ISSN:0975-9646



# Recommender System Using Clustering Based On Collaborative Filtering Approach

**Jyoti Pandey**
*P. G. Student*
*Department of Computer Engineering*
*Smt. Kashibai Navale College of Engg*
*Pune, India*

Prof. **M. R. Patil**
*ME Computer, PhD. Pursuing*
*.Department of Computer Engineering*
*Smt. Kashibai Navale College of Engg*
*Pune, India*

***Abstract-*** **Recommendation process plays an important role in many applications as W.W.W. Recommender systems uses the users, items, and ratings information to predict how other users will like a particular item. An important response to the information overload problem is provided by the recommender system, as it presents users more personalized and practical information services. In the recommender systems field Collaborative Filtering (CF) is one of the most successful technique. For users based on their neighbor's preferences Collaborative filtering creates better suggestions. Current CF suffers from poor accuracy, scalability, data sparsity and big-error prediction. As the number of users and items in recent years are growing rapidly poses some key challenges for recommender systems. In this paper, we borrow ideas of object typicality from cognitive psychology and propose a novel Object typicality-based collaborative filtering recommendation method (OTCO), It outperforms many CF recommendation methods on recommendation accuracy (in terms of MAE) in MovieLens data set, In our propose method main approach is to cluster the all items into several item groups by applying k-means clustering algorithm. To help users to search items more easily and to improve the accuracy and quality of the recommendation.**

***Keywords- Recommender System, Collaborative Filtering (CF), Object Typicality, user group and item group.***

## I. Introduction

Recommendation systems found their application in the field of e-commerce and internet where items suggest to a group of user on the basis of their requirement based on their area of interest. A recommendation system is an information filtering system that built a model from the characteristic of an item according to the rating or prediction, given by a user to an item. Recommendation system has an important component in social media sites (such as Amazon, IMDB, Movie Lens), social sites giants such as Amazon have been greatly gained from the capability of their recommenders in accurately delivering the correct item to the correct user [17]. Collaborative filtering (CF) is an important and popular technology for recommender system. CF methods are classified into user-based CF and item-based CF. The basic idea of user-based CF approach is to find out a set of users who have similar favor patterns or interest to a given user and the basic idea of item-based CF approach is to find out a set of items having highest correlation with the given item. In reality, people may like to group items into categories, and for each category there is a corresponding group of people who like items in the category [18]. Cognitive psychologists find that objects (items) have different typicality degrees in categories in real life [19], [20], [21]. But these collaborative filtering methods have facing some problems like-

- Data Sparsity.
- Recommendation accuracy.
- Big-error Prediction.
- Scalability.

An important feature of the Object typicality-based collaborative filtering recommendation (OTCO) is that it finds the "neighbors" of users by measuring users' similarity based on typicality degrees of user in user groups, which differentiates it from previous methods. Object typicality-based CF method has the advantages:

- It works well even with data sets,
- It can reduce the number of big-error predictions.
- It improves the accuracy of predictions

## II. Existing Work

### A. *Object Typicality*

Object membership is different from object typicality, according to the study of cognitive psychology. The object typicality is a measure of the goodness degrees of objects [19], and the object membership is a measure of degrees of objects belonging to a concept. Psychologists find that people generally are more interesting in typical objects than a typical one in concepts [17]. Object typicality always depends on salient properties shared by most of the objects of that concept, which generally include both necessary and unnecessary salient properties for defining the concept prototype [3], In the prototype view of concepts [7], a concept is represented by a prototype abstracted by the list of property that consists of the objects salient properties that are classified into this concept. An instance typicality can be determined by the number of its properties which it shares with the concept prototype. For example, the prototype of the concept "bird" will probably appear in the property "can-fly", because most birds can fly. So birds that can fly will be more typical than those that cannot. A prototype of a concept is abstracted to be a feature list, and is considered as the best example of the concept .Although the prototype view can explain many different aspects of how properties and concepts are represented in mind of human's, there are also some situations in which it fails to give a thorough explanation. For example, to represent the concept "animal" there is virtually no prototype. It cannot explain the co-occurring relations of an instance among its properties.

If an object is more similar to the prototype of the concept than only it is considered as more typical in a concept. Vanpaemel et al. [22] propose a model that extends the prototype view. It is considered that a concept is represented by some abstractions (prototypes) concluded from exemplars of the concept. An object is considered to be an instantiation of an abstraction that is most similar to it. Typicality of an object is determined by matching its properties with those of the abstraction that is most similar to it and shows that both the prototype model and exemplar model are special cases of the model they propose, and such a combined model is better than the prototype model and exemplar model.

