Threshold based KNN for fast and more accurate recommendations

Siddharth J. Mehta
Computer Engineering Dept.
Nirma University
Ahmedabad, Gujarat
sid93jm@gmail.com

Jinkal Javia
Computer Engineering Dept
Nirma University
Ahmedabad, Gujarat
javiajinkal@gmail.com

***Abstract*— Recommender systems attempt to predict the preference/ratings that a user would give to an item. Traditional collaborative filtering give recommendation to a user based on its similarity of ratings with the ratings of other users in the system. But they face issues such as sparsity, cold start problem, first rater problem and scalability. In the proposed framework, a user is being recommended by filtering K random users whose similarity is crossing some threshold and applying collaborative filtering only on those users. For the users/items visiting for the first time, demographic information is used. In it, demographics of users/item visiting for the first time are compared with users/item in system and discarding that user/item if a single mismatch is found. This framework has less MAE as compared to KNN or user based collaborative filtering, takes very less time to recommend as compared to above mentioned algorithms, as only K neighbors need to be considered.**

Keywords*— *Cold Start Problem, Collaborative Filtering, Demographic Information, Recommendation System, and Threshold K nearest neighbor (KNN)

I. INTRODUCTION

Recommender systems support users in personalized way for the identification of product based on the history of the user that can be useful or interesting in the large space of possible product [1]. Recommender system is an information filtering technology, commonly used on e-commerce Web sites to present information on items and products that are likely to be of interest to the reader. In presenting the recommendations, the recommender system uses details of the registered users' profile, opinions and habits of their whole community of users and compares the information to reference characteristics to present the recommendations. As a research discipline, recommender systems has been established in the early 1990's and since then it has shown enormous growth in terms of algorithmic developments as well as in terms of deployed applications. Practical experiences from the successful deployment of recommendation technologies in e-commerce contexts (e.g., amazon.com and netflix.com) contributed to the development of recommenders in new application domains.

Mentioned below are the types of recommender system [9]:

A. Non-personalized recommender
B. Collaborative filtering
C. Demographic recommender
D. Content filtering
E. Knowledge based recommender
F. Hybrid recommender

A. Non-personalized recommender:

This type of recommendation is identical for each customer. They are generally based on popularity measures of product like no. of clicks, no. of purchases, no, of likes, most rated, highest rated etc.

Merits & Demerits: It is the easiest to implement but it lacks personalization and it might not appeal to each and every customer [9].

B. Collaborative filtering:

Collaborative Filtering exploits similar consumption and preferences pattern between users. It maintains the database of many users' ratings for variety of items. For a given user, it finds other similar users whose ratings strongly correlate with the current user [10]. Recommending items rated highly by these similar users, but not rated by the current user. Almost all existing commercial recommenders use this approach [4].

Collaborative filtering follows the following approach:

- It takes the consumption matrix as the input wherein the (i, k) entry indicates the user i has consumed the item k or not.
- Weight all users with respect to similarity with the active user.
- Select a subset of the users (neighbours) to use as predictors.
- Normalize ratings and compute a prediction from a weighted combination of the selected neighbours' ratings.
- Present items with highest predicted ratings as recommendations. It is based on explicit ratings, where the user i rates the item k.

Consider the table I for example:

978-1-4799-8349-0/15/$31.00 ©2015 IEEE

TABLE I: User – Item matrix

User	Item1	Item2	Item3	Item4	Item5	Item6
User1	D	A	B	D	?	?
User2	A	F	D	-	F	-
User3	A	A	A	A	A	A
User4	D	D	-	C	-	-
User5	A	C	A	C		A
User6	F	A	-	-	-	F
User7	D	-	A	-	A	-

The above table consists of the set of users and the ratings given to the items by corresponding users. In the above table '-'indicates that ratings are not yet provided by the user. Consider the case of User1. Based on the data available the system has to predict the ratings of Item5 and Item6 for User1 in order to decide whether to recommend or not the item to User1. It is apparent from the table that User1 and User7 have more or less the same taste and since User7 has rated A for Item5 there are high chances of User1 rating it A or B and hence the item can be recommended to User1. The other way to infer is User1 and User2 have opposite taste and hence the item liked by User1 will probably be hated by User2. Hence, there are chances that User1 likes Item5 as it is rated F by User2.

