An Algorithm to Improve Accuracy of Recommendation System

Riddhi Patel[1] Radha Pathak[2] Prof. Hitul Patel[3]
[1,2,3]Department of Computer Engineering
[1,2,3]Swaminarayan College of Engineering and Technology, Kalol, GTU, India

Abstract— Recommender system is a new technology which are the most recently used over the internet. It apply KDD process to analyse recommendation for information ,items or services during a active interaction. There are various techniques and algorithms used for recommender systems. In this paper we discuss collaborative filtering technique Tyco using the relations among users groups and item groups,Cluster based data mining k-means algorithm, Map reduce for speed up computing and aprior algorithm. Item-based techniques first recognized the user matrix to identify relationships between items and then use those relationships indirectly to evaluate recommendation for use. Item based algorithm has lower time cost than other CF methods. Further, it can obtain more accurate predictions with less number of big data error predictions.

Key words: Recommendation system, collaborative filtering technique Ty Co, User-based algorithm, Item-based algorithm implementation

I. INTRODUCTION

A recommendation system help users that have no enough capability to check the, potentially huge, number of options[1] Recommender systems provide a proper and ranked lists of items by predicting what the most authentic items are based on the user's preferences and constraints.[1]

Recommendation system like the algorithm k-nearest neighbour based on the success are spreading on the web . The increasing growth in the amount of available information and the number of users has increased many key challenges to the recommendation system Recommender system provides high quality recommendations for millions of users and achieve high coverage in the face of data inadequate.

In collaborative filtering systems the huge of participants access large amount of work in the system. New technologies recommender system for lage scale problems Item-based algorithm the user-item matrix to identify relationships between different items, and then use these relationships to indirectly evaluate recommendations for use.[1]

E-commerce a new way to the recommender system for various online websites. Users can recommend their purchase items and add to the shopping cart list and provide their reviews after buying it.. Recommendation Engine will analyze large Datasets of User purchase pattern and recommends items to online users based on Machine Algorithms. There are three product recommendation techniques are:

- User Behaviour
- User Similarity
- Item Similarity

1) Any user select an item any related item or most recommend item given to user, which are selected by most of the users
2) Items can be recommended by the similar users and it is difficult to scale because of the dynamic nature of the different users.
3) A very fast and easy item-based recommendation approach recently available when users have provide ratings.[1]

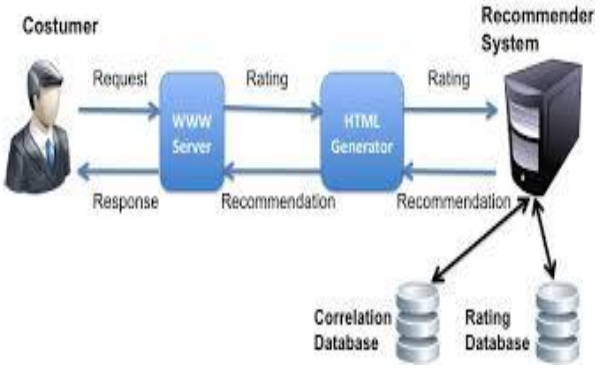


Fig. 1: Recommendation system

II. BACKGROUND AND RELATED WORK

A CF technique *TyCo*, users of neighbour are selected on the bases of user groups rather than co-rated items.

In a collaborative filtering technique, Assume there are U sets of users and O sets of items. Various items can be clustered into one item group is set of same items. For e.g, large movie sets can be clustered into groups like war movies, action and comedy movies etc. Particular one movie belongs to groups of different degrees.

The definition of an item group is given in the following:

Definition An item group denoted by *ki* is a fuzzy set of objects, as following:

$ki = \{Owi,11\ ,Owi,22\ ,\ \cdot\ \ \cdot\ \ \cdot\ ,Owi,h\}$

where *h* is the number of items in *ki*, *Ox* is an item and *wi,x* is the grade of link of item *Ox* in *ki*.

