Content Based and Collaborative Filtering for Online Movie Recommendation

Archana T. Mulik

Abstract - this research paper highlights the importance of content based and collaborative filtering to suggest item for the customer such as which movie to watch or what music to listen. Recommendation system plays an important in increasing sale of the product, customer satisfaction, increase sale of diverse product etc.
In order to increase sale of the product, every organization concern with increase the new customer and retain the existing customer with the organization. Traditional business had limitation of geographical location, but with the new era business spread all over the world. With the help of technological innovation e-business grows rapidly. Customer purchase the item using online store, the only limitation is to search the item in the store on its own; no helping hand available online, in this scenario product recommender system is very useful.

Index Terms— Content, Collaborative Filtering, Ranking, Similarity, cluster, Rating.

I. Overview of PROPOSED system

Recommender system techniques are classified into three categories: content based, collaborative and hybrid approaches. Content based approach recommends items similar to the user preferred in the past. Collaborative filtering approach suggests items that users with similar preferences have liked in the past. Hybrid approach can combine both content based and collaborative filtering approaches. The proposed system uses hybrid approach. Generally recommender system performs the following two tasks while providing recommendations to each user. First, the ratings of unrated items are predicted based on the available information using some recommendation algorithm. And second, the system finds items that maximize the user's utility based on the predicted ratings, and recommends them to the user.

Item based collaborative filtering technique

This technique uses the set of items the active user has rated and computes the similarity between these items and target item i and then selects N most similar items $\{i_1,i_2,\ldots,i_N\}$. Item's corresponding similarities also $\{s_{i1},s_{i2},\ldots,s_{in}\}$ are also computed. Using the most similar items, the prediction is computed.

Item similarity computation

1. Similarity computation between two items i (target item) and j is to first find the users who have rated both of these items.
2. There are number of different ways to compute similarity. The proposed system uses adjusted cosine similarity method which is more beneficial due to the subtracting the corresponding user average from each co-rated pair. Similarity between items i and j.

II. Related Work

Adomavicius G., Y. Kwon proposed, a number of item ranking techniques. These ranking techniques can generate suggestions that have higher aggregate diversity for all users while maintaining the recommendation accuracy. In this proposed approach they have considered additional factors, such as item popularity, when ranking the recommendation list to increase recommendation diversity with minimum accuracy loss. These studies say that the recommendation's quality can be computed along a number of dimensions, and only the accuracy of recommendations is not sufficient to find the most appropriate items for each user. One of the goals of recommender systems is to provide more diverse recommendations [6].

A. Ghose, and P. Ipeirotis proposed two ranking mechanisms for ranking product reviews: a consumer-oriented ranking mechanism ranks the reviews according to their expected helpfulness, and a manufacturer-oriented ranking mechanism ranks the reviews according to their expected effect on sales. Ranking mechanism combines econometric analysis with text mining techniques in general, and with subjectivity analysis in particular. To decide whether to buy a product, consumer as expected attracts to reading reviews. However, for a single product the more number of reviews are typically published makes it difficult for individuals to find the best reviews and realize the true quality of a product based on the reviews. Similarly, the manufacturer of a product needs to identify the reviews that control the customer base, and examine the content of these reviews. They showed that subjectivity analysis can give useful information about the helpfulness or benefit of a review and about its impact on sales. Their results can have a number of implications for the market design of online opinion forums [7].

Neal Lathiax Showed that temporal diversity is an important criterion for quality of recommender systems, by showing how CF data changes over time and performing a user survey. Then they evaluated three CF algorithms from the point of view of the diversity in the sequence of recommendation lists they produce over time and examine how a number of characteristics of user rating patterns affect diversity. They then proposed and evaluated set methods that maximize temporal recommendation diversity without extensively penalizing accuracy. However, current evaluation techniques pay no attention to the fact that users continue to rate items over time: the temporal characteristics of the system's top-N recommendations are not investigated. In particular, it is useless of measuring the extent that the same items are being recommended to users over and over again [8].