B. *Recommender System*

Any recommendation system consists of two basic entities user and items which help in decision making. There has been many works on recommendation systems and most of these works focus on developing new methods of recommending items to users, e.g., works in [13] [05]. Currently, recommendation methods are mainly classified into-

1) *Content-based Recommendation Systems*: The inspiration of these kind recommendation methods comes from the fact that people had their subjective evaluations on some items in the past and will have the similar evaluations on other similar items in the future [7]. These kind recommendation methods predict the preferences of active users on items based on the preferences of other similar users or items.
2) *Collaborative Filtering Recommendation Systems*: These kinds of recommendation methods predict the preferences of active users on items based on the preferences of other similar users or items. For the reason that collaborative filtering methods do not require well-structured item descriptions, they are more often implemented than content-based methods [7] and many collaborative systems are developed in academia and industry.
3) *Hybrid Recommendation Systems*: Several recommendation systems use a hybrid approach by combining collaborative and content-based methods, which helps to avoid some limitations of content-based and collaborative systems. A naive hybrid approach is to implement collaborative and content based methods separately, and then combine their predictions by a combining function, such as a linear combination of ratings or a voting scheme or other metrics.

## III. PROPOSED SYSTEM

A. *Problem Statement*

The main problem found in our existing system is a user with the recommendation on an item based on the other items with high correlations (i.e., "neighbours" of the item). It is significant to find users' (or items') neighbour in the similar set, but it is difficult to find correlation between users and items.

B. *Model*

In this paper, we borrow the idea of object typicality from cognitive psychology and propose a typicality-based CF recommendation approach named OTCO. The basic mechanism of object typicality-based CF recommendation is as follows: First, we cluster all items into multiple item groups. Second, we form a user group corresponding to each item group (i.e., a set of users who like items of a particular item group), in each of the user groups with all users having different typicality degrees. Third, we build a matrix known as user-typicality and measure similarities of users' based on users' typicality degrees in all user groups so as to select a set of "neighbours" of each user. Then, we predict the unknown rating of a particular user on an item based on the ratings of the "neighbours" of that user on the item. We propose a technique for an error-correction to suggest similar terms for the query keywords and return answers of the similar terms. To help users formulate high-quality queries, as user's type in keywords, we suggest keywords that are topically (popularly) relevant to the query keywords and we propose a query expansion-based technique to recommend users relevant keywords.

*Advantages of Proposed System:*

1. It improves the predictions accuracy when compared with previous recommendation methods.
2. It reduces the number of big-error predictions.
3. It works well with sparse training data sets also.
4. Users will find relevant patents more easily and improve user search experience.
5. Users can modify their keywords and interactively issue queries.

*Explanation of figure 1-* The ratings are collected from the super user and are represented in the form of user-item matrix. The ratings of the super user are compared with other users in the rating database and their similarity are computed using Pearson's correlation coefficient. Using the similarity values we cluster the users based on k-means clustering approach and find the top k-neighbors for producing recommendations. The recommendations from the top k neighbors are the products that the super user has not accessed yet that are given high ratings by their top neighbors. Proposed architecture follows using some parameters.

