6. EVALUATION OF THE RECOMMENDATION SYSTEM

Domain specific recommender systems are built to satisfy the growing demand from the users. However, the recommendation approaches or the algorithms used for designing such recommender systems should be evaluated to validate the recommendations made by these recommender systems. In order to evaluate the different recommendation approaches, the property of the recommendation engine to be evaluated such as the accuracy of prediction, scalability, robustness, etc. should be identified based on the application for which the recommender system is being built. This chapter explores the different approaches available for evaluating the recommendation engines and applies the selected evaluation metrics for evaluating the four different recommendation approaches applied in this research work.

6.1 APPROACHES FOR EVALUATING RECOMMENDATION

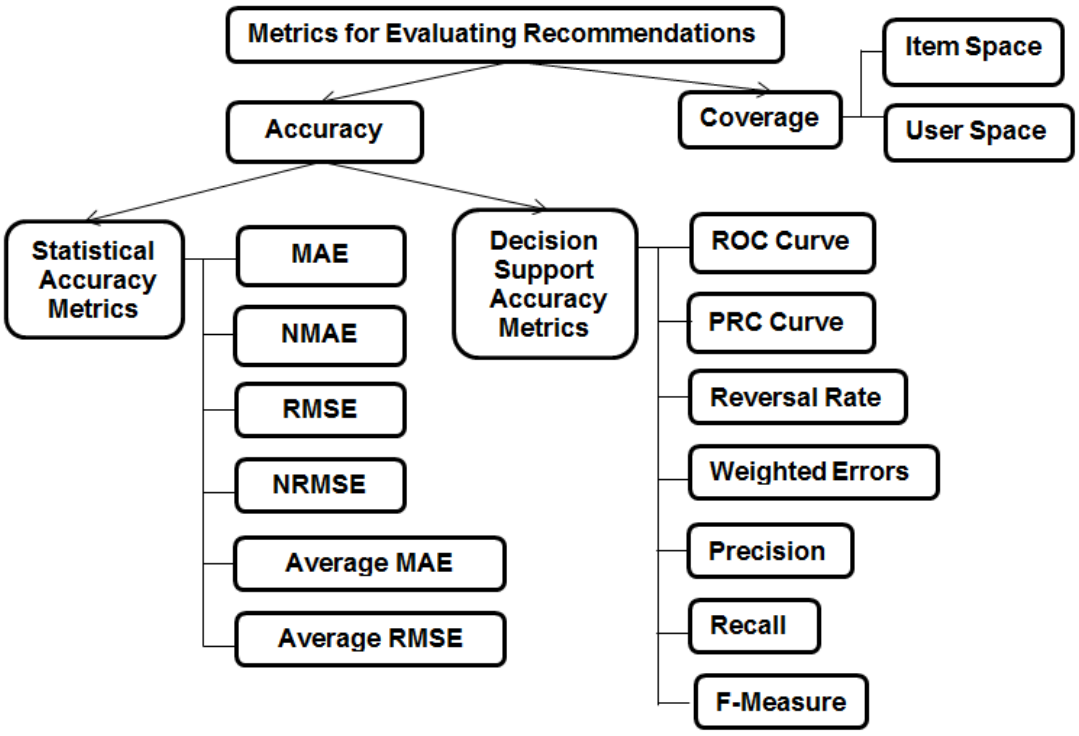


Figure 6.1 Metrics for Evaluating Recommendations

Many approaches have been proposed for evaluating the recommendation systems as given in Figure 6.1. The method of evaluation is usually selected based on the various factors like the type of data, the research issue considered, feasibility of collecting the information required for evaluation, etc. Some of the approaches discussed in the literature for evaluating the recommender systems are discussed below in detail.

The recommendation algorithm quality can be measured based on two types of metrics namely the accuracy and the coverage [147,148]. Based on the type of filtering technique used, the metric to be evaluated has to be chosen.

a) Metrics for Measuring Accuracy of Recommendations

The Accuracy of the recommendations made is defined as the fraction of recommendations that are correct out of the total recommendations that is possible. The accuracy metrics can be further divided into statistical and decision support accuracy metrics [148]. The choice of the metric should be made based on the dataset type, features and the recommender system task like whether it would rank the items or would recommend the top 5 or 10 items etc.

