Diversity in Recommendation System: A Cluster Based Approach*

Naina Yadav, Rajesh Kumar Mundotiya Anil Kumar Singh, and Sukomal Pal

Indian Institute of Technology (BHU), Varanasi, India
{nainayadav.rs.cse18, rajeshkm.rs.cse16, aksingh.cse, spal.cse@}@iitbhu.ac.in

Abstract. The recommendation system is used to process a large amount of data to recommend new item to users, which are achieved using the many developed algorithms. Hence, it is a challenging task for lots of online applications to establish an efficient algorithm for a recommendation system that follows a good trade-off between accuracy and diversity. Diversity in recommendation systems is used to avoid the overfitting problem as well as excellent skill, which provides a recommendation based on increasing the quality of user experiences. In this paper, we proposed a methodology of recommendation to the user with diversity. The impact of diversity on the system leads to user experience for new items. The aim of this paper is to provide a brief overview of diversification with state of the art. A further similarity measure based on heuristic similarity measure "proximity impact popularity" is used to provide a new model with the better-personalized recommendation. The proposed approach gives profitability to many applications for better user experience and diverse item recommendations.

Keywords: Recommender System, Proximity Impact Popularity, Diversity, Accuracy

1 Introduction

Recommender System is the software tool that is designed for analyzing the user's past experiences and gives a list of suggestions form a large amount of information. The better opinion or recommendation leads an efficient system that is developed for better user experience[1]. Many recommendation algorithms are developed to learn the user's past behavior after that recommendation are generated as per their preference history[2]. The recommendation system is a technique that is used to provide suggestions to the user for the selection of the item. These suggestions are based on various decision-making processes, i.e., choice of items to buy, screening for the movie from a set of movies similarly to other online application[3]. Different types of recommendation algorithms work according to their respective domain and the knowledge used for users; at the end,

* Supported by Indian Institute of Technology (BHU), Varanasi.

different types of prediction algorithms are used for the generation of recommendation. There are many recommendation algorithms are defined as collaborative filtering, content-based recommendation system, and hybrid recommendation system. The collaborative filtering algorithm is based on information filtering or finding of co-related patterns using different techniques involving collaboration among diverse users and items[28][4]. Another method apart from collaborative filtering is content-based algorithms where the system tries to recommend items to users that are similar to the other user's past preferences[5]. The similarity between items and users is calculated using the different similarity metrics i.e., cosine similarity measure, Pearson correlation coefficient, etc. The similarity between different user and item is calculated using the content information provided by the user. Both collaborative filtering and content-base approach have their pros and cons. The collaborative filtering approach suffers from cold start problem, which means for the new item, and the new user recommendation generation is impossible, similarly in the content-based recommendation generation, specification of specific content description is confusing. On the other hand, the collaborative filtering approach suffers from sparsity, which means the existing number of items exceed the amount a person can explore. In content-based filtering techniques, sometimes difficulty in distinguishing between personal information of user[7]. In the past, evaluation in recommendation system depends upon the accuracy, which means how much relevant items are recommending to the user[6]. But nowadays, too many other evaluation measures are obtained for a sound recommendation system, diversity, serendipity and novelty is defined for a better recommendation system. In our proposed algorithm, we described diversity as a performance measure for our recommendation model, which gives users a diverse recommendation.

2 Related Work

Most of the recommendation system follows the same steps for recommendation generation. It starts with information analysis of items and users followed with user model generation, which stores information processed from information analysis after that, these models are using for recommendation generation[8]. When recommending items to users, it is essential to consider many performance metrics and not just the accuracy of a recommendation prediction. There are variety of metrics for recommendation evaluation[9].

– **Diversity** - Diversity is inclusion of different types of item set in recommendation for user which is different from their past preferences. Diversity is calculated using intra list similarity measure.

$$Diversity = \frac{1}{2} \sum_{i_j \epsilon u} \sum_{i_k \epsilon u} sim(i_j, i_k) \tag{1}$$

$sim(i_j, i_k)$ is the similarity measure between two item i_j and i_k commonly rated by the user u.

意外发现

– **Serendipity** - Serendipity is the measure of how surprising or relevant recommendations are generated for the user. Serendipity is calculated as the difference of the probability an item i that is recommended for a user u and the probability that item i is recommended for any other user.

$$Serendipity = \sum_{u} \frac{RS_u \cup E_u}{|E_u|} \tag{2}$$

where RS_u is the recommendation generated for user u and E_u is the items set of user u and $|N|$ defines the complete item set of user.