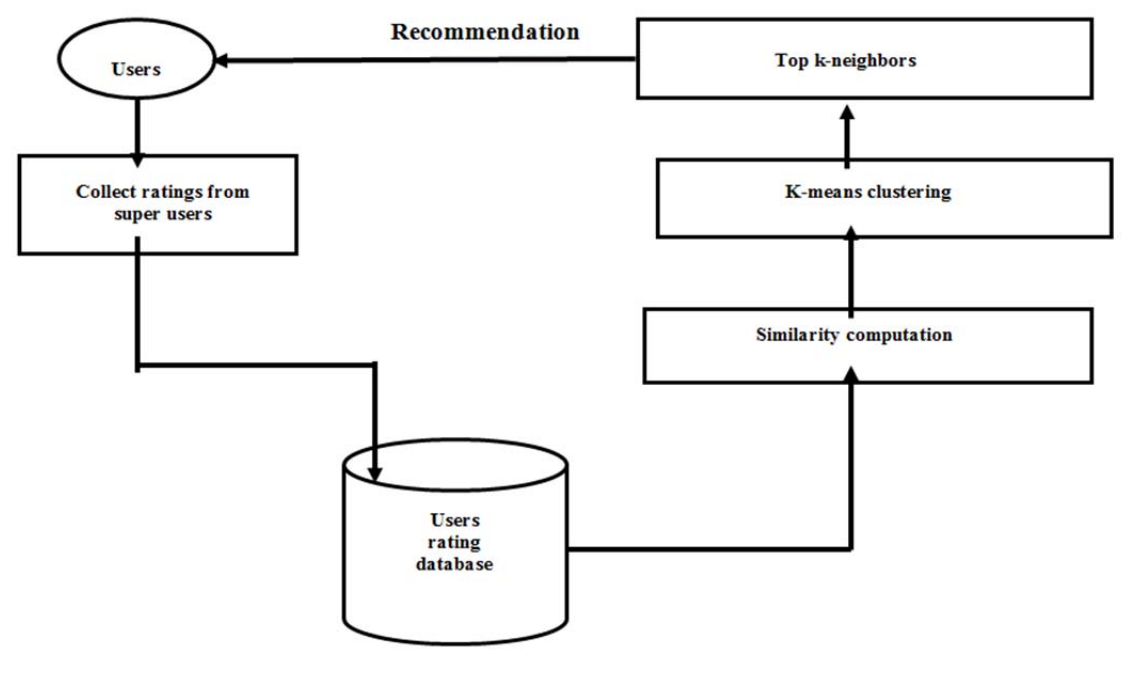


Fig 1 Architecture of Recommendation System using Collaborative Filtering

1) *Similarity*: The similarity of items and users can be measured by Pearson Correlation Coefficient (PCC). PCC obtain better performance [7] [14].

$$S_{a,b} = \frac{\sum_u (r_{u,a} - \overline{r_a}) \times (r_{u,b} - \overline{r_b})}{\sqrt{\sum_u (r_{u,a} - \overline{r_a})^2} \sqrt{\sum_u (r_{u,b} - \overline{r_b})^2}}$$

where,
$\overline{r_a}$ and $\overline{r_b}$ = the average rating given to item 'a' and 'b' respectively.
$r_{u,a}$ = rating given by user u on item 'a'.
$r_{u,b}$ = rating given by user u on item 'b'.

2) *Prediction:* Item-based CF predicts an active user u likes the active item I [7] [14].

$$P_{u,a} = \overline{r_a} + \frac{\sum_{b \in I_u} S_{a,b} (r_{u,b} - \overline{r_b})}{\sum_{b \in I_u} S_{a,b}}$$

where,
$P_{u,a}$ = predicted rating on the item a by the user u
$\overline{r_a}$ and $\overline{r_b}$ = the average rating given to item a and b respectively
$r_{u,b}$ = rating given by user u on item y.
$S_{a,b}$ = similarity of items.

C. *Results and Discussion*

1) *Data set description:* To evaluate our recommendation method, we use MovieLens data set, as this data set has been widely used in previous papers such as in [1]. We extract keywords of movies from the Internet Movie Database (IDMB), and regard such keywords as the descriptions of movies. It contains over 60,000 ratings provided by more than 900 users on 1,500 movies. Since the dataset is collected from social media, the expansion of the rated items (i.e., movies) is very large, leading to a sparsity value (i.e., ratio of known and all potential ratings in the item user matrix). The Sparsity level of the data set is 0.9555. Each users has rated at least 15 movies, and the ratings follow the 1 (bad) to 5 (excellent) numerical scale.

2) *Metrics: To* measure data accuracy, we use the mean absolute error (MAE) metric and it is defined as the average absolute difference between actual ratings and predicted ratings. MAE is a measure of deviation of recommendations from real user-rated ratings, and it is most commonly used and very easy to interpret. It is computed by averaging the all sums of the absolute errors of the n corresponding ratings prediction pairs, and can be formally defined as follows:

$$MAE = \frac{\sum_{i=1}^{n} |x_i - y_i|}{n}$$

where, n is the number of rating-prediction pairs, $x_i$ is an actual user-specified rating on an item, and $y_i$ is the prediction for a user on an item given by the recommender system. A lower MAE value means that more accurately the recommendation method can predict users' ratings. Thus, smaller the MAE values the recommendation is better.