- **Statistical Accuracy Metrics**

The Statistical accuracy metrics are popularly used to measure the accuracy of the predicted ratings in recommender systems. The predicted ratings are directly compared with the actual user rating and the accuracy of a filtering technique is evaluated. Some of the popular metrics used to measure the rating prediction accuracy of the recommender systems are discussed here.

Mean Absolute Error (MAE) is the commonly used metric and it measures the deviation of the recommendations made from that of the user's specific value. The accuracy of the recommendation engine is inversely proportional to the MAE value. It is computed using the formula given below in Equation 6.1.

$$MAE = \frac{1}{N} \sum_{u,i} |p_{u,i} - r_{u,i}| \quad \text{..... (6.1)}$$

Where, $P_{u,i}$ indicates the predicted rating for item 'i ' by user 'u';

$r_{u,i}$, indicates the actual rating of the user for the item;

N indicates the total number of ratings available for the particular item set.

Root Mean Square Error (RMSE) metric is another metric used widely. It calculates the absolute error and the accuracy of the recommendation engine is inversely proportional to the RMSE value and hence lower values indicate high accuracy as can be calculated using the formula given below in Equation 6.2.

$$RMSE = \sqrt{\frac{1}{n} \sum_{u,i} (p_{u,i} - r_{u,i})^2} \quad \text{..... (6.2)}$$

Normalized MAE (NMAE) and Normalized RMSE (NMRSE) is the normalized versions of MAE and RMSE in which the ranking of the algorithms is normalized based on the ratings.

Average MAE and Average RMSE is used if the test sets are unbalanced. Unbalanced test sets influence the MAE and the RMSE values heavily because of the errors even in few items, if there are very frequent. Hence, it is preferred to calculate the MAE and RMSE values of each item separately and then calculate the average of MAE and RMSE.

- **Decision Support Accuracy Metrics**

Decision support accuracy metrics are used for evaluating the recommendation accuracy of usage prediction or when a very high quality item has to be selected from a set of available items [149,151]. For example, the objective of some recommender systems would be to recommend some items that the user may interested to use and not to predict the ratings of the items. The Decision support accuracy metrics can be used to evaluate such recommender systems. Some of the popular metrics under the

Decision support accuracy metrics category include Precision, Recall and F-measure, Receiver Operating Characteristics (ROC) curve, Precision Recall Curve (PRC), Reversal rate, weighted errors, etc. To evaluate the usage prediction of a recommender system, the selected items of a test user that is kept unrevealed is compared with the predicted items for the same test user by the recommender system. The possible results that could be obtained in this comparison can be classified as shown in Table 6.1 below and given in the work of Shani et al [150].

Table 6.1 Classification of Possible Results of Recommendations

	Recommended	**Not Recommended**
Used	True Positive (Tp)	False Negative (Fn)
Not used	False Positive (Fp)	True Negative (Tn)

The True positives (T_P) are the relevant items that are to be recommended and are recommended correctly. Whereas, the False negatives (Fn) are relevant items that are to be recommended but are not recommended correctly. The False positives (Fp) are the non relevant items that should have not been recommended but were recommended to the users. The True negatives (Tn) indicate the non relevant items that should have not been recommended and were actually not recommended to the users either. Based on these classifications of possible results of recommendations, the metrics like include Precision, Recall and F-measure are calculated as given below.

Precision metric (P) is also called as the positive predictive value and is defined as the fraction of retrieved instances that are relevant to the query as given below in Equation 6.3.

$$P = \frac{\text{No of Correctly Recommended Items}}{\text{No of Recommended Items}} \quad \ldots\ldots (6.3)$$

Recall metric (R) is also known as the sensitivity. It is defined as the fraction of relevant instances that are retrieved successfully as below in the Equation 6.4.

$$R = \frac{\text{No of Correctly Recommended Items}}{\text{No of Interesting Items}} \quad \text{..... (6.4)}$$

F1 metric is also known as the F-score or the F-measure and it is a measure of the accuracy of a test. The F1 score of a test is calculated by considering both the precision p and the recall r of the test. An F1 score is best when its value is at 1 and worst when it is at 0. It is calculated as given below in Equation 6.5.

$$F1 = \frac{2 * \text{Recall} * \text{Precision}}{\text{Recall} + \text{Precision}} \quad \text{..... (6.5)}$$