Thus, collaborative filtering performs the filtering based on correlation with other users having similar tastes.

Collaborative filtering techniques include:

- *User-User Collaborative Filtering:*
  In this the neighbour having the similar taste to that of the user is chosen.
- *Item-Item Collaborative Filtering:*
  In this the similarity amongst the item is pre-computed via ratings.

Merits & Demerits: It does not require any knowledge about the product or the user; hence it is independent of the domain. They can recommend any item to the user because they do not look at the preferences of the user. But, the dataset on which collaborative filtering is to be performed is very large and sparse [9]. It also suffers from cold start problem and first rater problem.

C. Demographic recommender:

The demographic information like age, gender, geography, economic status, education, etc are made to determine the group of people that would like a particular item. They predict the group of users based on personal attributes. This demographic information is provided by the user while registering/enrolling for the service, hence no effort is needed for collecting it.

Merits & Demerits: The advantage is that as the ratings of user are not used and hence new user would also be recommended items [9]. Knowledge about items is not required, hence it is domain independent. But, the disadvantage is that customer with unusual taste and different opinions are not catered well. Recommending them is difficult in such cases. Another disadvantage is that, using demographic data leads to privacy issues.

D. Content based filtering:

Content based Filtering characterizes the affinity of users to certain features (content, metadata) of their preferred items. The filtering is done on the basis of the attributes of the item. For example, in the case of movies, who are the actor, director, genre and so forth [11].Machine Learning Algorithm is used to induce a profile of the user's preferences from examples based on a feature description of content. There are various other classification technologies under this area to decide whether the item is applicable to the user. In content based filtering the opinions of other users are not taken into account. This obviates the need for the maintaining the data relating to the preferences of other users. Such filtering is able to cater well to the users with unique taste. Also, the chief advantage is that when the new item arrives in the pool of items it does not face the cold start problem as in the case of collaborative filtering and hence can recommend the new and unpopular item.

Merits & Demerits: It works well under the condition that feature vector is represented properly [9]. But for each and every product, each product will have to be assigned vector manually if none are available. This could be a tiresome work. As new user would not have any profile of preferences, this filtering does not work on them and hence they suffer from cold start problem as well.

E. Knowledge based recommender:

These recommenders use the knowledge of user and items to recommend items that meet user requirements. It usually involves a series of questions whose answers depict their preferences, and by matching these preferences with those of items, recommendations are made.

Merits & Demerits: Biggest advantage is that, it does not have to store any information about user [9]. No ratings or demographic information is needed to be stored for recommending. If the preferences of customer changes, it is easy to adjust. But, every time you need to enter your preferences as the system does not learn them which can be dull or monotonous task. High domain knowledge is required about items in the system, which as times might not be possible

F. Hybrid Recommendations:

Hybrid recommender is combination of more than one technique such as Content based Filtering, Collaborative Filtering and Demographic information [3]. Therefore Hybrid

Filtering enjoys the advantages of all those techniques incorporated by getting rid of the limitations imposed by single approach.

II. MAJOR ISSUES

The major issues encountered in traditional Collaborative Filtering are listed below:

1) *Cold Start:* There needs to be enough other users already in the system to find a match. The matrix is always very large as well as very sparse as all the users do not consume most of the items. The challenge is how can we complete the missing entries in the matrix or how can we predict the ratings for the items that were not yet consumed by the user when the new user visits for the first time.

2) *Sparsity:* If there are many items to be recommended, even if there are many users, the user/ratings matrix is sparse, and it is hard to find users that have rated the same items [1].

3) *First Rater:* The system cannot recommend an item that has not been previously rated. Therefore problem arises when the new items are added, as no corresponding ratings are available [1].

4) *Popularity Bias:* Recommendations cannot be provided to someone with unique tastes as it tends to recommend popular items.

III. PROPOSED FRAMEWORK

Recommender system suffers from issues such as cold start problem, first rater problem and scalability problem [2].Generally in e-commerce website there are extremely large number of users and items and on the contrary numbers of ratings available are considerably low since users' consume only limited number of items out of total items available. This sparsity problem has negative impact on the efficiency of collaborative filtering algorithm. Also, since we have to deal with huge amount of data, the computational speed is low.

