u,j}-R_j)}{\sqrt{\sum_{u\in U}(R_{u,i}-R_i)^2}\sqrt{\sum_{u\in U}(R_{u,j}-R_j)^2}} \qquad (1)$$

In which U = { u1, u2,. . . ,un } represents the set of users .Ru,i is the rating provided by user u for the item i. Ri , Rj is the average rating value for the item i , j. The range value [−1, 1] of PCC, −1 means complete dissimilarity and value 1 complete similarity between item i & j. Negative values may decrease the recommendation accuracy.

In the proposed system, the PCC is used as a traditional similarity measure to find neighbor items. But classical method is not applicable to find out the indirect similarity between the items that were located at third level of the item neighbor tree. We therefore use a new similarity measure to calculate the similarities value of target items and indirectly similar items. The proposed similarity measure can be seen in equation.

sim(Ia,Ib)=(∑u⟦W*L⟧)/(∑u/w) (2)

In which Ia a indicates the target item and Ib indicate the indirect neighbor item. W and L will be computed by Eqs.(3) and (4).Since items Ia and Ib will be compared through a joint direct neighbor like - Ix -, p represents a combination of w and L between direct neighbors Ia, Ix and also Ix and Ib

$$L_{ij} = \frac{\propto * |U_i \cap U_j|}{|U_i \cup U_j|} \qquad (3)$$

In which | Ui∪Uj | indicate the sum of sets of visited users visited by items i and j which are direct neighbor items and |Ui ∩ Uj | represents common visited users of items i and j .

$$Wij = 1 - \frac{\sum |U_i \cap U_j| (R_i - R_j)^2}{\beta * |U_i \cap U_j|} \qquad (4)$$

Two popular prediction methods like the Sim_pred and Avg_pred are used to predict the ratings provided by user to the target items on the basis of the neighbor item rated by the users.

Sim_Pred(p,j)=R(i+∑(J∈N)⟦Sim(i,j)*(R(p,j) - RJ ⟧)/ (∑(J∈N)/ sim(i,j))⟧) (5)

Avg.Pre(P,i)=1/n*∑(i=1)^nR_ip (6)

3.7 Recommendation

Traditional method of recommendation makes use of the rating value between 4 to 5 recommend the item to the customer. The rating values are in the range of 1 to 5. Rating value 1 means the user is not interested in the item and 5 means most interested in the item.

4. Result

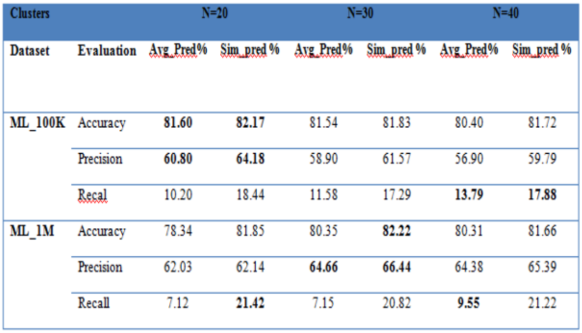
In the proposed system we use the K-means clustering with different number of clusters, Pearson Correlation Coefficient (PCC) with different numbers of neighbors. The Result of the proposed neighbor item similarity subspace clustering CF (NISSC) is compared with the existing method like K-means Clustering & PCC on ML_100k, ML_1M datasets. The performance of the system is measured in terms of precision, recall, accuracy.

$$Accuracy = \frac{\sum True\ Positive + \sum True\ Negative}{\sum Total\ Population} \quad (7)$$

$$Recall = \frac{\sum True\ Positive}{\sum Really\ Positive} \quad (8)$$

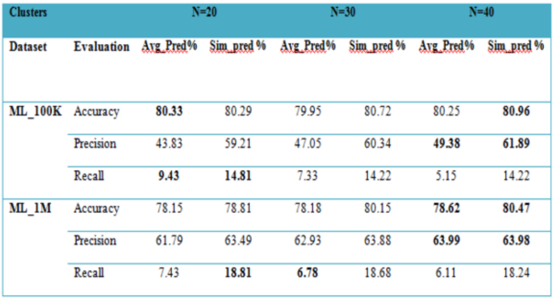
$$Precision = \frac{\sum True\ Positive}{\sum Test\ Outcome\ Positive} \quad (9)$$

