Machine Learning Based Recommendation System

Shivangi Singla[1] , Vinod Maan[2]
[1,2] Mody University Of Science and Technology

Abstract- *For making decisions for their personal interest on which the information for people based in their daily life that is the task of the recommendation system. It plays important role in the internet world. This paper describes various algorithms and techniques of recommendation system. Now-a-days, recommendation systems have been using the field of artificial intelligence named as machine learning. There are number of algorithms used in the machine learning and choosing the best algorithm out of number of algorithm is a very difficult task. Recommendation system have been divided into mainly three categories named as Collaborative filtering, Content based filtering and Hybrid filtering. In this, we will describe the Collaborative filtering with their types. This paper precise these techniques and their limitations.*

Keywords- Recommendation System, Machine learning, Collaborative, Content Based, Hybrid

I. INTRODUCTION

Recommendation Systems(RS) are helpful for the users for searching new services or items which includes , books, transportation, music or even people, depend on information about the recommended item or the user[1]. The main goal of the recommendation system is to make decision with maximum profit and minimum risk. Recommendation system , now-a -days are used in many companies like Google, LinkedIn, Netflix, Twitter.

The beginning of the Tapestry proposed in the mid of 1990's the field of Recommendation system begins. For the recommendation of the documents Tapestry was designed where there was many users and newsgroups[4].

There are three types of recommendation system, named as, Collaborative Filtering, Content Based Filtering and Hybrid Approach. Collaborative filtering systems collects the information from the community of users, while Content based filtering approach depends on what user likes according to the user's field of interest and Hybrid approach defines the combination of both approaches.

Machine learning is the field where there are many methods for computation which uses experience to improve the performance and predictions are being made.

Michalski, R. S., Carbonell describes Machine Learning as follows: “A computer program is said to learn from experience E with respect to some class of tasks T and performance measure P, if its performance at tasks in T, as measured by P, improves with experience E”. The study has been started since 1950s[2].. Today , there is a huge number of machine learning algorithms which are used in different fields now-a- days. Machine learning is classified into three types: Supervised Learning, Unsupervised Learning, reinforcement Learning. The algorithms which are used in recommendation system are included in these categories of machine learning[2].

II. RECOMMENDATION SYSTEM

For past few years, recommendation system is a platform for many online systems. But now-a-days , it is playing a very crucial role in many more systems which includes, portals, blogs, search engines, web pages etc. Recommendation system has mainly two components: Item and User , which can be put onto top of another system[4]. The domain of the recommendation system contains users for them item's preferences are expressed.. Rating is to be defined as the preference for an item given by the user. It is frequently represented by (User, Item, Rating) triple. With the help of all these triples a matrix is made called as rating matrix. Item data of system or both item and user data can be used to build the recommendation system. Items includes book, song, movie, news, procedure etc.

Items

	1	2	...	i	...	m
1	5	3		1	2	
2		2				4
Users :			5			
u	3	4		2	1	
:					4	
n			3	2		
a	3	5		?	1	

Fig.1 Rating matrix, i belongs to the item, each cell is described with ru, rating of user u.

Here a is the active user and to predict the missing rating ra [4]. This rating matrix is disperse, because mostly items are not rated by the user. Our main reason to build the rating matrix for predicting the rating given by the user to the unrated item. Activeuser is described as the current consideration for recommendation by user. The items which are rated highest are presented as recommendation[3].

Types of recommendation system:

1. Collaborative Filtering(CF):

It is also called Social Information Filtering. It maintains a database of a variety of items about users' rating. The similar rating which are rated highly whose items are recommended. It is divided into two main categories named as model based approach and memory based approach which are to be discussed in detail below .

2. Content Based Filtering(CB):

It needs item data individually. Items which are liked by the user in past and the attributes which are matched of the user are recommended item in this approach. Extracting features is a easy task in the case of structured data[3]. It has been categorized into two main algorithms named as TF-IDF and Naive Bayes. It has many advantages: transparency and its main effort is to make the profile of the user by considering the users rating and it has disadvantages too: analysis of the content is limited in nature and specialization is much more needed in this approach.

3. Hybrid Approach: Both collaborative and content based approaches are combined in this method. Advantages of both the approaches are included in this approach[5]. Their are different types of this types of approach which combines both the above mentioned approaches in it specified as Weighted Hybrid Approach, Mixed Hybrid and Cross-Source Hybrid Approach. Their are many issues included in this approach: Reliable Integration and Efficient Calculation[5].