In the proposed framework shown in Figure 3.1, collaborative filtering is combined with the demographic information of users and the item profile in order to generate a hybrid system. The proposed system is capable of addressing cold start, first rater and also to improve computations.

The Book-Crossing dataset available on Group lens is used. The system database consists of user, item and rating information. The user demographic information comprises of age and location. The item profile (in our case books) comprises of title, author, publisher and year of publication.

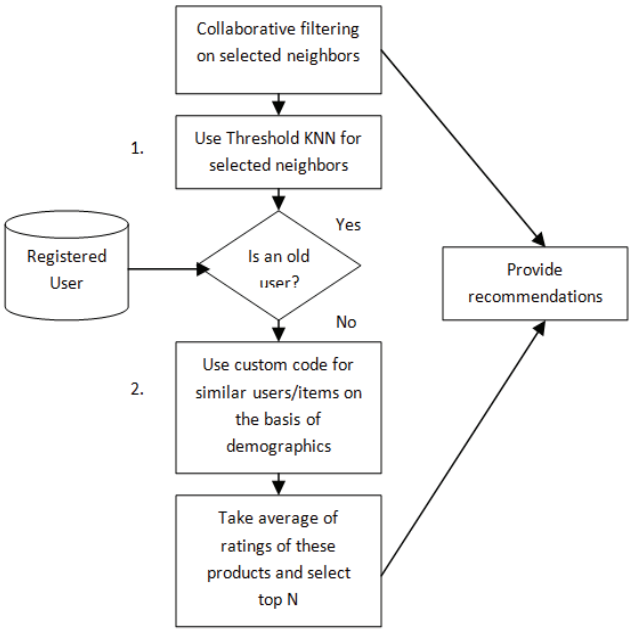

Fig 3.1

If the user is old user (rated a minimum of 5 items), a threshold based KNN algorithm (labeled as 1 in figure 3.1) is used to find the K neighbors for collaborative filtering. The algorithm works as shown in Figure 3.2:

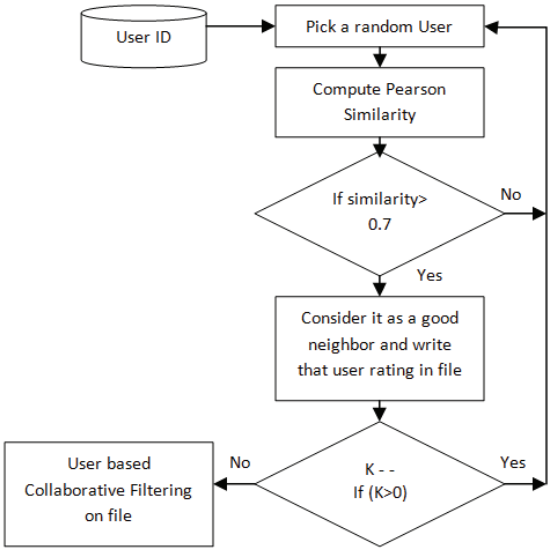

Fig 3.2

When a user wants recommendations, we compute Pearson correlation similarity between the active user (user that wants recommendation) and a randomly selected user. If the similarity is greater than 0.7, that randomly selected user is considered our neighbor for collaborative filtering. Then again a random user is taken and same process is repeated until we

get a total of K such neighbors. Formulae of Pearson Correlation similarity is given by equation (1):

$$s(u, v) = \frac{\sum_{i \in I_u \cap I_v} (r_{u,i} - \overline{r_u})(r_{v,i} - \overline{r_v})}{\sqrt{\sum_{i \in I_u \cap I_v} (r_{u,i} - \overline{r_u})^2} \sqrt{\sum_{i \in I_u \cap I_v} (r_{v,i} - \overline{r_v})^2}} \quad (1)$$

As compared to K nearest neighbor where similarity between all the user were computed and the best K users whose similarity was maximum were chosen for collaborative filtering or in threshold based neighborhood where similarity for all users were computed and those users that cross the threshold similarity were taken for collaborative filtering, in this case first K neighbors whose similarity is greater than 0.7 will be chosen, thus reducing computations and providing faster recommendations as compared to above traditional methods. However by experimenting it in Net Beans using Apache Mahout, we found better quality recommendations.