3) *Impact of number of user groups:* In OTCO, the number of user groups is the same as the number of item groups. To test the sensitivity of different number of user groups (i.e., n), we run experiments for various n from 5 to 30, and the best results (with the most suitable parameter) on MAE. Although the MAE values for some n (e.g., n = 25) are a little bigger than those for other n (e.g., n = 20). Thus, we regard recommendation quality under different n as stable by setting an appropriate $\gamma$ i.e. the threshold value.

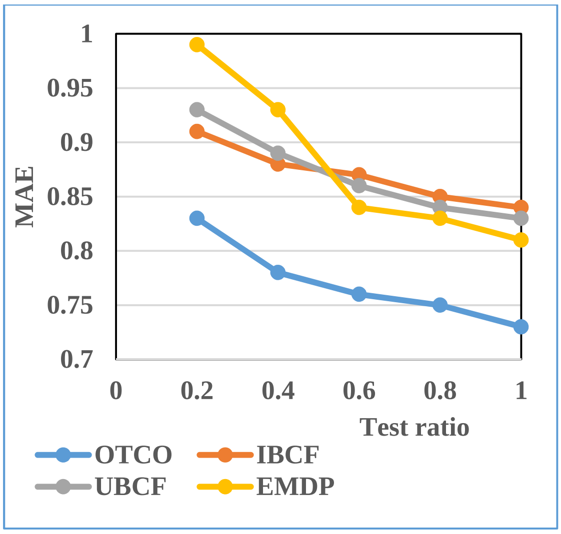


Fig 2 Comparison on MAE using various test ratio

TABLE 1: IMPROVEMENT OF OTCO FOR OTHER METHODS WITH SPARSE TRAINING DATA ON MAE.

| | Methods | $\chi$ = 0.2 | $\chi$ = 0.4 | $\chi$ = 0.6 | $\chi$ = 0.8 | $\chi$ = 1.0 |
|---|---|---|---|---|---|---|
| MAE | OTCO | 0.83 | 0.78 | 0.76 | 0.75 | 0.73 |
| | IBCF | 0.91 | 0.88 | 0.87 | 0.85 | 0.84 |
| | UBCF | 0.93 | 0.89 | 0.86 | 0.84 | 0.83 |
| | EMDP | 0.99 | 0.93 | 0.84 | 0.83 | 0.81 |

### D. *Comparison on recommendation quality*

We adopt three existing recommendation methods as baselines, which are user-based collaborative filtering (UBCF) with Pearson Correlation Coefficient, item-based collaborative filtering (IBCF) with Pearson Correlation Coefficient, and the EMDP method [25], to compare with the object typicality-based CF method. With the baseline methods with different train/test ratios on MAE. From figure 2, we can find that OTCO outperforms all other three baseline methods in all train/test ratios on MAE. Train/test ratio is denoted by$\chi$. For example, for train/test ratio $\chi$ =0:8, the MAE of OTCO is 0.75 while that MAE of IBCF is 0.85, MAE of UBCF is 0.84, and MAE of EMDP is 0.83. It shows clearly that object typicality-based CF method has higher recommendation accuracy than all compared methods.

The graph is showing the MAE and Test ratio estimation over the OTCO, IBCF, UBCF and EMDP. Here x axis shows the various test ratio and y axis shows the MAE.

## IV. CONCLUSION

In this paper, we investigate the recommendation system using clustering based CF approach and present a CF technique for improved recommendation system based on object typicality. The higher typicality degrees of users and items in the corresponding user and item groups, the recommendation scores are higher. The experiment shows the comparison of OTCO with IBCF, UBCF, and EMDP based on various test ratios and it shows that the object typicality-based method outperforms previous recommendation on recommendation quality. To deal with some of the challenges in CF, we propose a CF recommendation approach OTCO to cluster all the items into several item groups by using clustering algorithm.

## ACKNOWLEDGEMENT

The authors would like to thank the researchers as well as publishers for making their resources available and teachers for their guidance. We are thankful to the authorities of Savitribai Phule University, Pune and concerned members of IJCSIT, for their valuable guidelines and support.. We also thank the college authorities for providing the required infrastructure and support. Finally, we would like to extend a heartfelt gratitude to friends and family members.