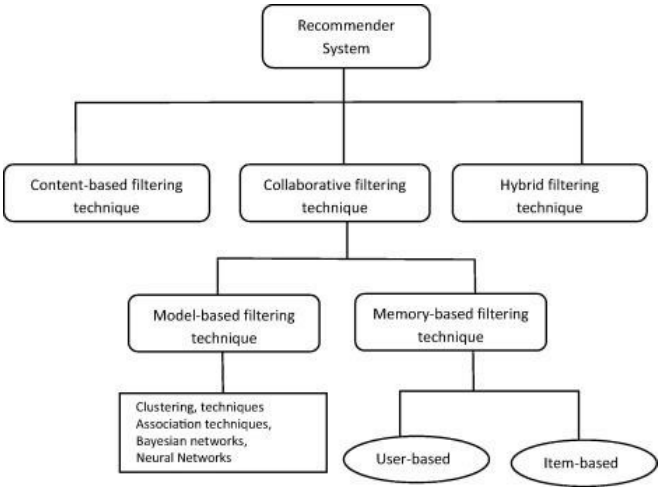

Recommender System
Content-based filtering technique
Collaborative filtering technique
Hybrid filtering technique
Model-based filtering technique
Memory-based filtering technique
Clustering, techniques
Association techniques,
Bayesian networks,
Neural Networks
User-based
Item-based

Fig. 2. Flow chart of Recommendation System[7]

3. Collaborative Filtering

Collaborative Filtering defines the evaluation of the items by the guidance of the other people's opinions. This term collaborative filtering is not been there for not more than a decade but the roots are there from centuries that is sharing the opinions with the other people. When processing information for recommendation the user data has been considered. For example, when user profiles are accessed in an online store, all the user data is accessed by the recommendation system, like age, country, songs purchased, city etc. Identification of the same music preference is shared by the system with the help of this recommendation system and then similar uses buy that recommended songs.

Collaborative models are divided into two types: Neighborhood-based and Model- based Approaches. Neighborhood approach is also called memory-based approaches.

Breese et al. described that memory-based algorithms require all items, users, ratings(the triple factor) are gathered in memory and patterns are rated by created the summary of the ratings periodically and offline in model based approach.

For ongoing algorithms, model based algorithm or hybrid approach is used with storing rating data in memory[3].

3.1 Neighborhood-based Collaborative filtering:

In this technique, placed on their similarity with the active user ,a subset of users are chosen and to produce predictions, a weighted combination of their ratings are being used.

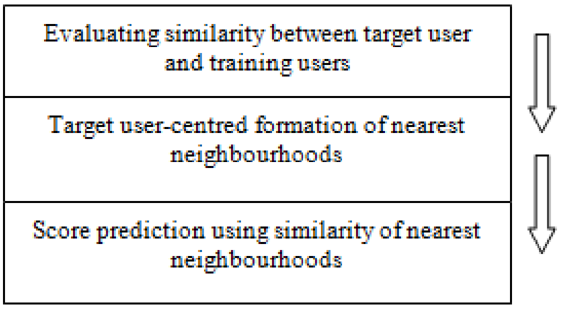

Evaluating similarity between target user and training users
Target user-centred formation of nearest neighbourhoods
Score prediction using similarity of nearest neighbourhoods

Fig. 3 Models of machine learning algorithms. The algorithm is summarized as follows:

1. Evaluating similarity
2. Generation of nearest neighbor
3. Predicting score

1.Existing similarity metrics:

The resemblance with training user and user that is active are evaluated for determining the items commonly rated. A weight for predicting the preference score is used for finding similarity[7]. From previous studies, various similarity metrics have been proposed which are Tanimoto Coefficient, Cosine Similarity, Pearson's Coefficient, Spearman's Rank Correlation described below one by one.

Tanimoto Coefficient: It specify the identity between two sets. The ratio between intersections of these sets is called Tanimoto Coefficient.

$$T(X,Y) = X \cup Y / (X+Y)-(XY) \qquad (1)$$

Here, X,Y are two sets[5].

This coefficient doesn't deal with user rating matrix but this is the case of precise rare data set is powerful.

Cosine Similarity: Also called as Cosine Coefficient or Vector similarity[2]. The cosϴ is calculated in between two points. It considers that in vector space model two points are specified for rating of users having common attributes.

$$COS(U1,U2) = \Sigma rU1*Ii*rU2/||U1||\ ||U2|| \qquad (2)$$

Here, U1,U2 are two users and Ii is item i.

Pearson's Coefficient: Between two variables the linear relationship is to be found and strength of that linear relationship defines the Pearson Correlation[3]. It is represented by r, its value ranges from[-1,0,1] where -1 denotes negative correlation, 0 denotes no correlation and 1 denotes positive correlation.

$$r(U1,U2) = \Sigma(rU1*Ii-U1)*(rU2*Ii-U2)/SU1*SU2 \qquad (3)$$

Here, SU1 is the standard deviation for U1 user and SU2 is for user2.