III. Proposed Work

The system is to increase the diversity of recommendations with only a negligible accuracy loss as well as recommend a sequence of items instead of a single recommendation and use consumer-oriented or manufacturer oriented ranking mechanisms.

While measuring recommendation quality, only accuracy is not sufficient. Therefore, using the item ratings and user profiles, recommender system has been proposed to provide diverse recommendations. The system algorithm derive recommendation using similarity computation, system predicted rating estimation, implementation of rank generation, item sequence generation, and implementation of consumer or manufacturer oriented ranking mechanism. The system proposes following steps:

1. It is necessary to estimate ratings for the items that have not been seen by a user. For recommender system, collaborative filtering, content based approaches will be used. In collaborative filtering approach, First system will compute the similarity between target item and other items using adjusted cosine similarity method. Thus, system will get most similar items with target item. System-predicted ratings i.e. unknown ratings for item will be calculated by weighted sum technique using previously calculated similarity computation results.
   In content-based approach, recommend items similar to those that a user liked in the past. Target item will be compared with items previously rated by the user. The profile of user contains tastes and preferences of this user. Cosine similarity method will be used to estimate rating of item by comparing user preferences present in user profile and item features that are represented as item attributes in item profile. Finally, we will combine the outputs obtained from both approaches i.e. collaborative filtering and content based approaches.
2. 2.Using item popularity-based parameterized ranking approach, ranks will be generated for items based on their popularity. User will get recommended list of top-N items. Recommendations will increase recommendation diversity while maintaining the accuracy.
3. 3. A consumer-oriented ranking mechanism will rank the reviews according to their expected helpfulness and a manufacturer-oriented ranking mechanism will rank the reviews according to their expected effect on sales with the help of text mining techniques examine the actual text of the review to identify which review is expected to have the most impact on sales.

Item similarity computation

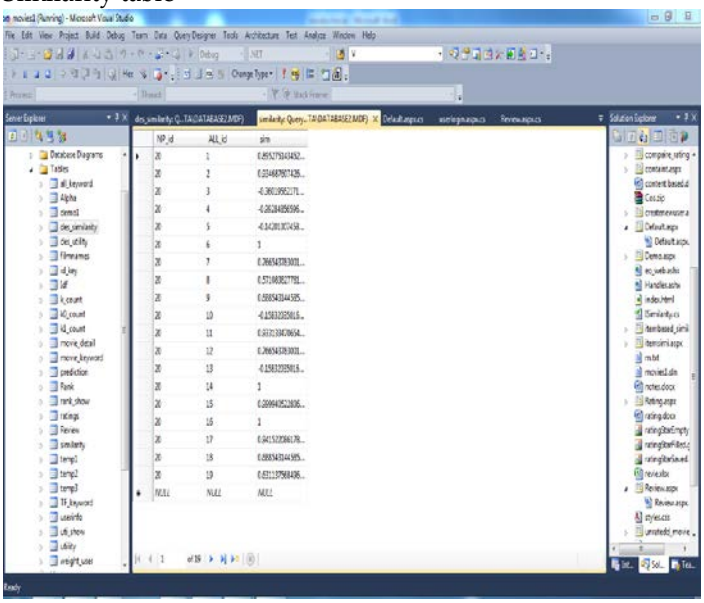
3. Similarity computation between two items i (target item) and j is to first find the users who have rated both of these items. There are number of different ways to compute similarity. The proposed system uses adjusted cosine similarity method which is more beneficial due to the subtracting the corresponding user average from each co-rated pair. Similarity between items i and j is given by [13]:

4.

