A LOCATION BASED MOVIE RECOMMENDATION SYSTEM

Pradnya Thakur[1], Prof. N.V.Karekar[2]
[1,2]Department of CSE, Gogte Institute of Technology, Belagavi

ABSTRACT
Recommendation systems now a days are personalized that help the user's in suggesting the list of items that are of user's interest. Collaborative Filtering technology is being used by many of the recommendation system. This method connects both user and item clusters that solve the problem sparsity and scalability of existing collaborative filtering technology. The proposed system finds the user who has rated for at least n movies, Least age user along with the Zip code is generated for the users who have rated for at least n movies, which is used to find the top 10 zip code location of the least age users through which the producer or theater owners can find out which age people are watching movies in which location through zip code and use this information to release the movies in the respective areas.
General Terms: Recommender system, Collaborative filtering, User and Item Clustering.

1. INTRODUCTION

Recommender systems are widely used these days to help a user find his desired products from a huge set of options available. They play a very important role in suggesting a relevant information to the users, that are used to give recommendation to most of the items. movielens.com is one of the famous movie recommender website. The information that is existing in internet is vast that make complicated for clients visiting different websites that are in search of objects or things that are interesting for clients. So, the system should learn to understand what the user like and produce reference that are of customers interest.

Recommender systems have emerged out of the need of information retrieval and persistent usage, which is also called as information filtering. This is used to implement core function that compares the item and are used to predict what items can be rated by the specific users. The most popular approach in recommender systems is Collaborative filtering[1]. That uses user item rating matrix and then recommends the items based on preference and the tastes of other users[1]. Recommendations in content based filtering are done on basis of items that are previously liked by the user.

Collaborative filtering has number of issues[2]. Sparsity is one of the major issues. As the items are massive whereas the users who rated the items is small. To get relieved from sparse problem, Factorization of Matrix is resourceful technique which is consistent and vigorous that helps to guess ratings of the user for a particular item for which the user has not rated.

2. PROBLEM DEFINITION

Collaborative filtering recommendation approach creates clustering technique for user and item and join the user and item clustering technology that finds similarities between end user and cluster center. The adjacent neighbors of the end user are established using K nearest neighbor algorithm that are rated for at least n movies and then finds the least age users with top10 zip code for which recommendation can be used for producer or theater owners to find in which location what age people are watching movies and perform the predictions to produce the recommendation that helps to resolve the problem of cold start.

Summary of Collaborative Filtering Process

The objective of collaborative filtering rule is to recommend new things or to guess the value of bound item for a selected user that supported for previous user ratings and therefore the opinions of different alike mind users.

In a collaborative filtering state, there's a listing of p users U and listing of q items I. Every user

U incorporates list of Items I_u that the user has expressed opinions regarding the items.

The user can clearly give the opinions based on rating score, among a numerical scale, or also can be derived from purchase record. It is probable for I_u to be null set. A user U_a € U are known as active user for which the task of collaborative filtering is to seek out the items which will be of 2 type.

- Predictions could be a mathematical value, P_{ai} that express the associated likelihood of an item I_j¢I_{ua} for an user that is dynamic U_a. The expected value is inside an equivalent range(Eg, from one to five) because of the estimated value that are provided by user U_a.
- Recommendation may be a list of N item , I_r € I, that can be liked by most of the dynamic user . The counsel list should get on things or item which is not obtained by the dynamic users. This boundary of collaborative filtering is understood as Top N Recommendation.

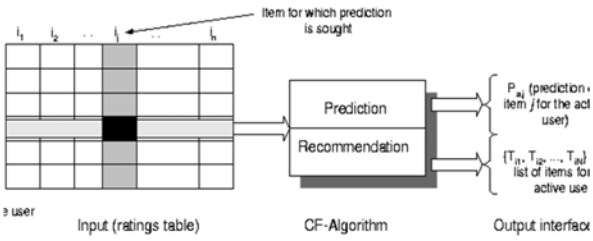


Fig 1.1 Collaborative Filtering Method

Figure1.1 describes the diagram of collaborative filtering method. Collaborative filtering algorithm shows the complete P*Q user item as a rating matrix, A. Each entry a_{ij} during A represent the preferred rating of last user on last item. Every single rating is at the intervals that has some numerical values and it can be 0 that indicates for which the user did not give any rating for item. This algorithm is separated into User based and item based algorithm.