Precision-Recall and ROC curves are the other useful methods that could be used. Precision-Recall curves are used to measure the proportion of recommended items that are actually preferred or selected by the user. Whereas, the ROC (Receiver Operating Characteristic) curves are used to measure, the proportion of items that are actually not preferred by the user but were recommended.

b) Metrics for Measuring Coverage of Recommendations

Coverage [149] metric is used when one has to measure the fraction of objects in the search space for which the system is able to provide recommendations. The two types of Coverage metrics are Item Space Coverage and User Space Coverage. The Item Space Coverage is also known as the Catalog Coverage. It is defined as percentage of all the relevant items that can be recommended to the user. User Space Coverage metric measures the user profile richness that is required to start generating recommendations to the particular user. For example, it is popularly used in evaluating the collaborative filtering algorithms to measure the number of items that a user should have rated to generate recommendations to that particular user.

6.2 SELECTED METRICS FOR EVALUATING THE PROPOSED RECOMMENDATION ENGINE

The metrics selected for evaluating the proposed recommendation that overcomes the cold start problem in recommending the cloud renderfarm services are Precision (P), Recall (R) and the F1 metric.

- **Precision (P)** metric is the ratio of the number of correctly recommended items (True positives) to the number of recommended items (True positives + False positives) as given below in Equation (6.6).

$$P = \frac{\#Tp}{\#Tp + \#Fp} \quad \text{..... (6.6)}$$

- **Recall (R)** metric is the ratio of the number of correctly recommended items (True positives) to the number of interesting items to the user (True positives + False negatives) as given below in Equation (6.7).

$$R = \frac{\#Tp}{\#Tp + \#Fn} \quad \text{..... (6.7)}$$

- **F1 Metric** is also known (1-Specificity) and is calculated by considering both the Precision (P) and the Recall (R) values as given below in Equation (6.8). An F1 score is best when its value is at 1 and worst when it is at 0.

$$F1 = \frac{\#Fp}{\#Fp + \#Tn} \quad \text{..... (6.8)}$$

The values of the above defined three metrics is calculated for each of the four different approaches, namely the Knowledge based filtering using ontology (K), Knowledge based method with MCDM (K+ MCDM), the integrated method that combines the results of the above two metrics with the average QoE of the service and re-ranks the services and the final approach evaluated is the filtering of services based

only on the average QoE metric. The experiments conducted and the results obtained are discussed below in the following sections.

6.3 RESULTS AND DISCUSSION

The results obtained for the three selected evaluation metric namely, the Precision (P), Recall (R) and the F1 metric for all the four approaches, namely, the Knowledge based filtering using ontology (K), Knowledge based method with MCDM (K+MCDM), the integrated method that combines the results of the above two metrics with the average QoE of the service and re-ranks the services (Integrated) and finally the approach to rank the services based only on the average QoE metric (QoE) are given below. It could be observed from Table 6.2 and Figure 6.1, that the Precision metric (P) value for the Knowledge based filtering using ontology (K) method calculates to the least value of 0.45. The Precision metric (P) value of Knowledge based method with MCDM (K+MCDM) calculates to 0.795 and QoE based recommendation method is 0.64.

Whereas, the Precision metric (P) value of the integrated method that combines the results of the Knowledge based filtering and the MCDM ranking with the average QoE rating of the services is used for generating recommendations is 0.894 and is the highest Precision metric (P) value compared to the other approaches.

Similarly, the Recall values of all the recommendation approaches are given in Table 6.3 and Figure 6.2. The Recall metric value (R) for the Knowledge based filtering using ontology (K) method calculates to the least value of 0.45.. The Recall metric value (R) of Knowledge based method with MCDM (K+MCDM) calculates to 0.795 and that of the QoE based recommendation approach is 0.64. Whereas, the Recall metric value (R) of the integrated method is also high and calculates to 0.894.