## REFERENCES

[1] Badrul Sarwar, George Karypis, Joseph Konstan, and John Riedl, ”Item Based Collaborating Filtering Recommendation Algorithms”, ACM, 2001 Hong Kong.
[2] Gouthami, Golamari. Jose Mary and Pulluri Srinivas Rao, “Ranking Popular Items by Naive Bayes Algorithm”, International Journal of computer science and information technology (IJCSI) Vol 4. No 1, Feb 2012, pp. 147-163.
[3] Gediminas Adomavicius, YoungOk Kwon” Improving Aggregate Recommendation. Diversity Using Ranking-Based Techniques”, IEEE Transactions On Knowledge And Data Engineering, Vol. 24, No. 5, May 2012.
[4] Xiang Cui,Guisheng Yin,” Method of collaborative filtering based on uncertain user interests cluster”, Journal Of Computers ,Vol.8,No.1,January 2013, pp. 186-193.
[5] Jian Chen, Jin Huang, Huaqing Min, “Easy Recommendation Based on Probability Model”, Proceeding in Fourth International Conferences on Semantics, Knowledge and Grid, IEEE, 2008, pp 441-444.
[6] Toine Bogers, “Movie Recommendation uses Random Walk over Contextual Graph”, 2010.
[7] Min,Jie tang and Juanzi Li, ”Typicality-based Collaborative Filtering Recommendation”,Ieee Transaction On Knowledge And Data Engineering, Vol. 26, No. 3, March 2014, pp. 766-779.
[8] Ralf Krestel, peter Fankhauser, Wolfgang Nejdl, “Latent Dirichlet Allocation for tag Recommendation”, ACM 2009.
[9] Jun Hu, Bing Wang, Yu Liu, De-Yi Li,”Personalized Tag Recommendation Using S
[10] Social Influence“, Journal Of Computer Science And Technology 27(3): 527-540 May 2012.
[11] Anastation Noulas, Salvatore Seellaoto, Neal Lathia, Cecilia Mascolo, “Random Walk Around the City: New Venue Recommendation in Location-Based Social Networks”, 2012.
[12] Shumeet Baluja, Rohan Seth,D. Sivakumar, Yoshi Jing, Jay Yagnik, Shankar Kumar, Deepak Ravichandran, Mohamed Aly, ”Video Suggestion and Discovery for YouTube: Taking Random walks Through the View Graph”, ACM , Beijing, China, April 2008.
[13] Wenpu Xing and Ali Ghorbani, ”Weighted PageRank Algorithm”, Proceedings of the Second Annual Conference on Communication Networks and Services Research (CNSR’04), 2004 IEEE
[14] N.S. Nati and T. Jaakkola, “Weighted Low-Rank Approximations, “Proc. 20th Int’l Conf. Machine Learning, pp. 720-727, 2003.
[15] B. Li, Q. Yang, and X. Xue, “Can Movies and Books Collaborate?: Cross-Domain Collaborative Filtering for Sparsity Reduction,” Proc. 21st Int’l Joint Conf. Artificial Intelligence (IJCAI), pp. 2052-2057, 2009.
[16] Y. Koren, “Factorization Meets the Neighborhood: A Multifaceted Collaborative Filtering Model,” Proc. 14th ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD ’08), pp. 426-434, 2008.
[17] Yi Cai, Ho-fung Leungy, Qing Li, Jie Tang and Juanzi Li Recommendation Based On Object Typicality Project No. City U 117608, October 2010.

[18] F. Ricci, L. Rokach, B. Shapira, P. B. Kantor, “Recommender Systems Hand-book”, New York, NY, USA: Springer-Verlag, Oct. 2011.
[19] K.M. Galotti, Cognitive Psychology In and Out of the Laboratory, third ed. Wadsworth, 2004.
[20] G.L. Murphy, “The Big Book of Concepts,” MIT Press, 2002.
[21] L.W. Barsalou, Cognitive Psychology: An Overview for Cognitive Scientists. Lawrence Erlbaum Assoc., 1992.
[22] S. Schiffer and S. Steele, Cognition and Representation. Westview Press, 1988.
[23] W. Vanpaemel, G. Storms, and B. Ons, “A Varying Abstraction Model for Categorization,” Proc. Cognitive Science Conf. (CogSci’05), pp. 2277-2282, 20
[24] H. Ma, I. King, and M.R. Lyu, “Effective Missing Data Prediction for Collaborative Filtering,” Proc. 30th Ann. Int’l ACM SIGIR Conf. Research and Development in Information Retrieval (SIGIR ’07), pp. 39-46, 2007.
[25] L. Backstrom and J. Leskovec, “Supervised Random Walks: Predicting and Recommending Links in Social Networks,” Proc. Fourth ACM Int’l Conf. Web Search and Data Mining (WSDM ’11), pp. 635-644, 2011.