Whenever a new user enters the system, cold start problem is addressed using the demographic information of the user. The system will search for other users already existing in the database with similar demographics as the user under consideration. The K nearest neighbor (KNN) algorithm with threshold value using Euclidian distance metric is used for this purpose. Any tuple whose value is not matching will be discarded and is not taken into consideration. The predicted ratings for the item will now be the average rating of the top N neighbors for that item. The same procedure is also used when new item is added to the existing system solving the first rater problem. The flowchart for finding nearest user/item based on demographics (labelled as 2 in figure 3.2) is shown in Figure 3.3.

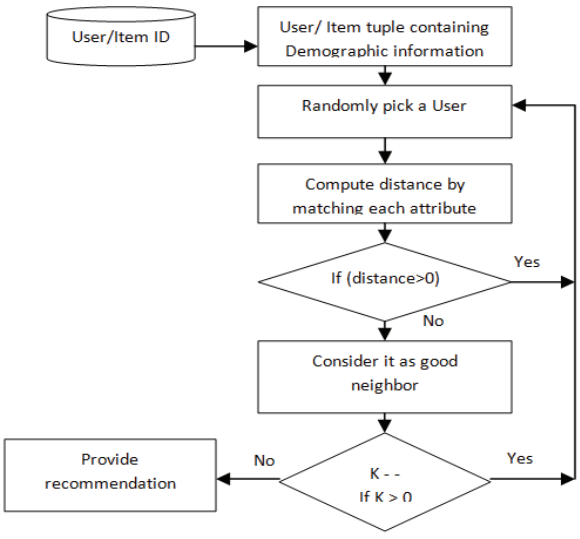

Fig 3.3

IV. OBSERVATIONS

The system can be evaluated using mean absolute error (MAE). MAE takes the mean of the absolute difference between the predicted rating pu, and actual rating ru, for all the held out ratings as described by equation (2):

$$MAE = \frac{1}{n} \sum_{u,i} |p_{u,i} - r_{u,i}| \quad (2)$$

The comparative analysis for results obtained for MAE for both Traditional Algorithm and Proposed Hybrid Approach are provided in the table II and Figure 4.1 below:

TABLE II: Results of MAE experimented with different neighbors

User ID	K = 7		K = 5		K = 3	
	Traditional Method	Proposed Method	Traditional Method	Proposed Method	Traditional Method	Proposed Method
1	0.95832	0.94164	0.96502	0.94754	0.981077	1.19244
2	0.98242	0.85427	0.9479	1.03615	0.95997	0.986878
3	1.2032	0.611908	0.96259	0.81389	0.987909	1.221844
4	0.96152	1.06758	0.96618	0.62463	0.967379	1.209233
5	0.9775	0.77322	0.9252	1.0138	0.99822	1.074246
6	0.944783	0.819847	0.963943	0.779355	0.975999	1.052224
7	0.943984	0.910714	0.946767	0.672508	0.946538	0.956547
8	0.947131	1.093655	0.971439	1.365897	0.983097	0.903822
9	0.941109	1.02021	0.968706	0.879986	0.98727	0.87595
10	0.941207	0.89962	0.982583	0.790082	0.956825	0.625507

