Prediction Accuracy Comparison Of Similarity Measures In Memory Based Collaborative Filtering Recommender Systems

Jayvardhan[1], Samuel V Thomas[2], Madan Lal Yadav[3]

[1]*MTech CS&E, ASET, Amity University, India*

[2]*MTech CS&E, ASET, Amity University, India*

[3]*Asst. Professor, ASET, Amity University, India*

***Abstract*— At the core of recommender systems are the processes which make predictions, such as ratings for an item which a user may assign to it, and recommends the most preferred item based on these ratings back to the user. Numerous design and implementation issues exist in building such intelligent systems. The success of such systems, are highly dependent on predictive accuracy of the underlying methodologies on which recommender engine is built upon. In this paper we have explored various similarity measures and analyzed their effect on predictive accuracy when applied in building neighbourhood based collaborative filtering recommender systems.**

***Keywords*— collaborative filtering, data mining, prediction, recommender systems, similarity measures.**

I. Introduction

The plethora of information on internet and the success of e-commerce websites have given rise to an indispensable need of systems capable of filtering information in meaningful ways. From the diverse variety of products and services offered at websites, users find it difficult to make decision in selecting items. Various information retrieval systems and tools are available to facilitate this process. Amongst the popular ones, there are Recommender Systems (RSs) which is built to provide suggestion of most preferred items to the user from all the available alternatives.

RSs are information processing software that provides suggestion of items to its users. In RSs item is used as generic term to denote product or service that are recommended to its user. The idea originated from individual behaviour of decision making being influenced by the choices of their peers or persons of similar tastes. The researchers tapped this idea to provide richer personalised recommendations to its users in form of ranked lists of items.

Goldberg et al. the developers of one of first RSs, Tapestry, in their paper [2] coined the term collaborative filtering (CF). CF emerged as one of the most successful approaches for providing personalized recommendations. The fundamental assumption of CF is that people who like the same thing are likely to feel similarly towards other things. The Memory Based CF algorithm uses entire or sample of user-item database (which typically stores the user ratings for the corresponding items) and loads it into the memory to generate prediction in order to make recommendations. The item recommendation task is described in section III. The Neighbourhood-based CF algorithm is a most popular mechanism for implementing memory based CF algorithm. It identifies the neighbours (where every user is part of group with similar taste) for the active user by calculating weight or similarity which is distance between two users or items and produces prediction by taking weighted average of all the ratings. In this approach similarity computation between users or items is important step.

The prediction quality of recommender system based on collaborative filtering technique is highly dependent on the precision of similarity between users or items. Thus choice of similarity computation measure is crucial step in building this kind of recommender systems. In this paper we have build neighbourhood based CF Recommender System for user based and item based approach over a dataset using various Similarity Measures and compared them on their predictive accuracy in order to facilitate selection of better similarity measure in building such systems.

II. Related Work

RSs became an independent research area in mid of 1990s [1, 2]. Xiaoyuan Su and Taghi M. Khoshgoftaar in their paper [3] have provided a comprehensive introduction on collaborative filtering technique. Sarwar et al. in [4] have analysed item based recommendation generation algorithms and have examined different techniques for computing item-item similarity. In another work [5] Sarwar et al. have analysed quality and performance of recommendation on two different dataset using collaborative filtering techniques. Herlocker et al. in [6] have presented very detailed and extensive techniques for evaluating collaborative filtering recommender systems.

III. Problem Definition And Similarity measures

In order to define the problem for comparing predictive accuracy of various similarity measures in memory based Collaborative Filtering Recommender Systems we need to first formalise the task of item recommendation, prediction and similarity measures.