3.CHALLENGES OF THE EXISTING SYSTEM[10][11]:

Sparsity : Even when the users are active, few ratings of the items are available in user item rating database. In collaborative filtering algorithm similar events are calculated on the rated set of objects, hence there is less precision observed in sparsity.

Scalability : The complexity in algorithm depends upon the number of Products and the consumers in the system. Therefore growth in number of consumers and products can result either slow down or can require more number of resources to control the processor speed that require computation that are costly and can grow in huge as the users and items in information set.
Cold Start : The items cannot be suggested until the particular item is been rated by more number of users. When the system is not able to produce recommendations for the new user this can lead to a cold-start problem. The new user has to rate more number of items then after that collaborative filtering algorithm can produce accurate recommendations.

4. RELATED WORK

In mid 1990's the concept of Recommender systems were emerged. In Different recommender sites there was enormous increase in enlargement. It was considerable for different system to produce recommendation for users which are closer to the product of users concern. Recommender systems are widely used these days to help a user find his desired products from a huge set of options available. Mainly accepted method in recommender system is Collaborative Filtering. The system uses user item matrix and then suggests information based on preference and experience of different user[9]. Content based filtering provides recommendations on the basis of items user previously liked.
George, T.[5] designed an incremental, equivalent description of the co-clustering algorithm that make an resourceful instantaneous collaborative filtering structure. The analysis of planned method on huge movies or ratings on different data set show that its probable to get accurateness which can be compared to correlation and matrix factoring at a lot of lower process value.
They are obtain by describing products and consumers. For instance features in case of movie can be genre or that describe type of movie for example, thriller/drama/action etc. Characteristic of this item or product can depict an level to the item that has the particular character. when the user is considered, the above characters show the amount of curiosity a user has in the specific item. Singular value decomposition is a enhanced method comparable to the established collaborating model, as it splits up the rating matrix based on features. Existing

information partition and cluster algorithm are use to partition set of objects that is on ratings of user data. Prediction can be calculated separately with every group. After partitions are created which will get better quality and enhance the collaborative filtering system, but they have varied result on improving accurateness.

In collaborative Filtering, Clustering method has been studied by some of the researches. The cluster analysis gather some user information with the similar characteristics. This information use user rating information to calculate the comparison involving consumers or products. The information can be utilized for building service recommendation to the user. Hence, this mechanism was previously employed in many of the marketable system. There are two methods of collaborative filtering : User Collaborative filtering is a straight forward algorithmic interpretation which discovers other user whose previous ranking performance is same as that of present user and use that user ranking on different items to expect what the present user can like. To guess A users liking of an item for which that user did not rate, users collaborative filtering look for other user who has elevated accord with A user on item together they rated. As the user dataset increases collaborative filtering suffers from scalability problem. It is very much necessary to build the algorithms that are more efficient and scalable in ecommerce websites when there is increase in the user dataset. Item based collaborative filtering, take a main step in the way, which is extensively organized filtering method existing today. Item based filtering method finds similarity among the ranking pattern of the item. If two item belongs to similar user, they are similar users may have similar preferences for similar items.

Xue, G et al.[7] provide the approach, clusters that are produced from training information that offer the foundation for knowledge smoothing and neighbors collection. Which supply advanced accurateness and also improved effectiveness in recommender system. The experimental study on each of the datasets of every pic and movie lens show new projected method that it systematically perform alternatively user primarily base filtering algorithm.

