**Journal of Industrial Information Technology and Application** *ISSN(Online):2586-0852*
Vol.7 No.1

# A Collaborative Filtering Recommendation Algorithm Using FP-Growth Algorithm and K-means Clustering

**Sang Suh[*], Monika Singh, Chirag Dave**
*Department of Computer Science.*
*Texas A&M University - Commerce, Commerce, TX*

**Abstract**. Recommender systems are widely used in online e-commerce applications to improve user engagement and increase revenue. Many recommendation systems employ collaborative filtering technology, which has been proven to be one of the most successful techniques in recommender systems in recent years. With the gradual increase of customers and products in electronic commerce systems, a key challenge for recommender systems is providing high-quality recommendations to users in "coldstart" situations. We consider three types of cold-start problems: 1) recommendation on existing items for new users; 2) recommendation on new items for existing users; 3) recommendation on new items for new users. To solve the problems of scalability and sparsity in collaborative filtering, this paper proposed a solution to the cold start problem by combining the association rules with the clustering. First, the items are clustered based on user ratings and price. Then, the user profile is enriched with the association rules on clustered data. The proposed approach utilizes item clustering collaborative filtering to produce the recommendations. The recommendation, combining "association rules, user clustering, and item clustering collaborative filtering" is more scalable and accurate than the traditional system. For association rules, we use FP-Growth algorithm instead of Apriori to mine frequent items because FP-Growth algorithm only needs two scans compared with Aprior's multiple scans, which is more efficient.

**Keywords;** recommender systems, collaborative filtering, cold start, association rule, user clustering, item clustering, scalability, sparsity, accuracy, mean absolute error

[*] Corresponding author: Sang.Suh@tamuc.edu

Received: Jan 14, 2023; Accepted: Jan 15, 2023; Published: Mar 31, 2023

## 1. Introduction

Recommendation systems (RS) play an important role in today's times. It caters to a variety of different Domains such as books, movies, tv shows, media, News, E-commerce, restaurants, hotels, etc. Travel Industry has gained maximum benefit from the recommendation system which caters to a wide variety of subcategories in the travel industry such as Restaurants, Hotels and other accommodations, places to visit, points of interest, and many others like this [7,8]. Hotel recommendation systems have gained huge popularity especially post covid times because of the rebounding travel industry.

Hotel recommendation systems recommend the hotel to the user based on the preferences of the user. Conventional RS methods are categorized into three groups, such as Collaborative filtering approaches recommend based on user-item historical interactions such as rating or feedback; content-based approaches extract the mutual information between the user and system to suggest the recommendation [15]. Most RS rely on the collaborative filtering approach, which is a feedback or review approach. The third RS is hybrid RS which integrates multiple ways to state the recommendation [1,10]. Most RSS rely on the rating history of users to the items, which supports learning the user preferences, item characteristics, and some additional correlation information between users to an item can be predicted using such information [4,9].

As shown in Figure 1, Cold start problem (new user, new item) is one of the major issues which hinder the performance of recommenders [5, 19]. In the case of a new user, the number of ratings will be very less. This implies that the user profiles (consist of ratings given to the items) will be very short. The new user will be given non-personalized recommendations until an adequate number of ratings are collected for the user. For a new item updated in the system, initially there will be no ratings. The possibility that this item will be recommended to the user is minimal. These problems should be addressed because the initial recommendations given to a new user plays an important role in deciding the user satisfaction and retention. Only if good quality recommendation is given, the users will come back to the site.

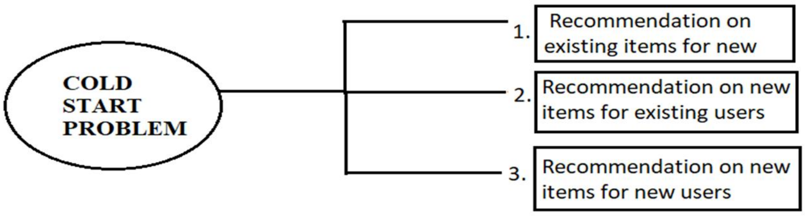

COLD
START
PROBLEM
1. Recommendation on existing items for new
2. Recommendation on new items for existing users
3. Recommendation on new items for new users

Figure 1. Types of cold-start problem

In this paper, we proposed a personalized recommendation approach that joins the association rules technology, user clustering technology and item clustering technology. The items are clustered based on user ratings and price. Then, the user profile is enriched with the association rules on clustered data. The proposed approach utilizes item clustering collaborative filtering to produce the recommendations. The recommendation joining association rules, user clustering and item clustering collaborative filtering is more scalable and more accurate than the traditional one.

## 2. Related Works

With the development of the internet and electronic commerce systems, there are amounts of information arriving that we can hardly deal with. Thus, personalized recommendation services exist to provide us with useful data employing some information filtering technologies. Information filtering has two main methods. One is content-based filtering and the other is collaborative filtering [1, 2]. Collaborative filtering (CF) has proved to be one of the most effective for its simplicity in both theory and implementation.