– **Novelty** - Novelty is fundamental qualities of recommendation system by which effectiveness and the new item was added to the recommendation list, which also leads to good accuracy.

$$Novelty = \frac{U_x}{U_i} \tag{3}$$

U_x is the item set which is unknown to user and U_i is the item set that is likes by the user U[10].

2.1 Diversity in Recommendation System: State of art

Diversity in recommendation systems is introduced to solve the problem of overfitting, which is, in the past few years, become a topic discussed by many researchers with informative publications. Diversity in recommendation has a twofold purpose: first is as mentioned above is overfitting problem, and second is user satisfaction with recommendation using diversification. K Bradley, B Smyth is the author who described diversity as a new type of diversity, preserving retrieval algorithm based on a similarity measure, which is proficient in delivering substantial improvements in recommendation diversity outwardly compromising recommendation similarity[11]. They propose diversity in three main strategies, which include retrieval of the k-item set from the complete item set using Bounded random selection. They also introduce a diversity measure as an intra list similarity measure. N Lathia, S Hailes et al. describe diversity with the time constraint. They proposed that recommendation grows with time over time as new users and items introduced to the system. Author calculate diversity using the collaborative filtering(CF) approach by giving the user a list of top-n recommendation. The formula for diversity calculation is -

$$Diversity(L_1, L_2, N) = \frac{\frac{L_2}{L_1}}{N} \tag{4}$$

where L_1, L_2 is for ranked list generated by CF algorithm and N is the total number of item in the set. Fleder, Daniel M. el al examines the effect of recommender systems on the diversity of sales. To measure sales diversity, they adopt the Gini coefficient. Gini coefficient is used to evaluate using a simulated

environment for user purchase tracking. The Gini coefficient for sales is defined as recommendation and diversity bias[13].

$$diversitybias(G) = 1 - 2\int_0^1 L(u)dx \tag{5}$$

Clarke, Charles LA et al. presents a methodology that is used for evaluation and comprehensively rewards novelty and diversity. They define diversity as a part of nDCG measure to avoid the ambiguity problems[14].

$$G(K) = \sum_{i=1}^{m} j(d_k, i)(1-\alpha)^{r_i,k-1} \tag{6}$$

on the another end Hu, Rong, et al. propose an approach based on a user study, that was conveyed to analyze an organization interface, which clubs recommendations into classes, with a standard list interface to perceived categorical diversity. They calculate diversity by a survey conducting between 20 participants[15]. Vargas, Sal, et al. proposed a methodology based on Binomial framework for genre diversity in recommender systems. They also propose an efficient greedy optimization technique to optimize Binomial diversity[16].

$$BinomDiv(R) = Converage(R) \times NonRed(R) \tag{7}$$

Hu, Liang, et al. stated that recommendation generation has diversified by using session contexts information with personalized user profiles. The author uses session-based wide-in-wide-out networks that are intended to efficiently learn session profiles across a large number of users and items[18]. Karakaya et al. proposed diversification using reranking algorithms that are utilized to aggregate diversity using the ranked list of recommendations[20]. Wilhelm, Mark, et al. proposed a diversified recommendation for the live YouTube user feed page. The author uses a statistical model of diversity based on determinantal point processes with set-wise optimization of recommendations[19]. Mller et al. use topic diversity in news recommendations using different diversity metrics in social science and democracy news[21]. Most of the researchers also focus on the other aspect of diversity in a recommendation, which includes serendipity and accuracy and their effects on diversity. Kotkov et al. proposed a serendipity oriented, the greedy reranking algorithm which improves serendipity of recommendations using feature diversification[22]. Apart from a tradeoff between diversity and other metrics of recommendation, Matt, Christian et al. described different types of diversity in the recommendation. The author proposed algorithmic recommendation diversity, perceived recommendation diversity, and sales diversity and identified different recommendation algorithms and user perception effects on sales recommendation[24]. Bag, Sujoy et al. proposed a model for online companies with personalized assistance to their consumers. The author suggested a prediction model for the profitability of online companies by recommending various items to users[25]. Recently Antikacioglu et al. give two different system-wide diversity metrics. The author in this proposed approach is considered as subgraph selection on a bipartite graph, which represents user and item[26].