Spearman's Rank Correlation: It is similar as Pearson's correlation but with one difference, Spearman's Rank Correlation uses the rank of scores[5]. It is better than Pearson's Correlation because preference score for Collaborative Filtering is normalized in range.

$$r(U1,U2) = 6*\Sigma(rank(rUm)-rank(rUm))^2/n*(n^2-1) \qquad (4)$$

2.Formation of Nearest Neighbor:

Different algorithms have been given by different Collaborative Filtering researchers to improve performance. This is the second step of generation of nearest neighborhoods after similarity evaluation. It includes various techniques for choosing the nearest neighborhoods like K-Means Classification, a graph algorithm.

3.Prediction of Preference Score:

Final step of memory-based CF is to predict the preference score for non-rating items of the target user[2]. There are different methods have been proposed and most general algorithm is called Weighted Mean. The performance evaluation has two types: prediction accuracy and recommendation quality.

$$PSU1,Ii = \Sigma\ sim(U1,NNUi)rNNUi,Ii / sim(U1,NNUi) \qquad (5)$$

Here, PSU1,Ii is defined as the score which is predicted of user U1 of item i, NNUi stands for the nearest neighbor and sim stands for the similarity between U1 user and its nearest neighbor[9].

It has two types of approach named as User based approach and Item based approach.

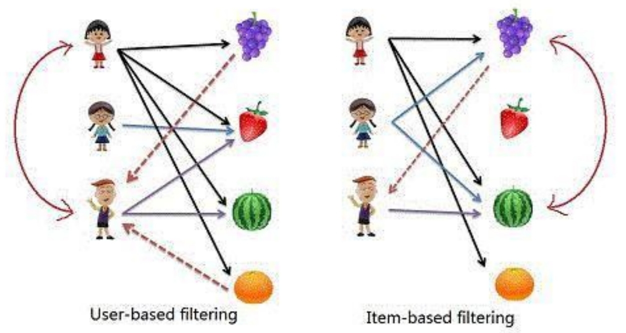


Fig. 5 Difference between User based and item based filtering

Advantages and Disadvantages of Neighborhood-based:

They are domain-independent, increasing quality of recommendation, no content analysis is required. These are some of the advantages. The disadvantages includes "cold start" problem which arises due to the Collaborative filtering methods which depends on acceptable performance of user from the past[8]. If a new user is added then it is necessary that the new user should rate the items so that new user would be able to get the better recommended items[2]. Sparsity is one of the disadvantages which describes due to past user in a network.

3.2 Model Based Approach

In this, a theoretical model has been proposed for rating the user behavior. From the available rating data, parameters of the model are estimated and predictions are being made by the project[4]. Model based algorithms are designed to remove the disadvantages of the memory based or neighborhood based approach. With the estimation of statistical models, the recommendations are being provided for user ratings. There are different algorithms with the help of which we can apply this approach: K-Means CF and Cluster Model[8].

3.2.1. K-Means CF:

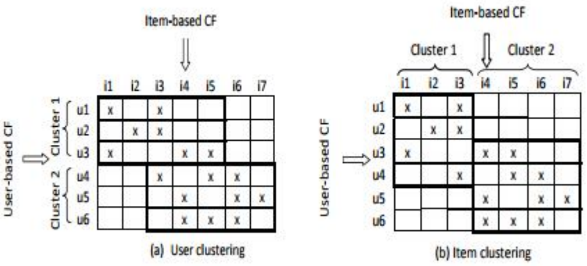
There is a wide application of K-Means method in data mining, machine learning and statistics. K-means is the part of unsupervised learning used to clarify the problem of clustering[5]. The input which is given to the K-means algorithm is determined by the distances enclosed by items which forms the group of similar items named as cluster and here distance stands for difference within items. Input parameter which is used here is k, which is the number of clusters used. The algorithm is iterative, the centroids of the clusters are to be calculated and that item is to be reassigned whose centroid distance is closest[6].

The algorithm is described as:

1. K points are positioned into the space expressed by the objects that are really being grouped. Basic set of centroids are defined by these points.
2. Allow every object to the set which has nearest centroids.
3. Recalculation of the regions of the K centroids are made when each and every object have been authorized.
4. Step 2 and 3 has to be repeated as far as there is no movement in centroids. Segregation of the objects into many sets have been composed and minimized metric can be calculated[6].