Table: 6.2 Precision (P) Metric Values

No of Recommendations	K	K+MCDM	Integrated	QoE
50	0.33	0.602	0.725	0.414
100	0.351	0.637	0.753	0.442
150	0.431	0.65	0.767	0.554
200	0.422	0.668	0.784	0.573
250	0.383	0.751	0.792	0.623
300	0.433	0.759	0.879	0.628
350	0.44	0.789	0.89	0.632
400	0.45	0.795	0.894	0.64

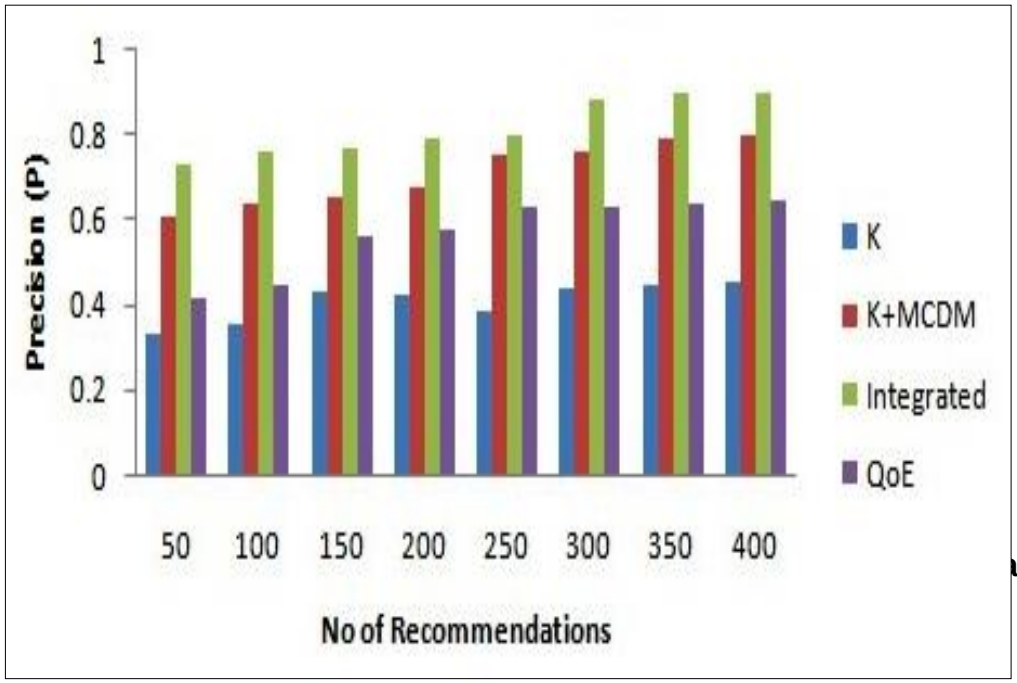
Table: 6.3 Recall (R) Metric Values

No of Recommendations	K	K+MCDM	Integrated	QoE
50	0.62	0.699	0.798	0.58
100	0.64	0.723	0.835	0.612
150	0.652	0.732	0.852	0.623
200	0.712	0.746	0.862	0.628
250	0.732	0.749	0.869	0.636
300	0.738	0.755	0.873	0.648
350	0.741	0.776	0.879	0.656
400	0.744	0.782	0.89	0.675

Table: 6.4 F1 Values Metric Values

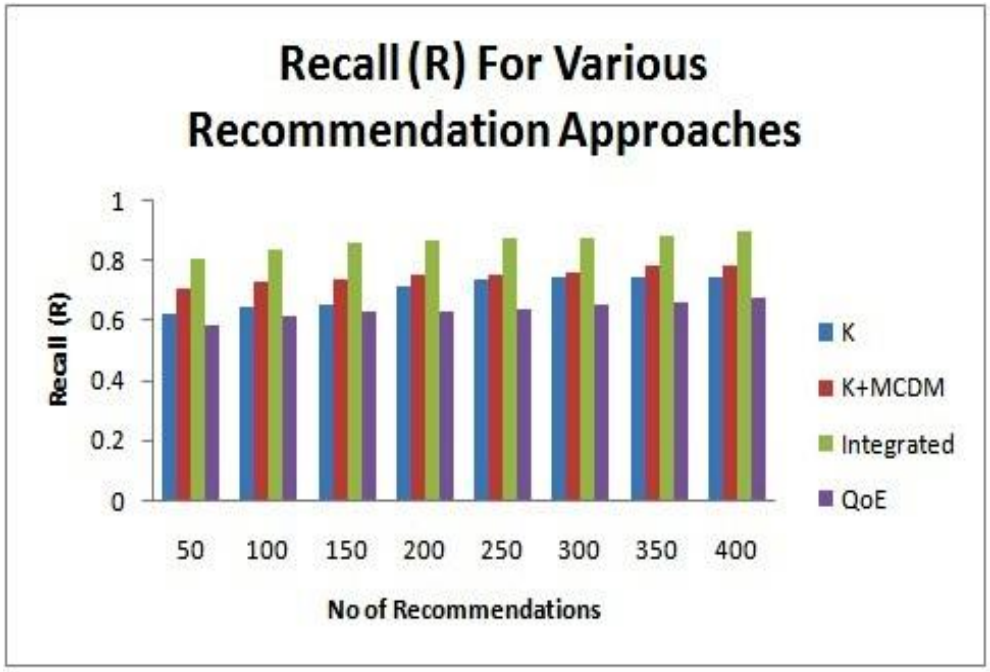
No of Recommendations	K	K+MCDM	Integrated	QoE
50	0.437	0.646	0.764	0.45
100	0.442	0.677	0.791	0.462
150	0.453	0.697	0.81	0.483
200	0.472	0.715	0.827	0.52
250	0.4825	0.789	0.838	0.54
300	0.49	0.8	0.845	0.63
350	0.495	0.82	0.86	0.65