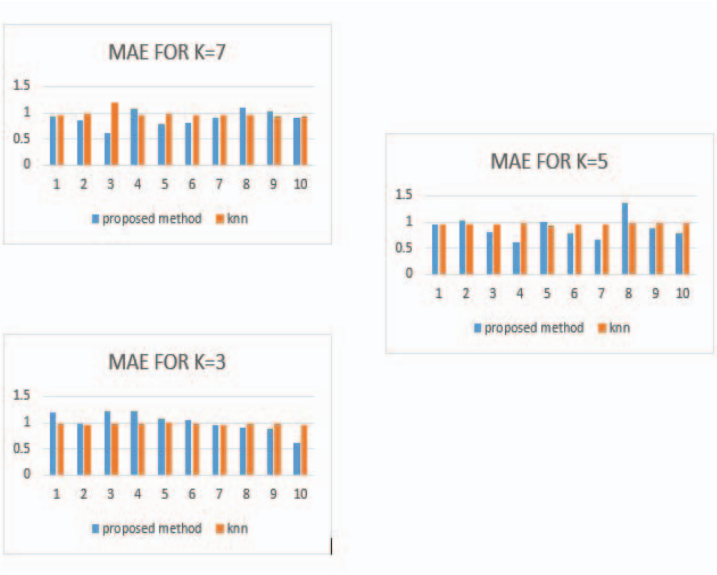

Fig 4.1

Above experiments were performed on 100K movie lens dataset available publicly. Displayed results suggest that for K=7 and K=5, the proposed algorithm provides better accuracy as compared to traditional algorithms. Proposed method also takes on average 32 users for computations, whereas traditional method required no of users equal to that in the system i.e. 978. Hence, proposed method required less computations thereby saving time.

The shortcomings of above experiments are that, because we are randomly selecting a user, it might be possible that

same user is selected more than once thereby giving more weightage to that user. It might be possible that the active user might be selected randomly. Another drawback is that, because we cannot consider only certain users in Apache Mahout for user based collaborative filtering; temporary files need to be generated for each of querying user. This might pose difficulty if the numbers of users active at a given instance of time are quite high.

V. CONCLUSION

In this paper, a novel framework is proposed that combines prediction using user/item based collaborative filtering with the demographic information of users and item profile. K nearest neighbor (KNN) algorithm with threshold value is used which takes into consideration the entries crossing the threshold value and provides first K neighbors. The results showed that threshold KNN dealt only with the entries crossing threshold i.e. reduced dataset and provided faster results without compromising the accuracy of results. Thus, the proposed hybrid system along with providing faster results productively solves cold start, first rater, and curse of dimensionality problem.

ACKNOWLEDGMENT

We would like to thank our faculty guide Mrs. Purvi Kansara to give us such an opportunity to prepare this research paper and inspire us to think innovatively to come up with solution to existing problems. This really helped us improve our ability in understanding the recommendation algorithms. Also big thanks to our university for providing us with such a platform where we can indulge in this kind of research work.

REFERENCES

[1] http://www.webopedia.com/TERM/R/recommender_systems.html

[2] Jyoti Gupta, Jayant Gadge, "A Framework for a Recommendation System Based On Collaborative Filtering and Demographics", *International Conference on Circuits, Systems, Communication and Information Technology Applications (CSCITA)*, 2014.

[3] Jihane KARIM, "Hybrid System for Personalized Recommendations", *IEEE*, 2014

[4] Xiaoyun Wang, Chao Zhou, "A Collaborative Filtering Recommendation Algorithm using User Implicit Demographic Information", *The 7th International Conference on Computer Science & Education (ICCSE 2012)* July 14-17, 2012

[5] Lin Chen, Richi Nayak,Yue Xu , "A Recommendation Method for Online Dating networks based on Social Relations and Demographic Information", *International Conference on Advances in Social Networks Analysis and Mining*, 2011

[6] Long Yun, Yan Yang, Jing Wang, Ge Zhu, "Improving Rating Estimation in Recommender Using Demographic Data and Expert Opinions", *IEEE*, 2011

[7] Qian Wang, Xianhu Yuan, Min Sun, "Collaborative Filtering Recommendation Algorithm based on Hybrid User Model", *Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010)*

[8] Tian Chen, Liang He, "Collaborative Filtering based on demographic attribute vector", *ETP International Conference on Future Computer and Communication*, 2011

[9] Manisha Hiralall, Wojtek Kowalczyk, "Recommender systems for e-shops", *Business Mathematics and Informatics paper*, 2011

[10] Carlos Iv´an Ches˜nevar, Ana Gabriela Maguitman, Guillermo Ricardo Simari, "Argument-Based Critics and Recommenders: A Qualitative Perspective on User Support Systems", *Elsevier Science*, August 31, 2005

[11] Yan Chen, F. Maxwell Harper, Joseph Konstan, Sherry Xin Li, "Social Comparisons and Contributions to Online Communities: A Field Experiment on MovieLens", *Scientific Literature Digital Library and Search Engine*