In the first clustering methodology Lee, WS et al.[8] assume that every user is probably going to fit in one of p cluster of user and user's rating for every item is produced in every step the distribution that depends on the item and cluster that user belong.

In secondary methodology, hypothesis is created for every user is probably going to belong to at least one of the p clusters of users whereas every item is expected to belong to at least q clusters of items. Result's projected that projected strategies area are sensible in some way.

The numerical model of filtering algorithm and different algorithm are used to evaluate and approximate the model restriction that embody distinction of Gibbs sampling and K means cluster algorithm. L.H. Ungar[12,13] bestowed that this model may be simply comprehensive to handle cluster of object with multiple attributes that is higher than normal one.

5. IMPLEMENTATION

The process in collaborative filtering follows with the ratings of user given for a particular item, which means that user is interested in the item, system then tries to search for the alternative user who gave the same rating for the specific item in the system. Using this information, system tries to predict for item which the user did not give any rating.

Users are clustered depending upon the ratings given by user for each item and there is a center of cluster for each user groups. The nearest neighbors of the target user are found, Which is dependent on relationship between end user and center, smooth the prediction wherever necessary according to the algorithm used. Collaborative filtering uses Item cluster filtering technique to produce recommendation for the planned system which follows the method of connection both user and item clustering in recommender system which is found to be more scalable and accurate.

Set of User's with same ratings are recognized for user clustering technique by finding the average mean of new user in the specific cluster, the clusters are created and prediction are completed for end user in the system .

Subclass of collaborative filtering are recommender system, these system are used to guess the rating or specify the choice that user would provide for an item. These system became quite common in current years and are used in many of the applications. Objective of the system is to find out the rating for different movies given by user, so that recommendation of movies can be done for the least age users based on zip code location area.

- Notice all the users that gave rating for at least two movies, Movie x and Movie y.
- The above ratings got from user are used to form vector for Movie P and Movie Q.
- Between the vector compute the co-relation using formula.
- The movie that is corelated more may be recommended to other user who is interested to watch the movie.

To compute the similarity between the couple of movies we can assume if one user watch movie matrix, the movie similar to matrix movie can be recommended to same user, for example Blade runner movie is similar to matrix movie. To outline the similarity of movie correlation is used. The main idea behind this is to find all users who rated for both the movie. ratings of these movies is used to form vector, calculation of correlation is performed between the two vectors. Now the most correlated movie can be recommended to the user who is willing to watch the movie.

To Solve this problem chaining of two MapReduce job is done. Output of the first job is given as input to second. To chain multiple jobs multiple driven method is used.

1. In first step data set file is given as input to the first MapReduce job. After completion it will generate output1 file which will work as input to MapReduce job 2.
2. MapReduce job will wait for the completion of MapReduce job1, its output1 will work as input to MapReduce job 2. MapReduce job will generate final output.

MapReduce Job-1 : The first MapReduce job is to collect all the user rated both the items.

MapReduce Job-2 : Is used to find the similarity between item using correlation formula.

K means algorithm and K nearest neighbor is used as the essential algorithm. K is the numeral value given as key in to the rule which is used to specify the preferred quantity of cluster. K item from the k distinct cluster are taken from algorithm. Every residual item in cluster is compared with nearby cluster center. Clusters are evaluated again and again from the cluster center that are created in preceding pass and then clusters are re-evaluated.

To compute the linear co-relation among two different vector Pearson co-relation method is used to find rating between target item and items that are left behind.

$$sim(t,r) = \frac{\sum_{i=1}^{m} (R_{it} - A_t)(R_{ir} - A_r)}{\sqrt{\sum_{i=1}^{m} (R_{it} - A_t)^2 \sum_{i=1}^{m} (R_{ir} - A_r)^2}}$$

Rating R_{it} is for target item T given by user u, Rating for remaining item R_{ir} by user u. Average A_t is rating of target item for all user that are co-rated. A_r is average of left out item for user that is co-rated, and m is for all rating user for item t and r.