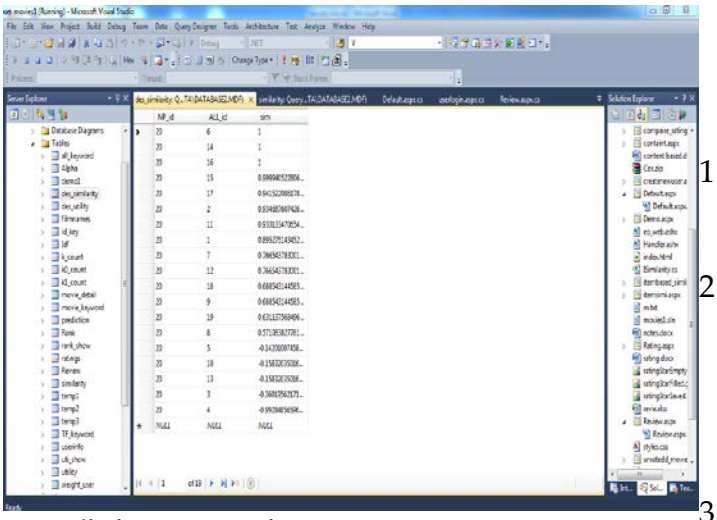
$$sim(i,j) = \frac{\sum_{u \in U}(R_{u,i} - \bar{R}_u)(R_{u,j} - \bar{R}_u)}{\sqrt{\sum_{u \in U}(R_{u,i} - \bar{R}_u)^2}\sqrt{\sum_{u \in U}(R_{u,j} - \bar{R}_u)^2}}.$$

Here $\bar{R}_u$ is the average of the u-th user's ratings.

Similarity table



Similarity in desending order table



A. Prediction computation

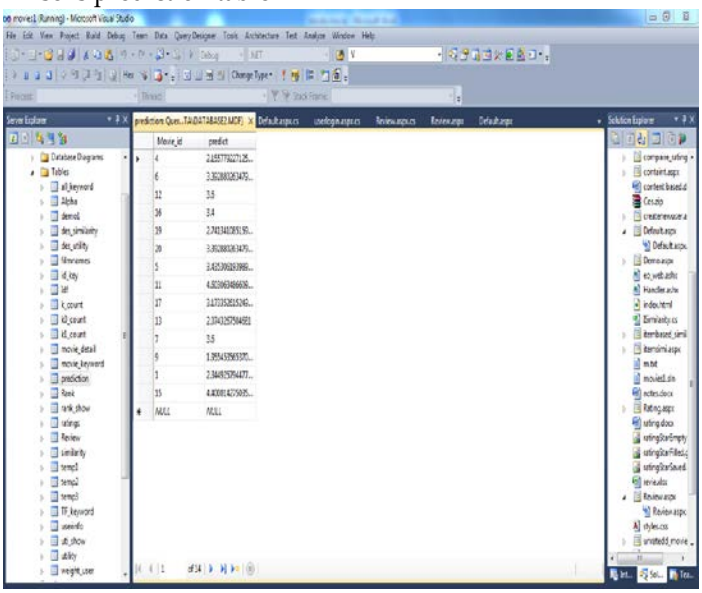
To obtain the predictions weighted sum approach is used.

1. Weighted sum computes the prediction on an item i for a user u by computing the sum of ratings given by the user on the items similar to i. Each ratings is weighted by the corresponding similarity $s_{i,j}$ between items i and j.
2. That weighted sum is scaled by sum of the similarity terms to make sure the prediction is within the predefined range. Prediction on an item i for user u is given by [13]:

$$P_{u,i} = \frac{\sum_{\text{all similar items, N}} (s_{i,N} * R_{u,N})}{\sum_{\text{all similar items, N}} (|s_{i,N}|)}$$

1st Select top 5 values from des_similarity table Implement above formula and calculate predication

Thise is prediction table



End collabirity algo.

B. Algorithm for item ranking

Predicted unknown ratings, calculated in previous steps, are used for item ranking

1. Ratings of items are integers between 1 and 5, where high value represents most liked item. Thus items greater than 3.5 rating as highly ranked i.e. threshold for high ratings(T_H).
2. According to standard ranking method, predicted rating value is used as ranking criteria. Rank of item i is equal to its predicted rating value as follows [1]:

$$rank_{\text{Standard}}(i) = R^*(u, i)^{-1}.$$

Where $R^*(u, i) = P(u, i)$.