Table 1. Recommendation by K-means Clustering Method



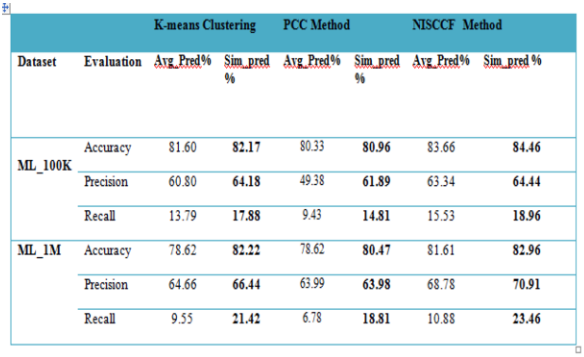
Clusters		N=20		N=30		N=40	
Dataset	Evaluation	Avg_Pred%	Sim_pred %	Avg_Pred%	Sim_pred %	Avg_Pred%	Sim_pred %
ML_100K	Accuracy	**81.60**	**82.17**	81.54	81.83	80.40	81.72
	Precision	**60.80**	**64.18**	58.90	61.57	56.90	59.79
	Recal	10.20	18.44	11.58	17.29	**13.79**	**17.88**
ML_1M	Accuracy	78.34	81.85	80.35	**82.22**	80.31	81.66
	Precision	62.03	62.14	**64.66**	**66.44**	64.38	65.39
	Recall	7.12	**21.42**	7.15	20.82	**9.55**	21.22

Table 2. Recommendation by PCC Method



Clusters		N=20		N=30		N=40	
Dataset	Evaluation	Avg_Pred%	Sim_pred %	Avg_Pred%	Sim_pred %	Avg_Pred%	Sim_pred %
ML_100K	Accuracy	**80.33**	80.29	79.95	80.72	80.25	**80.96**
	Precision	43.83	59.21	47.05	60.34	**49.38**	**61.89**
	Recall	**9.43**	**14.81**	7.33	14.22	5.15	14.22
ML_1M	Accuracy	78.15	78.81	78.18	80.15	**78.62**	**80.47**
	Precision	61.79	63.49	62.93	63.88	**63.99**	**63.98**
	Recall	7.43	**18.81**	**6.78**	18.68	6.11	18.24

Table 3. Comparisons of Existing Methods with Proposed NISCCF (Neighbor Item Similarity Subspace Clustering CF) Method



		K-means Clustering		PCC Method		NISCCF Method	
Dataset	Evaluation	Avg_Pred%	Sim_pred %	Avg_Pred%	Sim_pred %	Avg_Pred%	Sim_pred %
ML_100K	Accuracy	81.60	**82.17**	80.33	**80.96**	83.66	**84.46**
	Precision	60.80	**64.18**	49.38	**61.89**	63.34	**64.44**
	Recall	13.79	**17.88**	9.43	**14.81**	15.53	**18.96**
ML_1M	Accuracy	78.62	**82.22**	78.62	**80.47**	81.61	**82.96**
	Precision	64.66	**66.44**	63.99	**63.98**	68.78	**70.91**
	Recall	9.55	**21.42**	6.78	**18.81**	10.88	**23.46**