A. Item Recommendation Task

Taking clue from [7] let the set of users in the system be denoted by U and the set of items by I. R denotes the set of ratings stored in the system given by users for items and S records the set of possible values([1,5] or {like, dislike}) for a rating. Any user $u \in U$ can assign only one rating for a particular item $i \in I$ denoted by r_{ui}. The subset of users who have rated an item i is denoted by U_i and the subset of items rated by a user u by I_u. Apart from this the set of items which have been rated by two user's u and v is denoted by I_{uv} and likewise the set of users that have rated both items i and j is denoted by U_{ij}. The task of item recommendation primarily focuses on recommending a user $u \in U$ a new and not yet experienced item $i \in I$ that may be relevant to the user's interest. This task is accomplished by calculating a function f : U x I → S that predicts the rating f(u,i) of a user u for a new item i. This function is then used to recommend to the active user u_a an item i^* for which the estimated rating has the highest value:

$$i^* = \arg\max_{j\varepsilon I} f(u_a , j). \tag{1}$$

B. Prediction

In neighbourhood based CF system [8] a subset of nearest neighbour, based on similarity, of active user are chosen and a weighted aggregate of their similarity is used to generate prediction using the following formula:

$$P_{a,i} = \bar{r}_a + \frac{\sum_{u \in U}(r_{u,i} - \bar{r}_u) * w_{a,u}}{\sum_{u \in U}|w_{a,u}|} \tag{2}$$

where $\bar{r}_a$ and $\bar{r}_u$ are the average ratings for the user a and user u for all other rated items. $w_{a,u}$ is the weight between the user a and u which is used to calculate similarity correlation. There are different ways to calculate this weight which is discussed in the following section.

In the case of item-based collaborative filtering recommendation [4] the method of similarity computation between two items i and j is to first isolate all the users who have rated these items and then to apply similarity computation technique. The Prediction generation formula used in this case is:

$$P_{u,i} = \frac{\sum_{n \in N} r_{u,n}\, w_{i,n}}{\sum_{n \in N} |w_{i,n}|} \tag{3}$$

where the summations are over all other rated items $n \in N$ for users u, $w_{i,n}$ is the weight between items i and n, $r_{u,n}$ is the rating for user u on item n.

C. Similarity Measures

In order to compute similarity between users different similarity calculation techniques have been used. The following sub-sections discuss some of the used similarity computation technique.

1. Pearson Correlation Similarity: In statistics correlation is used to find how well two random variables are related. Correlations are important because they discover a predictive relationship between the concerned variables which can be exploited into practice. Pearson Correlation, more formally known as Pearson Product Moment Correlation or PPMC, is one of the most important correlation based measure which is widely used. In simplest form Pearson Correlation is the measure of the linear correlation between two values X and Y giving value between +1 to -1 inclusively. In user-based collaborative filtering recommender system, the similarity weight between user u and v is calculated using following formula:

$$w_{u,v} = \frac{\sum_{i \in I_{uv}}(r_{ui} - \bar{r}_u)(r_{vi} - \bar{r}_v)}{\sqrt{\sum_{i \in I_{uv}}(r_{ui} - \bar{r}_u)^2 \ \sum_{i \in I_{uv}}(r_{vi} - \bar{r}_v)^2}} \tag{4}$$

where I_{uv} the set of items which have been rated by two user's u and v. r_{ui} rating given by user u to item i and $\bar{r}_u$ is mean of ratings given by user u.

2. Euclidean Distance Similarity: Euclidean distance is another famous distance measuring formula used in mathematics. The formal definition of Euclidean distance measurement states that it is the square root of the sum of squared differences between corresponding elements of two vectors. This concept of measuring distance is adopted in calculating similarity between users by considering users as elements in Euclidean space whose coordinate value is their preference. The similarity metric then calculates the distance d between two such user points using following formula:

$$d_{uv} = \sqrt{\sum_{i=1}^{n}(x_{ui} - x_{vi})^2} \tag{5}$$

where i is item, u and v represents users, x_{ui} and x_{vi} are ratings given by user u and v for item i. But this equation alone doesn't form the similarity metric as the larger value indicates more distance making the users less similar. So it's required to minimize this value so that it users can be found to be more similar. Therefore the value d_{uv} is minimized and similarity for each pair of users is calculated and returned by following formula:

$$\frac{1}{1 + d_{uv}} \tag{6}$$

which gives value in the range of 0 and 1.