Users who divides same interests on an item group could form a community and we name such a group as a user group.Typicality degree have been changed to different user groups. Otherwise, for each item group *ki*, we present a related user group to some degrees. Suppose Bob and Amen are interested in action movies not in comedy while Amy and Alice are interested in romantic movies very much but not like war movies.Bob and Amen are typical users in the user group of users who enjoy war movies, but not typical users in the user group belonging to romance movies.

All rights reserved by www.ijsrd.com

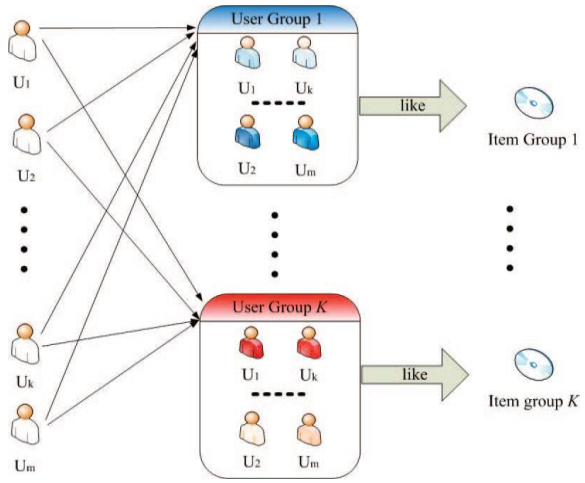


Fig. 2: The Relations among Users, User Groups And Item Groups

– An example of TyCo technique is shown in figure below:

	g_1	g_2	g_3	g_4	g_5	g_6
U_1	0. 87	0. 75	0. 92	0. 12	0. 32	0. 28
U_2	0. 34	0. 21	0. 38	0. 89	0. 85	0. 94
...	...	...	...	...	...	...
U_k	0. 81	0. 79	0. 89	0. 15	0. 29	0. 31
...	...	...	...	...	...	...
U_m	0. 41	0. 22	0. 35	0. 90	0. 88	0. 92

Fig. 3: An example of user typicality matrix based on TyCo

	i_1	i_2	...	i_k	...	i_n
U_1	5	?	...	3	...	4
U_2	?	?	...	4	...	5
...	...	...	...	...	...	...
U_k	2	5	...	?	...	3
...	...	...	...	...	...	...
U_m	5	4	...	2	...	?

Fig. 4: An example user rating matrix in traditional CF

III. APRIOR ALGORITHM

Aprior algorithm is a abstraction that is used for mining the frequent items sets for various transaction of databases by association rule.It allows the sets which are the part of transaction databases.For e.g customers choose group of items like books,various websites used recently etc.[2]

The steps of aprior algorithm are as follows:

1) uk: ua itemset of size k
2) Lk : frequent itemset of size k
3) L1 = {frequent items};
4) for (k = 1; Lk !=_ ; k++) do begin
5) uk+1 = users generated from Lk;
6) for each transaction t in database do
a. increment the count of all users in uk+1 that are
7) contained in t
8) Lk+1 = candidates in uk+1 with min_support
9) end
10) return _ k Lk;

A. *Advantages*

- It is uncomplicated.
- Its performance is simple.

B. *Disadvantages*

- Users set can be accomplished by many scan over the database.
- Its accomplishment takes more memory, space and time.
- Frequent item sets are similar as the number of database passes. [2]

IV. PROPOSED WORK

A. *Proposed Algorithm*

Item-based collaborative filtering is a model-based algorithm for production of recommendations. In the algorithm, the community of similar items and various different items in the dataset are calculated by using one of a number of similarity measures, and then these similarity values are used to predict ratings for user-item pairs not available in the dataset.

B. *Input*

Datasets with user rating and item rating are considered. Item-based collaborative filtering evaluate matrix user vs item and find out its similarity using distance formula and originate each user's preference value.

C. *Output*

User with predicted item and its preference value. Similar values between items are evaluated by watching all the users who have rated both the items. As shown in the diagram the similarity between two items depend upon the ratings provided by the items for users who have rated both of them.