400	0.54	0.86	0.89	0.7



aches

Figure 6.2 Precision (P) for Various Recommendation Approaches



hes

Figure 6.3 Recall (R) for Various Recommendation Approaches

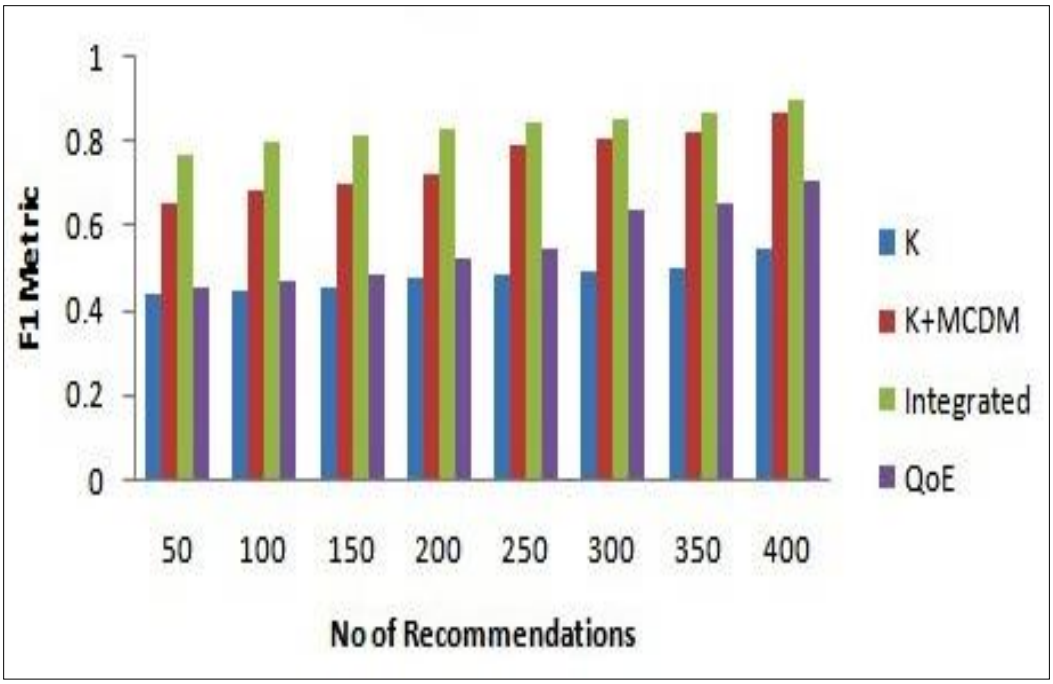


Figure 6.4 F1 Values for Various Recommendation Approaches

The F1 values of all the recommendation approaches given in Figure 6.3 and Table 6.4. The The F1 values of the the Knowledge based filtering using ontology (K) method also evaluates to be the lowest at 0.54, the QoE based recommendation approach at 0.7, the Knowledge based method with MCDM (K+MCDM) at 0.86. The F1 values of the integrated method is also the highest at 0.89.

Thus, It is evident from the above discussions, that the proposed integrated methodology, overcomes the cold start problem, as recommendations could be

generated with acceptable precision and accuracy. The following chapter concludes the research work and discusses the limitations of this work and the scope for further work.