3. Tanimoto Coefficient Similarity: Tanimoto Coefficient also known as Jaccard Similarity Coefficient measures similarity between datasets, and is defined as the size of intersection divided by the size of the union of the datasets. The formula is:

$$T(X, Y) = \frac{X \cap Y}{X \cup Y} \tag{7}$$

where X and Y defines elements in datasets. The adaptation of Tanimoto Coefficient in context of calculating similarity in

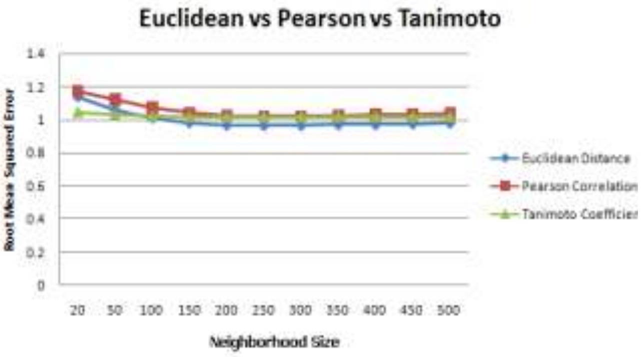

Figure 1 Combined Result for User Based CF showing comparative accuracy performance between three similarity metrics

collaborative filtering recommender system is very interesting. The important fact about Tanimoto Coefficient is that it ignores what preference value is given to the item. This means that it only takes into consideration that whether the user has expressed some preference for the item or not. This makes Tanimoto Coefficient as an important similarity metric because it can be used in context where user's preference for the item is not very detailed. For example irrespective of giving ratings on the scale of 1 to 5 a user can assign values like good or bad. Thus the only considerable thing is the relationship between user and item which is recorded as transaction by the system.

IV. DATASET, EVALUATION METRIC AND EXPERIMENTAL RESULT ANALYSIS

In this section used dataset, accuracy evaluation metric and result analysis is presented.

A. Dataset

In order to perform any empirical analysis pertaining to information system need of consistent dataset is indispensible. The empirical analysis in this paper is carried out by implementing memory based collaborative filtering technique on the movie ratings dataset. The dataset has been procured from Grouplens Research website[1]. The used dataset consists of 100,000 ratings from 943 users on 1682 movies.

B. Evaluation Metric

The task of item recommendation, prediction generation and similarity measurement was formalised and detailed in section III. However, one key aspect in implementing such system is about analyzing how well the system is generating prediction for which accuracy is considered in evaluating the performance of recommender system. For this purpose some evaluation metric [6] is needed. Mean Absolute Error (MAE) and its variation Root Mean Squared Error (RMSE) falls under the category of predictive accuracy metric. For purpose of evaluating predictive accuracy in this paper RMSE metric has been used owing to the fact that it penalises less accurate predictions more heavily compared to MAE. The RMSE metric, which measures the square root of the average of the squares of the difference between predicted and actual rating values, is calculated using following formula:

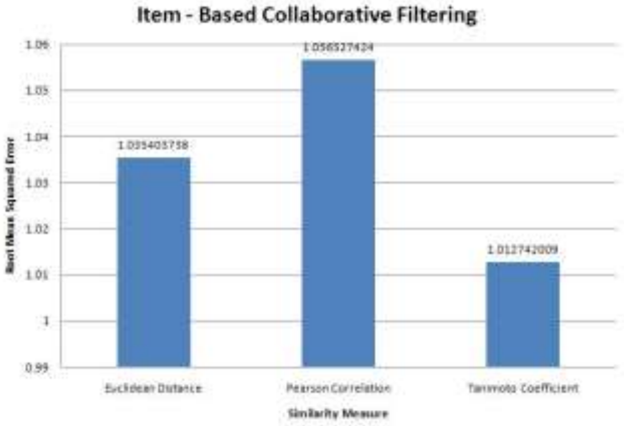

Figure 2 Combined Result for Item Based CF showing comparative accuracy performance between three similarity metrics

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(p_i - r_i)^2}{N}} \qquad (8)$$

where i is item, p_i is predicted rating and r_i is the actual rating for item I. N is the number of total items under consideration. In analyzing the accuracy the lower value of RMSE presents better performance of recommender engine.