Choosing Cluster Center

In Item based collaborative filtering main vital phase is to look for the neighbor of target item. whenever additional new members and products are brought in the record In ancient memory base filtering it was required to look for the entire records to find the neighbor of target items then it experiences very poor scalable problem. When the groups of items are created it forms cluster center, In this center mean ratings of whole items are present in that cluster through which the neighbor of target items can be selected in that item center. To seek out the resemblance among the target items and center Pearson cor-relation method is used, the item that is more same within the center is taken as further applicant.

Selecting Neighbor's

After choosing the target item from the closest cluster center, it is necessary to calculate the resemblance among the item in cluster center and target item. similar items can be chosen using on cosine measure method, that come across angle among the vector of rating as target item and remaining item.

$$sim\ (t, r) = \frac{\sum_{i-1}^{m} R_{it} R_{ir}}{\sqrt{\sum_{i-1}^{m} R_{it}^{\ 2} \sum_{i-1}^{m} R_{ir}^{\ 2}}}$$

Generating Recommendation

After getting the ratings of all the items , the calculation should be done on the weighted average of neighbor's ratings, weighted by their similarity to the target items. The ratings of the end user and end item can be calculated follows:

$$P_{ut} = \frac{\sum_{i=1}^{c} R_{ui} \times sim(t,i)}{\sum_{i=1}^{c} sim(t,i)}$$

R_u is the ratings of the end user to neighbor items, sim is the similarity of the end item and neighbor user, for entire co rated item, there is number for all ratings of user to the different items.

6. SYSTEM DESIGN

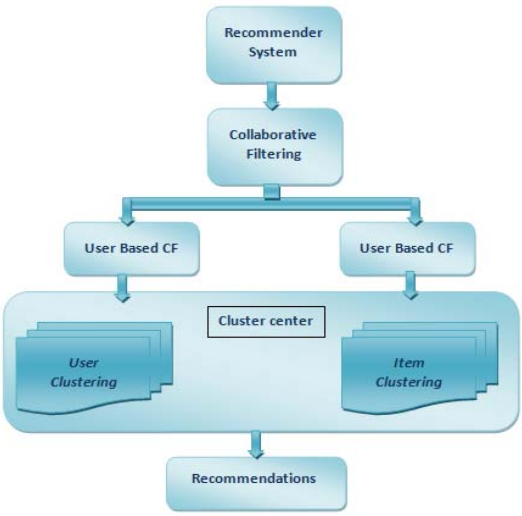


Fig 6.1 System design Architecture.

The System Design Architecture consists of the following Modules :

1. Recommender System

Recommender system plays an important role in suggesting relevant information to the users that can help people find interesting things that are widely used in electronic commerce systems.

2. Collaborative Filtering

Collaborative filtering is an effective method of recommender system. Many websites use filtering technique in their recommender system to illustrate the browsing expertise for every individual user.

3. User base Collaborative Filtering

User base collaborative filtering proposes the better method to get the convinced user's fascinating items that is to get other user having same concern.

4. Item Base Collaborative Filtering

This method is used to predict the item after doing enquiry between the similar items and other items that are related to the other users.

5. User Clustering

Clusters are created for user supported on user's rating on each item and every user cluster contains a cluster center. Similarity among the end user group center and the closest neighbor of end user may be found and sleek the prediction wherever required.

6. Item Cluster

This technique works by recognizing collection of item who emerge to have same rating.

7. Cluster Centers

In Cluster centers, when the items are grouped item center is created. Here we represent an average rating for all the item in that cluster center. From the item center we can prefer target item neighbor. To find relationship among the target item and item center Pearson co-relation is used. The recommender system which combines user and item clusters in filtering method is more precise than the existing system.

7. EXPERIMENTAL RESULT

Data set

The experiment is conducted on MovieLens dataset, the dataset is downloaded from the site (http://grouplens.org/datasets/Movilens/) to estimate the presentation.