3. According to item popularity based ranking method, item ranking is based on their popularity from lowest to highest. Popularity is represented by number of known ratings that each item has. Rank of item i is as follows [1]:

$$rank_{\text{ItemPop}}(i) = |U(i)|, \text{ where } U(i) = \{u \in U \mid \exists R(u, i)\}.$$

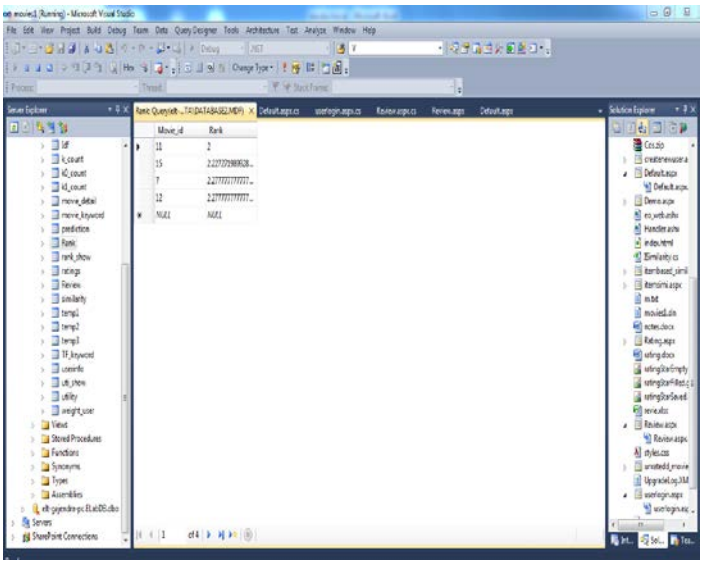
In proposed ranking method, ranking threshold is used. Ranking threshold $T_R \in [T_H, T_{max}]$ where T_{max} is highest possible rating i.e. T_{max} =5. T_R allows to user to choose a certain level of recommendation accuracy. Using standard ranking and Item popularity ranking methods, Item popularity based parameterized ranking method for item i with ranking threshold T_R is given by [1],

$$rank_X(i, T_R) = \begin{cases} rank_X(i), & \text{if } R^*(u,i) \in [T_R, T_{\max}] \\ \alpha_u + rank_{\text{Standard}}(i), & \text{if } R^*(u,i) \in [T_H, T_R) \end{cases}$$

$$\text{where } I_u^*(T_R) = \{i \in I \mid R^*(u,i) \geq T_R\}, \alpha_u = \max_{i \in I_u^*(T_R)} rank_X(i).$$

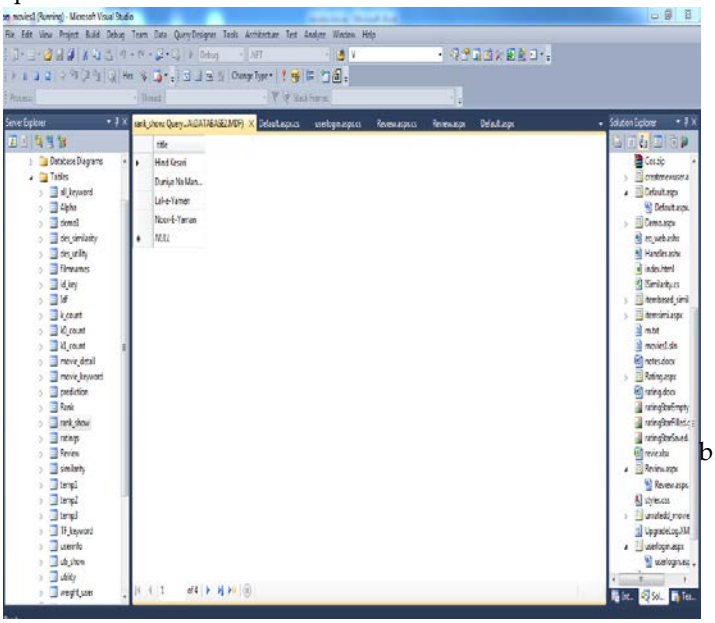
Where X=Itempop.