C. Result Analysis

In user based collaborative filtering technique we have tuned the system on neighborhood size of k-nearest neighbors which influences the prediction quality. The result has been presented in fig. 1 which plots predictive accuracy of system in various similarity measure approaches on the dimensions of root mean squared error and neighborhood size.

The Pearson Correlation oriented User Based Collaborative filtering recommender system shows best prediction around neighborhood size of 250 for which its prediction accuracy on RMSE factor was around 1.0208. The worst accuracy was recorded when neighborhood size was 20. This shows that Pearson Correlation similarity metric is not suited for sparse dataset. The Euclidean Distance based Similarity measure has shown better result as compared to Pearson Correlation. The best RMSE accuracy score on this dataset was 0.9659 at the neighborhood size of 250. Also in case of data sparsity this metric was too found not very effective. The Tanimoto Coefficient similarity metric was found to be moderately good as compared to Pearson Correlation in the context that it even performed well for sparse dataset. However for combined

[1] http://www.grouplens.org/node/73

comparison of these metrics upon the used dataset Euclidean Distance Similarity Measure emerged to be better option in implementing user based collaborative filtering recommender system.

In analysis of Item Based Collaborative filtering Recommender System the prediction accuracy of Tanimoto Coefficient was found to be better as compared to other. The RMSE score for Tanimoto Coefficient was 1.01, followed by Euclidean Distance similarity score of 1.035 and then was Pearson Correlation with 1.056 which is presented in fig.2. The most significant aspect was that as compared to User based CF algorithms Item based CF algorithms were very fast in computing similarity weight and prediction. This makes the choice of Item based approach more favorable over User Based approach.

V. Conclusion And future work

In Recommendation System accuracy of prediction serves as the most prominent source for building trust of customers. In this paper we explored the importance of three different similarities metric on two important variants of memory based collaborative filtering technique. We compared the prediction accuracy performance of these systems with respect to discussed similarity metric and seen its influence on the accuracy of prediction. However it should be noted that the performance of recommender engines are also dependent on data sparsity which immediately makes the monitoring and tuning of these systems a requirement. Scalability issue is another challenge that needs to be addressed in building these systems. This gives an opportunity to implement new algorithms and test new design approach and further improvise prediction accuracy of these systems.

References

[1] Anand, S.S., Mobasher, B.: Intelligent techniques for web personalization. In: Intelligent Techniques for Web Personalization, pp. 1–36. Springer (2005)

[2] Goldberg, D., Nichols, D., Oki, B.M., Terry, D.: Using collaborative filtering to weave an information tapestry. Commun. ACM 35(12), 61–70 (1992)

[3] Xiaoyuan Su and Taghi M. Khoshgoftaar, "A Survey of Collaborative Filtering Techniques," Advances in Artificial Intelligence, vol. 2009.

[4] Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2001, April). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web* (pp. 285-295). ACM.

[5] Sarwar, B., Karypis, G., Konstan, J., & Riedl, J. (2000, October). Analysis of recommendation algorithms for e-commerce. In *Proceedings of the 2nd ACM conference on Electronic commerce* (pp. 158-167). ACM.

[6] Herlocker, J.L., Konstan, J.A., Terveen, L.G., Riedl, J.T.: Evaluating collaborative filtering recommender systems. ACM Transaction on Information Systems 22(1), 5–53 (2004).

[7] Ricci, Francesco, and Bracha Shapira. *Recommender systems handbook*, pp. 108-109. Springer, 2011.

[8] Herlocker, J. L., Konstan, J. A., Borchers, A., & Riedl, J. (1999, August). An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 230-237). ACM.