3 Proposed Approach

The recommendation system is a beneficial decision support tool, and nowadays, they are an inevitable part of any user's daily life and web services. Basic building blocks of recommendation systems are user and item. Numerous algorithms described for recommendation system are based on the feedback of user-provided to the item in terms of review, tag, and rating or track the user behavior in terms of his likes and dislikes for items based on these information algorithms gives predictions. This article aims to provide a recommendation to the user with item diversification. Diversity in the recommendation system is for delivering the item to the user, which is different from user preferences. Suppose a user's preference for movie genres is action and science fiction. Still, for some variety, he wished to watch movies belongs to the family genre; algorithms like collaborative filtering give movies belonging to the genre action and science fiction frequently, which is not relevant in terms of recommendation generation. Recommendation system using a large amount of data, several machine learning algorithms are used for recommendation generation. In the proposed approach use K- means clustering algorithm to develop a personalized movie recommendation system with MovieLens dataset. K-means clustering algorithm is an unsupervised learning technique which is used for categorizing data and it depends on the hyperparameter k which denotes the number of clusters that is used for data classification.

3.1 Diversification Algorithm

Diversification in our approach is achieved using k-means algorithm with similarity measure algorithm Proximity-Impact-Popularity (PIP). PIP similarity measure is described by the Ahn, Hyung Jun. in 2008[17]. Steps in diversification algorithm as follows -

K-Means Clustering Algorithm is an unsupervised learning algorithm which is used for unlabelled data to classify them in clusters. Clusters in the k-means algorithm share the same set of properties. The algorithm works iteratively for each data point to cluster them using the features that are provided. K-means Algorithm work as follows:

- Define the number of the cluster as k.
- Randomly select k data point and calculate centroid without data shuffling.
- Keep iterating until centroid value is not going to change. It means data value assigned to the cluster is not changed.
- Compute the euclidean distance from one data point to other data points and assign data point to each cluster having minimum distance and then compute the centroid.

As per the algorithm we define cluster size as 20, it depends on the genre information, there are twenty distinct genre presents in the MovieLens dataset.

Proximity Impact Popularity (PIP) is a heuristic measure based on domain-specific data. PIP is more effective than any other similarity calculation because it overcomes the issue of the cold-start problem, which is an important issue in the recommendation system. PIP similarity calculation is based on three different factors i.e., Proximity, Impact, and Popularity.

$$SIM(u_i, u_j) = \sum_{k \epsilon C_{i,j}} PIP(r_{ip}, r_{jp}) \tag{8}$$

r_{ip}, r_{jp} is the rating value of item p rated by user i and j.

- Agreement - It is Boolean function Agreement(r_1, r_2) is based on value R_m

$$R_m = \frac{MaximumRating + MinimumRating}{2} \tag{9}$$

$$Agreement(r_i, r_j) = \begin{cases} False, & \text{if } (r_i > R_m \& r_j < R_m) \text{or} (r_i < R_m \& r_j > R_m) \\ True, & \text{otherwise} \end{cases} \tag{10}$$

- Proximity - A absolute distance between two rating is defined as -

$$D(r_i, r_j) = \begin{cases} |r_i - r_j|, & \text{if } Agreement(r_i, r_j) = True \\ 2 \times |r_i - r_j|, & \text{if } Agreement(r_i, r_j) = False \end{cases} \tag{11}$$

$$Proximity(r_i, r_j) = 2 \times (R_{max} - R_{min}) + 1 - D(r_i, r_j)^2 \tag{12}$$

- Impact - Impact(r_i, r_j) defined as -

$$Impact(r_i, r_j) = \begin{cases} (|r_i - r_m| + 1)(|r_j - r_m| + 1) & \text{if } Agreement(r_i, r_j) = True \\ \frac{1}{(|r_i - r_m| + 1)(|r_j - r_m| + 1)}, & \text{if } Agreement(r_i, r_j) = False \end{cases} \tag{13}$$

- Popularity - Let α is the average rating of item p given by all user.

$$Popularity(r_i, r_j) = \begin{cases} 1 + (\frac{r_i + r_j}{2} - \alpha)^2, & \text{if } (r_i > \alpha, r_j > \alpha) or (r_i < \alpha, r_j < \alpha) \\ Popularity(r_i, r_j) = 1 & \text{otherwise} \end{cases} \tag{14}$$

Recommendation Generation In the final step of our approach is recommendation generation, which is achieved by using these two algorithms discussed in section 3.1. From the k-means algorithm, we get the cluster information of the user, item, and rating. Clustering is based on the genre information of the movie. We are using MovieLens dataset, which contains 20 distinct genres, so we obtained 20 different clusters. We calculate similarities between a user from one cluster to another and then received an item set from another cluster. Then we calculate the predicted rating for that item set using the PIP algorithm and then recommend top - k item to the target user.