D. *Similarity measures*

There are a number of various mathematical aspects that can be used to calculate on the basis of similarity of two items:

1) *Cosine-Based Similarity*

This similarity is also known as vector-based similarity, this formulation views two items and their ratings as vectors, and identify the similarity between them as the angle between these vectors.Items are vectors in the m dimensional user space .

$$sim(i,j) = \cos(\vec{i},\vec{j}) = \frac{\vec{i}\cdot\vec{j}}{||\vec{i}||_2 * ||\vec{j}||_2}$$

2) *Pearson (Correlation)-Based Similarity*

This similarity is evaluated based on how many the ratings by common users for a pair of items change from total ratings for those items.

All rights reserved by www.ijsrd.com

– Using the Pearson-r correlation,

$$sim(i,j) = corr_{i,j} = \frac{\sum_{u \in U}(R_{u,i} - \bar{R}_i)(R_{u,j} - \bar{R}_j)}{\sqrt{\sum_{u \in U}(R_{u,i} - \bar{R}_i)^2}\sqrt{\sum_{u \in U}(R_{u,j} - \bar{R}_j)^2}}.$$

– R(u,i) = rating of user u on item i.
– R(i) = average rating of the i-th item.

3) Adjusted Cosine Similarity

This similarity measurement is a varied form of vector-based similarity where we take into the fact that different users have different ratings schemes; some users might rate items high rate and others might give items less ratings as a preference. To remove this drawback from vector-based similarity, we extract average ratings for each user from each user's rating for the pair of items in question.

– Each pair in the co-rated set corresponds to a different user.

$$sim(i,j) = \frac{\sum_{u \in U}(R_{u,i} - \bar{R}_u)(R_{u,j} - \bar{R}_u)}{\sqrt{\sum_{u \in U}(R_{u,i} - \bar{R}_u)^2}\sqrt{\sum_{u \in U}(R_{u,j} - \bar{R}_u)^2}}$$

– R(u,i) = rating of user u on item i.
– R(u) = average of the u-th user.

4) From Model To Predictions

we can predict the rating for large user-item pair by using the concept of weighted sum. Take all the items similar to our target item, and from those similar items and then select items which the present user has rated. We weight the user's rating for each of these items by the similarity between that and the target item. At last calculate the prediction by the sum of similarities to get a logical value for the predicted rating.

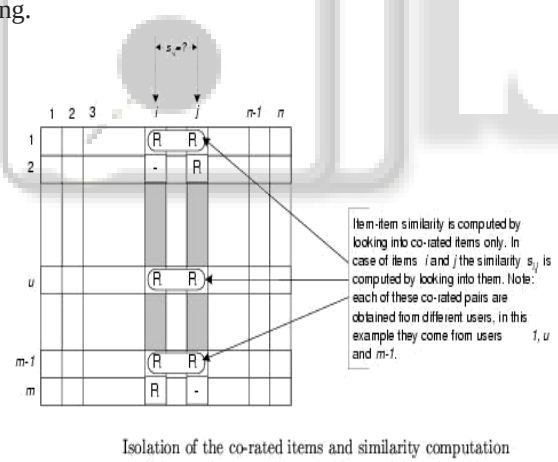
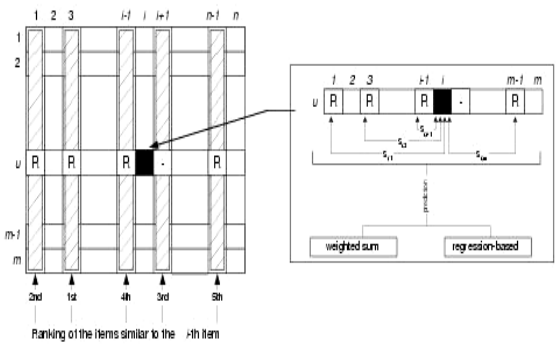


Fig. 5: Isolation of the Co-rated Items and Similarirty Computation

– Generating the prediction – the target users ratings and use techniques to obtain predictions.
– Weighted Sum – how the present user rates the same items.

$$P_{u,i} = \frac{\sum_{\text{all similar items, N}}(s_{i,N} * R_{u,N})}{\sum_{\text{all similar items, N}}(|s_{i,N}|)}$$

– The prediction process is as follows:



Based on accuracy we will decide similarity distance formula.