Calculate rank :table



Calculate alpha(u)

In rank_show table after implement rank formula add title of specific movie



End rank algo

C. Content based technique

In content based technique, recommender system suggests items to the user preferred in the past. The utility u(c,s) i.e. rating for user u of item s is estimated based on the utilities assigned by user c to items 'si' ∈ S (set of all items) similar to item s. Only the movies with high degree of similarity to user's preferences are would get recommended.

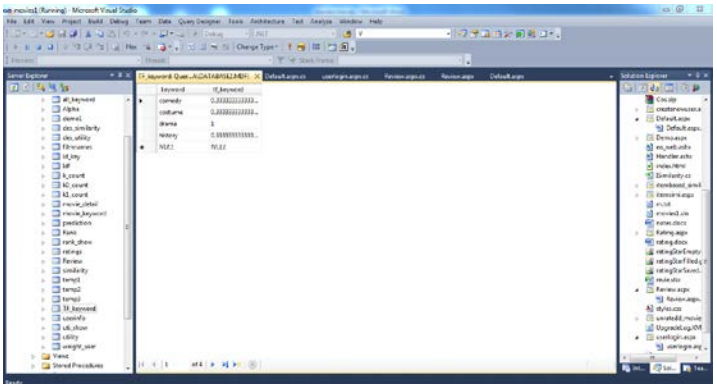
1. Item profile i.e. movie information table contains set of genres or keywords for characterizing item. User profile i.e. user information table contains taste and preferences of user. User preferences are obtained by previously rated items by that user.
2. To specify keyword weights, term frequency- inverse document frequency (TF-IDF) weighting measure can be used. N is the total number of items that can be recommended to users and keyword k_i appears in n_i of them. $f_{i,j}$ is number of times keyword k_i appears for item i_j or user u_j [11].

a. The term frequency of keyword k_i in item i_j or user u_j is defined as

$$TF_{i,j} = \frac{f_{i,j}}{\max_z f_{z,j}}$$

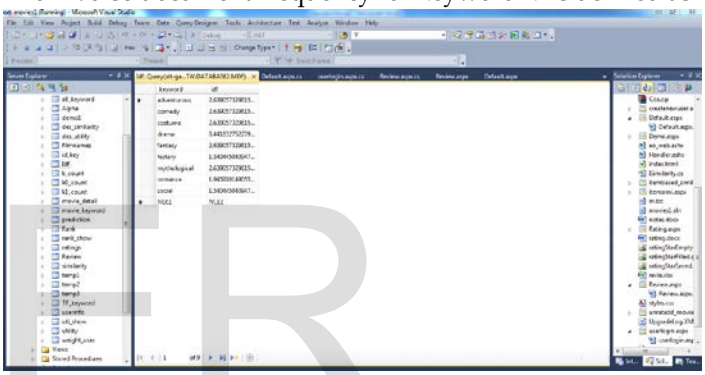
where maximum is computed over the frequencies $f_{z,j}$ of all keywords k_z that appear in the item i_j or user u_j.