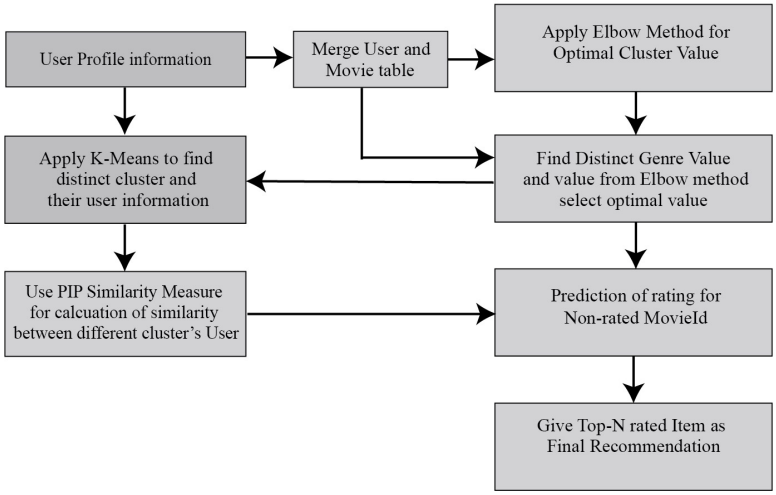

Fig. 1. Flow-Diagram For Praposed Appraoch

Table 1. Recommendation Generation Table

UserId	Recommendation Generation	Genre
12	Star Kid (1997)	Adventure—Children—Sci-Fi
12	They Made Me a Criminal (1939)	Crime—Drama
12	Someone Else's America (1995)	Comedy—Drama
12	Saint of Fort Washington, The (1993)	Drama
12	Prefontaine (1997)	Drama
31	Marlene Dietrich: Shadow and Light (1996)	Documentary
31	They Made Me a Criminal (1939)	Crime—Drama
31	Star Kid (1997)	Adventure—Children—Sci-Fi
31	Someone Else's America (1995)	Comedy—Drama
31	Saint of Fort Washington, The (1993)	Drama

4 Experiment and Results

In this section, we present results which are achieved using our proposed methodology as discussed in section 3.1. We present our top - 5 recommendation for the user id - 12 and 31 using Movielens 100k dataset. These users belong to different clusters, that clusters are achieved using the k-means algorithm over genre information. user- id 12 belongs to clusterId - 0, and user-id 31 belongs to clusterId - 12. For a number of the optimal clusters, we also applied the alleged

eblow method. It requires running the algorithm multiple times over a loop, with a growing number of cluster selection, and then contriving a clustering score as a function of the number of clusters. The optimal cluster value for a different pair of the genre is 7, 17, 22, and 27. Some of the cluster analysis of genre pair Drama & Action and Comedy & Action are shown in figure2, 3, respectively. The final recommendation for two different users with their respective genre are described in the table 1.

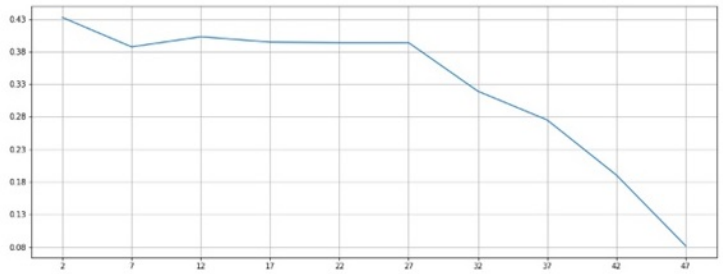

Fig. 2. Cluster Analysis of Genre- Drama & Action

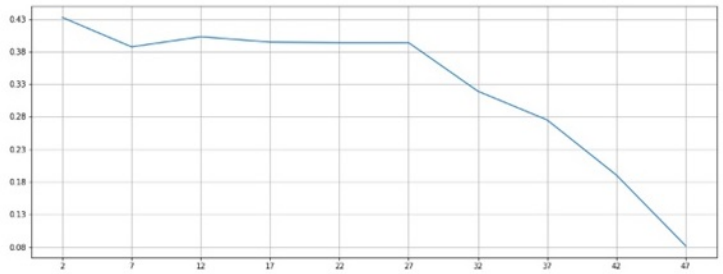

Fig. 3. Cluster Analysis of Genre- Drama & Comedy

5 Conclusion

The primary concern for any Recommendation system is to provide an accurate recommendation, but some times, this recommendation from the same domain

makes the user uninterested. So consider these limitations, we proposed an algorithm for diversification. In the future, we plan for directing serendipity and stability issues for the recommendation system with accuracy trade-off. Further, we want to examine the association of diversification methods with various deep learning approaches so that we can learn a suitable plan for diversification.