These datasets were composed from GroupLens investigation venture at institution of higher education of Minnesota. Movie Lens is internet based recommendation system. Weekly more than hundred of customers visit the website[16] to rate and obtain recommendation for movie.

This website has more than 50000 consumer who have convey their suggestions on more than 7000 different movie. This dataset consists of sufficient users to gain 1,00,000 rating from 6000 consumer on 4000 movie ie considered as items with every user having at least rated more than 30 film. Ratings are presented in numerical values ranging from 1 to 5 in which 4 and 5 represent optimistic rating and number 3 indicate average between 2 and 3 and 1 represent negative. Users are represented by their user id and movies ie items are also represented by item id. File format is as follows: u.item contains information about the items(movies). u.user file contains demographic data about the user with the list of (User identity | Gender | Age | profession id | Zipcode). All ratings be into other file rating file that contains (User id | Movie id | Rating | Mobile number). Each line represents one rating for one movie by one particular user. Movie information file contains details about the movie which includes (Movie id | Movie name & released year | type of movie) type of movie determines whether the movie is animated, action or comedy movie.

The Explanation of input data

Rating.dat - Which comprises of user Id, rating, Movie Id and timestamp.

Example : 1::1193::5::978300760

: 409::2406::3::976288876

Movies.dat - Which consists of movie Id, Title and genres

Example:1::Toystory(1995)::Animation|childre ns|comedy
2::Jumanji (1995)::Adventure|Children's|Fantasy

User.dat : Consists of User Id, Gender, age, occupation and Zip codes

Example: 1::F::1::10::48067
2::M::56::16::70072

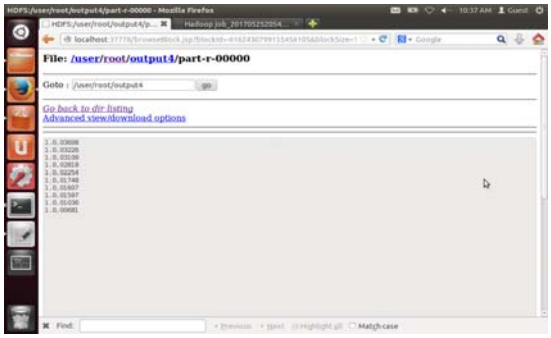
The Recommendation algorithm using Hadoop MapReduce framework is composed of three phases

Phase 1 : First we find a set of all User Ids who has rated for atleast N movies. In our model we have considered N=30. The algorithm takes the input Rating.dat as input file.

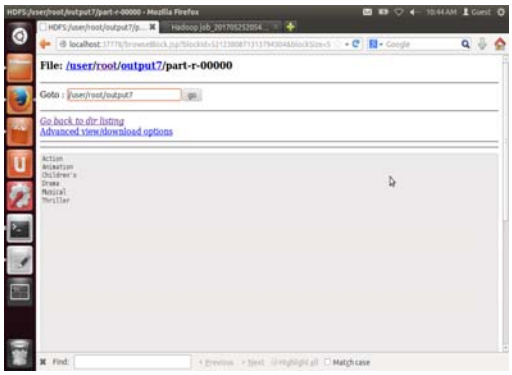
Output : Obtains output of user who has rated at least 30 movies. This condition is defined to show the user rate the movies regularly to find regular user.

Phase 2 : The goal of the work is to recommend movies to the newly joining user (Relevant movies). The existing model uses user-to-user social interaction to build community (Louvain's Community Detection CD)) and based on which recommendation is done. However for practical scenario this may not be efficient. For example consider for two users who live in different countries but they are connected through social network, if movies are releasing in certain countries where user 1 is located. The recommendation model recommends user 2 the same movie. However the movie is not release in user 2 country. As a result, the movie is practically relevant to user 2.