abow formula implement and save value to TF_keyword table



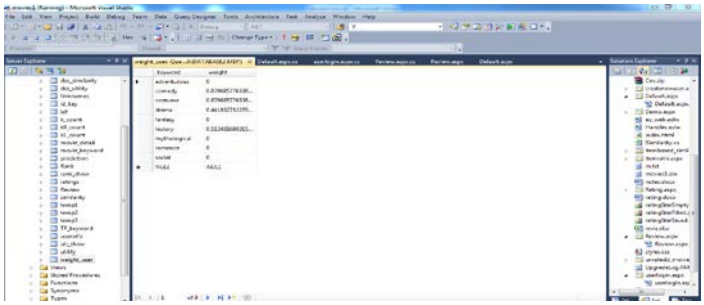
$$IDF_i = \log \frac{N}{n_i}$$

b. The inverse document frequency for keyword k_i is defined as



c. Thus the TF-IDF weight for keyword k_i in item i_j or user u_j is defined as

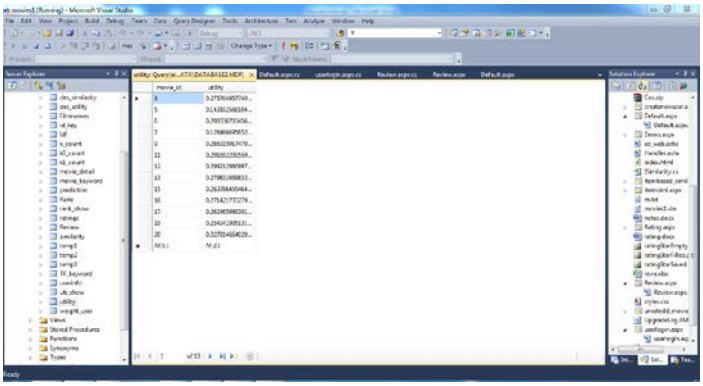
$$w_{i,j} = TF_{i,j} \times IDF_i$$



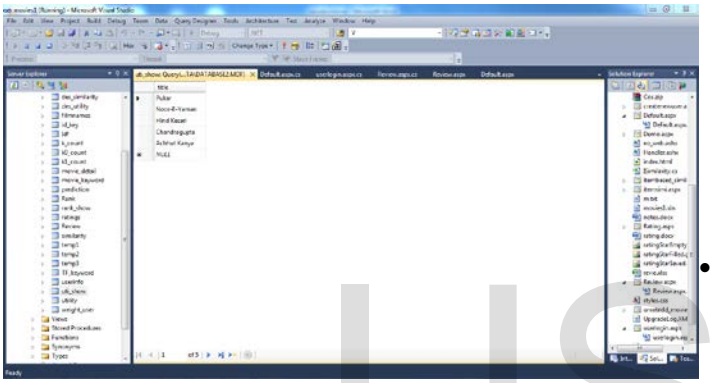
3. Utility for user c to item s i.e. u(c, s) is estimated using cosine similarity measure[11] as follows

$$u(c,s) = \cos(\vec{w}_c, \vec{w}_s) = \frac{\vec{w}_c \cdot \vec{w}_s}{\| \vec{w}_c \|_2 \times \| \vec{w}_s \|_2} = \frac{\sum_{i=1}^{K} w_{i,c} w_{i,s}}{\sqrt{\sum_{i=1}^{K} w_{i,c}^2} \sqrt{\sum_{i=1}^{K} w_{i,s}^2}}$$

Where K is total number of keywords.



Desending order utility,Arrange desending order utility, Get top 5 values and insert title of movie to the uti_show table



4. Items that have higher utilities with user's preferences will be recommended to user.

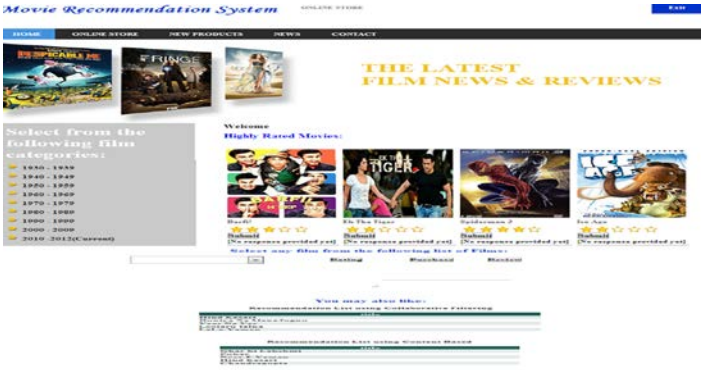
D. Item sequence generation technique

Each time when user visits shop, new list of recommended items is generated using user's past history. Markov Chain (MC) model can be used to predict the user's next preference based on the last sequential data. So transition matrix is estimated to get probability of buying an item based on last purchases of user.