Collaborative filtering approach collects user ratings on items to predict user preferences. Based on the opinions of other users who share similar interests, the approach filters items and makes recommendations. Using the collaborative filtering approach, people can help each other to perform filtering [5, 6]. In e-commerce recommender systems nowadays, collaborative filtering that make recommendations according to the shopping experiences of similar users is the most important and widely used approach [3].

Many researchers have proposed various kinds of CF technologies to make quality recommendations. All of them make a recommendation based on the same data structure as a user-item matrix having users and items consisting of their rating scores. There are two methods in CF, user-based collaborative filtering and item-based collaborative filtering [16, 17]. User-based CF assumes that a good way to find a certain user's interesting item is to find other users who have a similar interest. So, at first, it tries to find the user's neighbors based on user similarities and then combines the neighbor users' rating scores, which have previously been expressed, by similarity-weighted averaging. And item-based CF fundamentally has the same scheme as user-based CF. It looks into a set of items; the target user has already rated and computes how similar they are to the target item under recommendation [6, 8, 13]. After that, it also combines his previous preferences based on these item similarities. The challenge of these two CFs are cold-start and sparsity.

There are many examples in the literature of machine learning techniques being utilized in recommendation systems. Although, hybrid filtering was proposed, as a solution to the limitations of CBF and CF, hybrid filtering still does not adequately address issues such as data sparsity, where the number of items in the database is much larger than the items a customer typically selects, and grey sheep, which refers to a typical user. Further, a system may still be affected when recommending items to new users (cold starts). To this end, in [10], researchers proposed a simultaneous co-clustering and learning framework to deal with new users and items. According to their data-mining methodology, a cluster analysis approach is integrated in the hybrid recommendation system, which results in better recommendations [11, 12]. Such a system was built in order to deal with sparsity and scalability in both CF and CBF approaches. Researchers also used clustering techniques to create user segmentations prior to classification [16].

To solve the problems of cold-start and sparsity in collaborative filtering, in this paper, we proposed a personalized recommendation approach that joins the association rules technology, user clustering technology and item clustering technology. Association rules are used to create and expand the user profile so that it will contain more ratings/domains of interest to solve new user problems. Apriori and FP-growth algorithms are mainly used in mining association rules. Apriori algorithm constructs candidate sets by repeatedly scanning the original data sets, and then uses the candidate sets to mine frequent item-sets. Due to the large number of scanning candidate sets, the algorithm is inefficient and cannot mine quantitative rules [19]. FP-Growth algorithm is to mine frequent items by building a frequent pattern tree FP-Tree. Each node in the tree view corresponds to an item in the set of frequent items. Because FP-Tree data structure compresses the original data, FP-Growth algorithm only needs two scans compared with

Apriori's multiple scans, which is more efficient [20]. The clustering technique is used to group items and make predictions for items to solve new item problems. Our research shows that combining two techniques such as FP-growth for association rules and K-means for clustering are more efficient and also give better recommendations.

## 3. Fp-growth based collaborative filtering

In this section, an outline of the proposed approach for solving the cold-start problem in recommender systems is drawn. The approach is to combine two existing approaches in a sequential manner. First, the FP-growth algorithm is applied to expand the user profile. With the help of this expanded user profile, K-means clustering technique is applied for recommendation, focusing on new item recommendation.

### *3.1 - K-Means clustering algorithm*

K-Means is one of the unsupervised learning algorithms and the simplest way to solve the problem of clustering. This procedure follows the simple and convenient way to define a set of specific data through a certain number of clusters k predetermined. A set of vectors will be $x_j$, j = 1, ... n, divided into groups i (G) for i=1, ..., c. The cost function is based on the Euclidean distance between vectors vector $x_k$ in group j and c cluster centers. This procedure follows the simple and convenient way to define a set of specific data through a certain number of clusters assumes k cluster that was previously set. The main idea is to define the centroid k, one for each cluster. The next step is to take every point included in a particular data set to connect it to the nearest centroid. The centroid k changes their location step by step until there are no more changes were made.

$$j = \sum_{i=1}^{c} j_i = \left(\sum_{k,x_k} \in G_i \; ||x_k - c_i||^2\right)$$
$$1, if \; \left| ||x_j - c_i|| \right|^2 \leq ||x_j - c_k||^2 , for\ each\ k \neq i$$
$$0, \qquad otherwise \tag{1}$$

### *3.2 - Association Rule*

Association rule is a data mining process to determine all the associative rules that meet the minimum requirements minimum support (minsup) and minimum confidence (minconf) on a database. Both of these conditions will be used for interesting association rules in accordance with the limits defined, namely minsup and minconf. Association rule mining is a procedure to look for relationships between items in a dataset. Starting with the search for frequent itemset, namely the combination that most often occurs in an itemset and must meet minsup. This phase will be conducted searches combinations of items that meet the minimum requirements of the value of the support in the database. The calculation of the value of support an item A out of items A and B can be obtained by the following formula.

$$Support\ (A, B) = P(A \cap B)$$
$$P(A \cap B) = \frac{Number\ of\ transactions\ containing\ item\ A}{Total\ transaction} \quad (2)$$