3.2.2. Cluster Model:

The customers who are similar are to be found and are done by dividing the customer base into many segments by cluster models. This task is to be treated as a classification problem. The goal is to assign the user to the segment which contains the most similar customers[7]. Then it uses the ratings and purchases of the customers to generate recommendations in the segment. The segments are created using unsupervised learning algorithm and clustering algorithms, whereas, in some applications, segments are determined manually. These models have better performance and online scalability as compared with collaborative filtering[4]. This is because as if they are compared the user to a controlled number of segments rather than the entire customer base.



	i1	i2	i3	i4	i5	i6	i7
u1	x		x				
u2		x	x				
u3	x			x	x		
u4			x		x	x	
u5				x		x	x
u6				x	x	x	

(a) User clustering

	i1	i2	i3	i4	i5	i6	i7
u1	x		x				
u2		x	x				
u3	x			x	x		
u4			x		x	x	
u5				x		x	x
u6				x	x	x	

(b) Item clustering

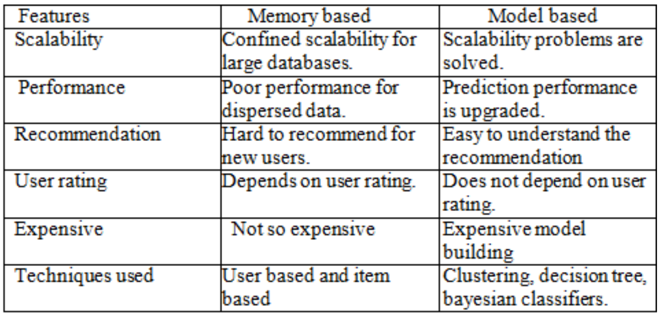
Fig. 6 Clustering Model

Two steps are defined for Traditional Collabortive Filtering methods where first is, closeness between two users can be calculated and neighborhood is analyzed and second is, rating of the neighbors are assembled and discover the recommendation for users which are acive[9]. Butin this method, pimarily users are being clustered then the traditional method is engaged to induce recommendations.

The algorithm suggests that:

1. Codify the social system of users with the description of set of users and their inteconnected relations.
2. Segregation of users into different clusters wiith the help of clustering methods.
3. Anticipation of the neighbors with the guidance of clusters.
4. Comparison between model based technique and memory based technique[8]:

Table no. 1



Features	Memory based	Model based
Scalability	Confined scalability for large databases.	Scalability problems are solved.
Performance	Poor performance for dispersed data.	Prediction performance is upgraded.
Recommendation	Hard to recommend for new users.	Easy to understand the recommendation
User rating	Depends on user rating.	Does not depend on user rating.
Expensive	Not so expensive	Expensive model building
Techniques used	User based and item based	Clustering, decision tree, bayesian classifiers.

IV. CONCLUSION

Recommendation system has three main types: collaborative filtering, content based filtering and hybrid approach. Many recommendation systems have been prepared on the basis of these types. Machine learning concepts have been used in the recommendation system to improve the performance of the recommendation system. Most of these methods have been able to resolve the problems. It is required to work on this research area due to information explosion so that exploring and providing new methods that can provide big range of applications in recommendation system field with the consideration of the privacy aspects and the quality. This field needs enhancement for present and future requirements with better recommendation qualities.

REFERENCES

[1] S. P. Sahu, A. Nautiyal, and M. Prasad, "Machine Learning Algorithms for Recommender System - a comparative analysis," International Journal of Computer Applications Technology and Research, vol. 6, no. 2, pp. 97–100, 2017.

[2] M. Mohri, A. Rostamizadeh, and A. Talwalkar, Foundations of machine learning. Cambridge, MA: MIT Press, 2012.

[3] M. D. Buhmann, P. Melville, V. Sindhwani, N. Quadrianto, W. L. Buntine, L. Torgo, X. Zhang, P. Stone, J. Struyf, H. Blockeel, K. Driessens, R. Miikkulainen, E. Wiewiora, J. Peters, R. Tedrake, N. Roy, J. Morimoto, P. A. Flach, and J. Fürnkranz, "Recommender Systems," Encyclopedia of Machine Learning, pp. 829–838, 2011.

[4] H.-W. Yang, Z.-G. Pan, Bing-Xu, and L.-M. Zhang, "Machine learning-based intelligent recommendation in virtual mall," Proceedings of 2004 International Conference on Machine Learning and Cybernetics.

[5] J. L. Herlocker, J. A. Konstan, A. Borchers, and J. Riedl, "An algorithmic framework for performing collaborative filtering," Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR 99, 1999.

[6] Akhmed Umyarov, Alexander Tuzhilin,"Improving Collaborative Filtering
Recommendations Using External Data", 2008 IEEE.

[7] Greg Linden, Brent Smith, and Jeremy York,"Amazon.com Recommendations
Item-to-Item Collaborative Filtering", IEEE Computer Society 2003.

[8] R. K.sorde and S. N. Deshmukh, "Comparative Study on Approaches of Recommendation System," International Journal of Computer Applications, vol. 118, no. 2, pp. 10–14, 2015.

[9] I. Portugal, P. Alencar, and D. Cowan, "The use of machine learning algorithms in recommender systems: A systematic review," Expert Systems with Applications, vol.
97, pp. 205–227, 2018.