1. Markov chain is stochastic process that undergoes transitions from one state to another where next state is only dependent on current state.
2. States in MC model represent the previous choices made by user. Thus set of states contains all possible sequences of user selections. Only sequences of at most k items can be considered to reduce state space size. Sequences are represented as $\langle x_1,\ldots,x_k\rangle$ which denote state in which last k selected items i.e. $x_1,\ldots,x_k$ by user are present.
3. The transition function shows the probability that a user with k recent selections $x_1, x_2, \ldots, x_k$ will select item x' next. $tr(\langle x_1, x_2,\ldots,x_k\rangle, \langle x_2, \ldots, x_k, x'\rangle$ is given by [9],

$$tr_{MC}(\langle x_1, x_2, x_3\rangle, \langle x_2, x_3, x_4\rangle) = \frac{count(\langle x_1, x_2, x_3, x_4\rangle)}{count(\langle x_1, x_2, x_3\rangle)}$$

Where $count(\langle x_1, x_2,\ldots,x_k\rangle)$ is number of items $x_1,\ldots,x_k$ sequence was observed in data set.



IV. *CONCLUSION*

Proposed recommendation technique which is based on content. Item popularity based parameterized ranking technique will ranks the items such that recommendation accuracy will be maintained and the diversity will be increased. Quality of recommendations will be improved using consumer/ manufacturer oriented ranking and item sequence generation techniques.

There are number of advantages of these systems due to which service providers may want to use this technology:

- Numbers of items sale will be increase: As compared to the number of items usually sold without any recommendation, recommender system is able to sell additional set of items. This is because of recommended items are as per the user's needs and wants.
- User satisfaction will be increase: User will find recommendations are interesting, useful and efficient. This system can also improve the experience of user.
- Sale of diverse items: This system is helpful to user to select the items that might be hard to find without using the recommendations.
- Better understanding of what the user wants: Recommender system uses user's preferences either collected explicitly from user profiles or predicted by system. This system uses this knowledge to improve the suggestions.

REFERENCES

[1] P. Resnick et al., "GroupLens: An Open Architecture for Collaborative Filtering of Netnews," *Proc. ACM 1994 Conf. Computer Supported Cooperative Work*, ACM Press, 1994, pp. 175-186

[2] J. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," *Proc. 14th Conf. Uncertainty in Artificial Intelligence*, Morgan Kaufmann, 1998, pp. 43-52.

[3] B.M. Sarwarm et al., "Analysis of Recommendation Algorithms for E-Commerce," *ACM Conf. Electronic Commerce*, ACM Press, 2000, pp.158-167.

[4] L. Ungar and D. Foster, "Clustering Methods for Collaborative Filtering," *Proc. Workshop on Recommendation Systems*, AAAI Press, 1998.

[5] M. Balabanovic and Y. Shoham, "Content-Based Collaborative Recommendation," *Comm. ACM*, Mar. 1997, pp. 66-72.

[6] Adomavicius, G., Y. Kwon. "Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques", IEEE Transactions on Knowledge and Data Engineering. 2011

[9]

[7] A. Ghose, and P. Ipeirotis, "Designing Novel Review Ranking Systems: Predicting Usefulness and Impact of Reviews," Proc. of the 9th Int'l Conf. on Electronic Commerce (ICEC), 2007.

[8] Neal Lathiax, Stephen Hailesx, Licia Caprax, Xavier Amatriainy, "Temporal Diversity in Recommender Systems", SIGIR'10, Geneva, Switzerland, July 19–23, 2010.