To address this work, Present location and age based recommendation model is used. This model takes the user data as input, the algorithm produces two output. The algorithm first find the least age person in each zip codes location, then find the number of person in that Zip code location, then it finds whether the person has rated the different movies atleast 30 times, then cumulated the overall age and obtain the output (Output3). Now using this information of output 3, We compute the mean age of the user in that zip codes, then find top 10 zip code location with least age user (we can consider the kids). Recommendation of movies based on age and zip code location is done.



Phase 3 : At last the genre of recommended movies is found by giving the input data file as movies.dat. We have used k-nearest neighbour algorithm for classifying or creating group/community.



8. FUTURE WORK

The Future work for recommendations system can be done through the zip code area that can be generated for the aged user or minimal aged user further along with the specific areas so that the predicted movies can be released according location based on the age of the users, using location based movie recommendation technique.

9. CONCLUSION

In this paper the Recommendation method connects both user and item clustering technique which is used to solve the problem of scalability and sparsity. The nearest neighbor of target user can be found using k means nearest algorithm and determine the predictions. Using this information the user that rated at least n movies are found and the least age of the users rating the cartoon movies are generated based on the users location area wise . Output generated is least aged user along with the Top 10 zip code, where these predictions can be given to producers or directors to recommend such types of movies in the mentioned zip code area where there are least aged users available and such users can watch the same predicted movies.

REFERENCES

[1] Breese J, Hecherman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence (UAI'98). 1998. 43~52.

[2] Yu Li, Liu Lu, Li Xuefeng, A hybrid collaborative filtering method for multiple-interests and multiple-content recommendation in E-Commerce, Expert Systems with Applications 28 (2005) 67–77.

[3] George Lekakos, George M. Giaglis, Improving the prediction accuracy of recommendation algorithms: Approaches anchored on human factors, Interacting with Computers 18 (2006) 410431.

[4] Sarwar B, Karypis G, Konstan J, Riedl J. Item-Based collaborative filtering recommendation algorithms. In: Proceedings of the 10th International World Wide Web Conference. 2001. 285295.

[5] George, T., & Merugu, S. A scalable collaborative filtering framework based on co-clustering. In Proceedings of the IEEE ICDM Conference. 2005

[6] M. O. Conner and J. Herlocker. Clustering Items for Collaborative Filtering. In Proceedings of the ACM SIGIR Workshop on Recommender Systems, Berkeley, CA, August 1999.

[7] Xue, G., Lin, C., & Yang, Q., et al. Scalable collaborative filtering using cluster-based smoothing. In Proceedings of the ACM SIGIR Conference 2005 pp.114–121.

[8] Lee, WS. Online clustering for collaborative filtering. School of Computing Technical Report TRA8/00. 2000

[9] Hyung Jun Ahn, A new similarity measure for collaborative filtering to alleviate the new user coldstarting problem, Information Sciences 178 (2008) 37-51.

[10] A. Kohrs and B. Merialdo. Clustering for Collaborative Filtering Applications. In Proceedings of CIMCA'99. IOS Press, 1999.

[11] Lee, WS. Online clustering for collaborative filtering. School of Computing Technical Report TRA8/00. 2000.

[12] L. H. Ungar and D. P. Foster. Clustering Methods for Collaborative Filtering. In Proc. Workshop on Recommendation Systems at the 15th National Conf. on Artificial Intelligence. Menlo Park, CA: AAAI Press.1998

[13] L. H. Ungar and D. P. Foster. A Formal Statistical Approach to Collaborative Filtering. Proceedings of Conference on Automated Leading and Discovery(CONALD), 1998.

[14] M. O. Conner and J. Herlocker. Clustering Items for Collaborative Filtering. In Proceedings of the ACM SIGIR Workshop on Recommender Systems, Berkeley, CA, August 1999.

[15] A. Kohrs and B. Merialdo. Clustering for Collaborative Filtering Applications. In Proceedings of CIMCA'99. IOS
Press, 1999.

[16] http://grouplens.org/datasets/Movilens/