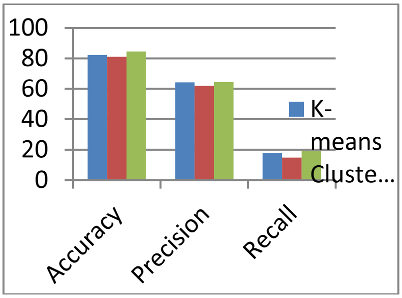


Figure 1. Similarity Prediction for Dataset ML_100K

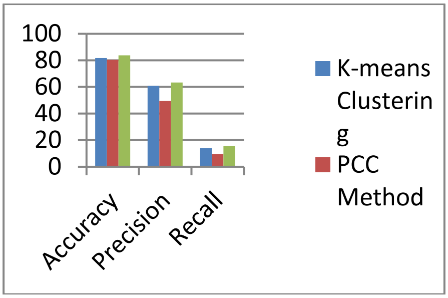


Figure 2. Average Prediction for Dataset ML_100K

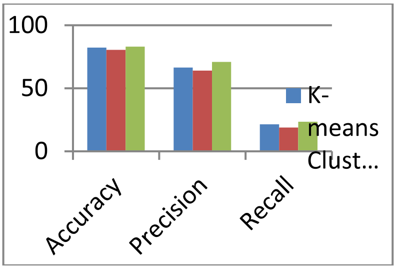


Figure 3. Similarity Prediction for Dataset ML_1M

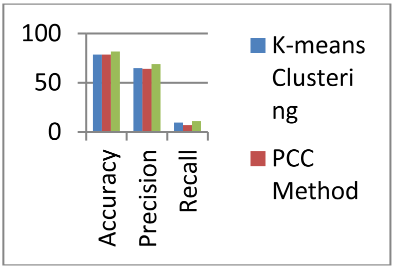


Figure 4. Average Prediction for Dataset ML_1M

5. Conclusion

Item similarity based collaborative filtering is one of the successful recommendation system than user based collaborative filtering because the number of users always more than the number of available item on e-commerce websites. The users are more dynamic over the e-commerce websites than the items. Due to this computational cost for finding the similarities between the users is more as compared to item-item to similarity. The contribution of proposed system is to solve the problem like sparsity, high-dimensionality & scalability that negatively affect the performance of the recommendation system. In this system the item set present at the first level is directly similar to the target item & target item is indirectly similar to the third level item. In proposed work the Pearson Correlation Coefficient is used to find out the direct similarities between the two neighbors & for indirect similarity calculation the new similarity method is proposed. The performance of the item based collaborative filtering is measures by using the performance metrics like Precision, Recall, and Accuracy. To evaluate performance of proposed system is tested on Movielens 100K, Movielens 1M dataset & compute performance metrics like Precision, Recall, and Accuracy in order to make comparison with the traditional system. The proposed method shows the classical result than the traditional collaborative filtering system.

References

[1] H. Liu, Z. Hu, A. Mian, H. Tian, X. Zhu (2014), A new user similarity model to improve the accuracy of collaborative filtering .Knowledge-Based Systems, Vol. 56, 156-166.

[2] J. Liu,Y. Jiang, Z. Li,X. Zhang, H. Lu (2016), Domain-Sensitive Recommendation with User-Item Subgroup Analysis.IEEE Transactions On Knowledge And Data Engineering, Vol. 28, No. 4, 939–950.

[3] B. Hammou, A. Lahcen, S. Moulinea, B. Hammou, A. Lahcena, S. Moulinea(2018), APRA: An approximate parallel recommendation algorithm for Big Data. Knowledge-Based Systems,Vol-157,10-19.

[4] X. Ma, H. Lu, Z. Gan , Q. Zhao (2016), An exploration of improving prediction accuracy by constructing a multi-type clustering based recommendation framework.Neurocomputing,Vol-191,388–397.

[5] M. Narayanan, A. Cherukuri(2016), A study and analysis of recommendation systems for location-based social network (LBSN) with big data .IIMB Management Review,Vol- 28, 25–30.

[6] H. Koohi, K. Kiani(2017),A new method to find neighbor users that improves the performance of Collaborative Filtering .Expert Systems With Applications ,Vol-83, 30–39.

[7] R. Katarya, O. Verma(2017),An effective collaborative movie recommender system with cuckoo search.Egyptian Informatics Journal, Vol.18, 105–112.

[8] Yi Cai, H. Leung, Q. Li, H. Min, J. Tang, and J. Li (2014),Typicality-Based Collaborative Filtering Recommendation. IEEE Transactions On Knowledge And Data Engineering, Vol. 26, No. 3,766-779.

[9] Nitin Kumar, Zhenzhen Fan* (2015), Hybrid User-Item Based Collaborative Filtering.Procedia Computer Science 60 1453 – 61

[10] T. wata, K. Saito, T. Yamada (2008), Recommendation Method for Improving Customer Lifetime Value.IEEE Transactions On Knowledge And Data Engineering, Vol. 20, No. 9,1254-1263

[11] Yi Cai, Ho-fung Leung, Qing Li, Senior Member, IEEE, Huaqing Min, Jie Tang, and Juanzi Li(2014), Typicality-Based Collaborative Filtering Recommendation.Ieee Transactions On Knowledge And Data Engineering,Vol. 26

[12] X. He, M. Gao, Min-Yen Kan, D. Wang (2017),BiRank: Towards Ranking on Bipartite Graphs,"IEEE Transactions On Knowledge And Data Engineering, Vol. 29, No. 1, 57-91

[13] H. Mashayekhi, J. Habibi, T. Khalafbeigi, S. Voulgaris, M. Steen (2015),GDCluster: A General Decentralized Clustering Algorithm.IEEE Transactions On Knowledge And Data Engineering, Vol. 27, No. 7, 1892-1905.

[14] S. Frémal, F. Lecron(2017), " Weighting strategies for a recommender system using item clustering based on genres,"Expert Systems With Applications Vol.77 ,105–113.

[15] M. Jiang, P. Cui, X. Chen, F. Wang, W. Zhu (2015), Social Recommendation with Cross-Domain Transferable Knowledge .IEEE Transactions on Knowledge And Data Engineering, Vol. 27, No.11, 3084-3087.

[16] V.D.Ambeth Kumar (2017), Efficient Routing for Low Rate Wireless Network a Novel Approach. International Journal of Image Mining, Vol. 2, Nos. 3/4, 2017.