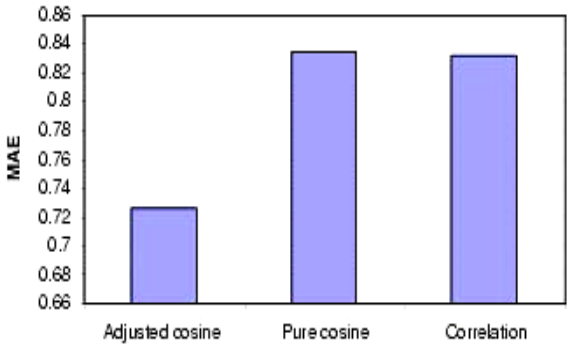


Fig. 5: Effect of similarity computation

The steps of item-based algorithm are as follows:

1) For every item i that u has no preference for yet
2) For every item j that u has a preference for
3) compute a similarity s between i and j
4) Add u's preference for j, weighted by s, to running average
5) Return the top items, ranked by weighted average.
6) Return the top items, ranked by weighted average.

The similarity functions calculated on the basis of Euclidean distance, Pearson coefficient, Log likely hood ratio and tanimo to coefficient shown in graph.

1) Euclidean distance: It calculates the similarity and distance between the two items. On the basis of mean absolute error(MAE) and RMSE values means prediction accuracy can be calculated.
2) Log likely hood ratio: LLR can assess the similarity objects moving around between users based or items. It focuses on the events that occur between the user or items.
3) Tanimoto coefficient: It measures the tendency of two user's preference values to move together to be relatively high or low relatively low on the same items.
4) Pearson coefficient similarity: A number in one series is to be measured relatively large when the corresponding number in the other series is high.

Implementation technical configuration and analysis based on the various existing algorithms as correation coefficient method based group Pearson (SCBPCC), the approximation low weighted rate (WLR), on the basis of collaboration (TCC), SVD ++ and table comparison of different state of the art - methods MAE (mean absolute error) is shown below.

Then I compare the existing algorithm Tyco (collaborative filtering based on Typicality) with the proposed algorithm with different parameters.According to

All rights reserved by *www.ijsrd.com*

the results of existing and proposed algorithm table 6.2.4 mention the value and displays the results to extract subsets of 500 users from data sets and select the first three users 100,200 and 300 to form joint training as ML100, ML200 and ML300 respectively, other users 200 are used as test users.

– Trial users vary the number of classified items provided by test users in training from 5.10 to 20 named Dada 5, given 10 and 20.

Training Set	Methods	Given 5	Given 10	Given 15
ML100	SCBPCC	0.874	0.845	0.839
	WLR	0.915	0.875	0.890
	CBT	0.840	0.802	0.786
	SVD++	0.925	0.911	0.916
	TyCo	**0.830**	**0.799**	**0.777**
	nCo	0.802		
ML200	SCBPCC	0.871	0.833	0.828
	WLR	0.941	0.903	0.833
	CBT	0.839	0.800	0.784
	SVD++	0.881	0.815	0.812
	TyCo	**0.830**	**0.775**	**0.775**
	nCo	0.831		
ML300	SCBPCC	0.870	0.834	0.819
	WLR	1.018	0.962	0.938
	CBT	0.840	0.801	0.785
	SVD++	0.885	0.815	0.802
	TyCo	**0.814**	**0.762**	**0.760**
	nCo	0.960		