References

1. Salton, Gerard, and Michael J. McGill. Introduction to modern information retrieval. mcgraw-hill, 1983.
2. Kunaver, Matev, et al. "Increasing Top-20 diversity through recommendation post-processing." Semantic Web Evaluation Challenge. Springer, Cham, 2014.
3. Ricci, Francesco, Lior Rokach, and Bracha Shapira. "Introduction to recommender systems handbook." Recommender systems handbook. Springer, Boston, MA, 2011. 1-35.
4. Terveen, Loren, and Will Hill. "Beyond recommender systems: Helping people help each other." HCI in the New Millennium 1.2001 (2001): 487-509.
5. Su, Xiaoyuan, and Taghi M. Khoshgoftaar. "A survey of collaborative filtering techniques." Advances in artificial intelligence 2009 (2009).
6. Gunawardana, Asela, and Guy Shani. "A survey of accuracy evaluation metrics of recommendation tasks." Journal of Machine Learning Research 10.Dec (2009): 2935-2962.
7. ano, Erion, and Maurizio Morisio. "Hybrid recommender systems: A systematic literature review." Intelligent Data Analysis 21.6 (2017): 1487-1524.
8. Kurapati, Kaushal, et al. "A multi-agent TV recommender." Proceedings of the UM 2001 workshop Personalization in Future TV. 2001.
9. Gunawardana, Asela, and Guy Shani. "A survey of accuracy evaluation metrics of recommendation tasks." Journal of Machine Learning Research 10.Dec (2009): 2935-2962.
10. Zhang, Liang. "The Definition of Novelty in Recommendation System." Journal of Engineering Science Technology Review 6.3 (2013).
11. Bradley, Keith, and Barry Smyth. "Improving recommendation diversity." Proceedings of the Twelfth Irish Conference on Artificial Intelligence and Cognitive Science, Maynooth, Ireland. 2001.
12. Lathia, Neal, et al. "Temporal diversity in recommender systems." Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval. ACM, 2010.
13. Fleder, Daniel M., and Kartik Hosanagar. "Recommender systems and their impact on sales diversity." Proceedings of the 8th ACM conference on Electronic commerce. ACM, 2007.
14. Clarke, Charles LA, et al. "Novelty and diversity in information retrieval evaluation." Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2008.
15. Hu, Rong, and Pearl Pu. "Helping Users Perceive Recommendation Diversity." DiveRS@ RecSys. 2011.
16. Vargas, Sal, et al. "Coverage, redundancy and size-awareness in genre diversity for recommender systems." Proceedings of the 8th ACM Conference on Recommender systems. ACM, 2014.

17. Ahn, Hyung Jun. "A new similarity measure for collaborative filtering to alleviate the new user cold-starting problem." Information Sciences 178.1 (2008): 37-51.
18. Hu, Liang, et al. "Diversifying Personalized Recommendation with User-session Context." IJCAI. 2017.
19. Wilhelm, Mark, et al. "Practical diversified recommendations on youtube with determinantal point processes." Proceedings of the 27th ACM International Conference on Information and Knowledge Management. ACM, 2018.
20. Karakaya, Mahmut zge, and Tevfik Aytekin. "Effective methods for increasing aggregate diversity in recommender systems." knowledge and Information Systems 56.2 (2018): 355-372.
21. Mller, Judith, et al. "Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity." Information, Communication Society 21.7 (2018): 959-977.
22. Kotkov, Denis, Jari Veijalainen, and Shuaiqiang Wang. "How does serendipity affect diversity in recommender systems? A serendipity-oriented greedy algorithm." Computing (2018): 1-19.
23. Wu, Qiong, et al. "Recent Advances in Diversified Recommendation." arXiv preprint arXiv:1905.06589 (2019).
24. Matt, Christian, Thomas Hess, and Christian Wei. "A factual and perceptional framework for assessing diversity effects of online recommender systems." Internet research (2019).
25. Bag, Sujoy, Abhijeet Ghadge, and Manoj Kumar Tiwari. "An integrated recommender system for improved accuracy and aggregate diversity." Computers Industrial Engineering 130 (2019): 187-197.
26. Antikacioglu, Arda, Tanvi Bajpai, and R. Ravi. "A new system-wide diversity measure for recommendations with efficient algorithms." arXiv preprint arXiv:1812.03030 (2018).
27. Yuan, Bowen, et al. One-class Field-aware Factorization Machines for Recommender Systems with Implicit Feedbacks. Technical Report. National Taiwan University, 2019.
28. Tewari, Anand Shanker, Naina Yadav, and Asim Gopal Barman. "Efficient tag based personalised collaborative movie reccommendation system." 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I). IEEE, 2016.