Table 1: Improvement in between Ty Co and proposed algorithm IBCF

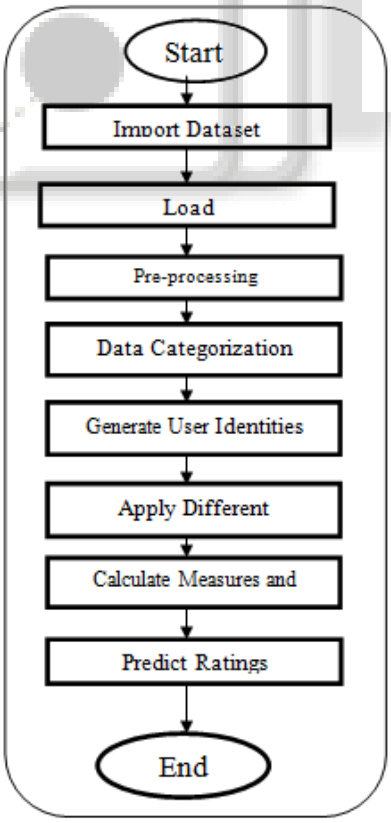


Fig. 6:

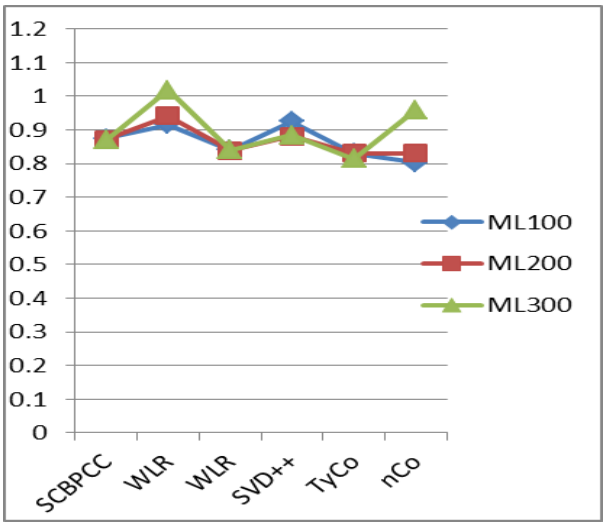


Fig. 7:

– Item cluster vs Error probability for various similarity shown in graph

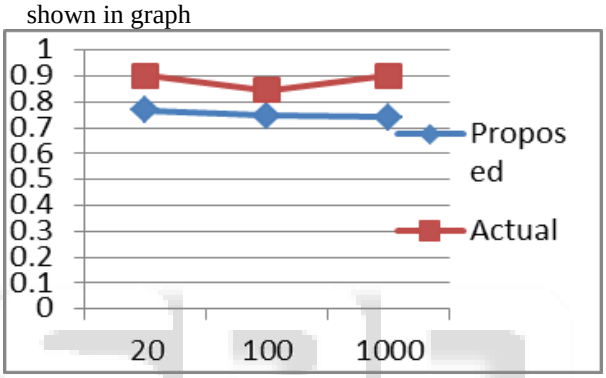


Fig. 8: Euclidean Distance

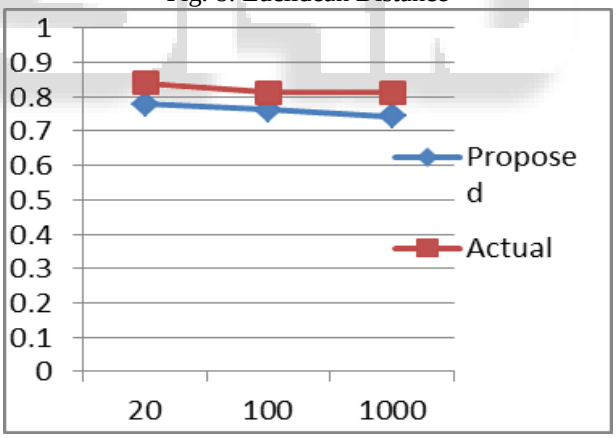


Fig. 9: Tanimoto Coefficient

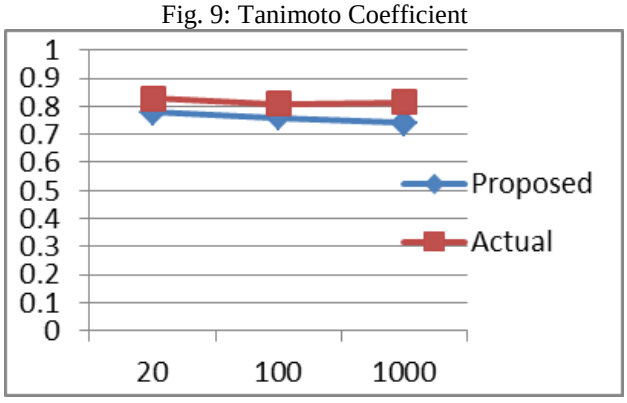


Fig. 10: Pearson Coefficient

All rights reserved by www.ijsrd.com

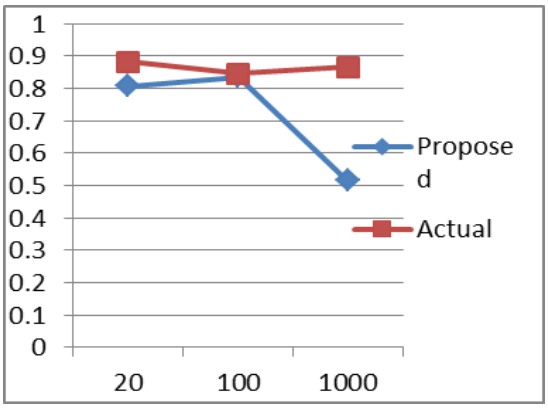


Fig. 11: Log Likely-hood

V. FUTURE WORK AND CONCLUSION

There are various possible future addition to our work.Item-based algorithm for various items chosen by users has been projected.According to the background and related work represents that different algorithms and techniques based on TyCo,Aprior,user and item based permitted different results.Item-based algorithm evaluated similarity between items and make recommendations.Items selected by the users changes rarely,so this often can be computed offline.User-based algorithm recommends items by finding similar users.It is difficult to scale because of the dynamic nature of users.So, This algorithm has better scope for recommendation system used over the internet.Performance of the system can be increased using UBCF and IBCF that generate the set of nearest neighbour of items or users based on user-item rating matrix.The proposed algorithm IBCF provide better efficiency for recommendation system used for different similarity measures for various web services.In future we can create web services based on saas solutions and map reduce technique using very large datasets.

REFERENCES

[1] Yi Cai_, Ho-Fung Leung†, Qing Li_, Huaqing Min_, Jie Tang_ And Juanzi Li,An Algorit', Department Of Computer Science And Technology, Tsinghua University, China, Ieee Transactions On Knowledge And Data Engineering.
[2] J.Refonaa, Dr. M. Lakshmi, V.Vivek , Analysis And Prediction Of Natural Disaster Using Spatial Data Mining Technique,'' 2015 International Conference On Circuit, Power And Computing Technologies.
[3] Shrikanth Narayanan, Ayush Jaiswal, Yao-Yi Chiang,' Yanhui Geng,' Craig A. Knoblock,' Pedro Szekely,' Integration And Automation Of Data Preparation And Data Mining, 2014 Ieee International Conference On Data Mining Workshop.
[4] Leila Ismail, Mohammad M. Masud,' Latifur Khan,' College Of Information Technology Uaeu,' Computer Science Department University Of Texas Dallas, Usa, Fsbd: A Framework For Scheduling Of Big Data Mining In Cloud Computing,' 2014 Ieee International Congress On Big Data.
[5] Carson Kai-Sang Leung*,Richard Kyle Mackinnon,' Fan Jiang,' Department Of Computer Science, University Of Manitoba, Winnipeg, Mb, Canada,' Reducing The Search Space For Big Data Mining For Interesting Patterns From Uncertain Data,' 2014 Ieee International Congress On Big Data.

All rights reserved by www.ijsrd.com