After all frequent items and large itemsets are obtained; you can find the minimum confidence (minconf) condition by using the following formula:

$$Confidence(A \rightarrow B) = P(A|B)$$
$$P(A|B) = \frac{Number\ of\ transactions\ containing\ A\ and\ B}{Number\ of\ transactions\ containing\ A} \quad (3)$$

### *3.3 - FP Growth*

The pattern of association of the functionality associated with the data transactions such as e-commerce activities for extracting data. The pattern of association will provide an overview of a number of attributes, or certain properties that often appear together in a given data set. Paradigm priori developed by Agrawal and Srikant, which is Apriori Heuristic: Every pattern with long pattern k that does not often appear (not frequent) in a data set, the pattern of length (k + 1) containing sub k pattern will not often appear also (not frequent). The basic idea of this a priori paradigm is to find the set of candidates to the length (k + 1) of a set of frequent patterns of length k, then match the number of occurrences of these patterns with the information contained in the database. Apriori algorithm will scan database repeatedly, especially if the amount of data is large enough. So, the FP-Growth algorithm which only requires twice scans the database to determine frequent itemset. The data structure used to find frequent itemset with FP-Growth algorithm is an extension of the use of a prefix Tree, commonly called is FP-Tree.

### *3.4 - Dataset*

We have used the Airbnb Hotel Recommendation dataset from Inside Airbnb website.

### *3.5 - System Design*

System design consists of several stages as shown in Figure 2. First, use the clustering process. Second, input the data to the FP-growth algorithm to enrich the new user's profiles. When scanning in FP-growth, if the number of data frequencies calculated is less than the frequent item set and the value of the relationship strength is measured by the lift ratio (LR)<1 then the item or combination of items will not be included in the next calculation. If it meets, the result of the rule can be determined. After that with K-means clustering technique, we can recommend hotels to new users. At last, the efficiency and accuracy of the result can be calculated with root mean square error (MSE) methodology.

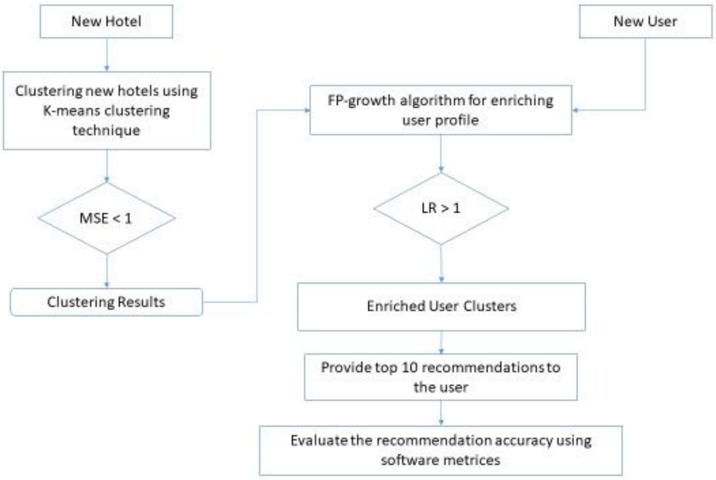

New Hotel
Clustering new hotels using K-means clustering technique
MSE < 1
Clustering Results
New User
FP-growth algorithm for enriching user profile
LR > 1
Enriched User Clusters
Provide top 10 recommendations to the user
Evaluate the recommendation accuracy using software metrices

Figure 2. Functional architecture of the proposed recommender System

## 4. Implementation

Our methodology for the implementation for the proposed work consists of five steps:

### *4.1 - Data collection and preprocessing*

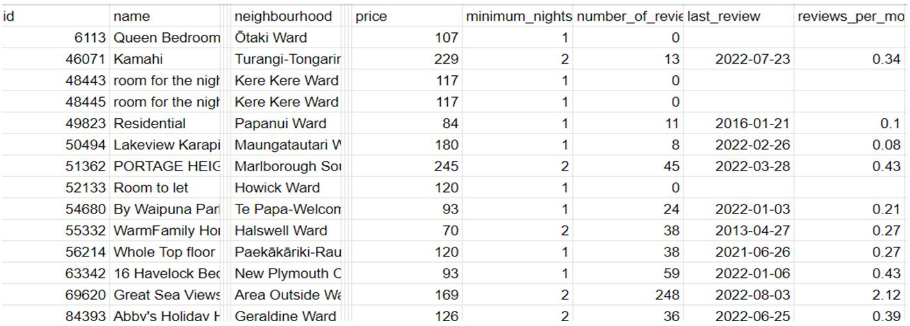
Dataset for the research is the Airbnb dataset collected from the Inside Airbnb site. The dataset size is 10MB (hotels: 2000, review: 100000). These dataset files are in .csv format. Each hotel has attributes such as id, name, host_id, host_name, neighbourhood_group, neighborhood, latitude, longitude, room_type, price, minimum_nights, number_of_reviews, last_review, reviews_per_month, calculated_host_listings_count, availability_365, number_of_reviews_ltm and license as shown in Table 1. Each user review has attributes such as listing_id, reviewer_id, reviewer_name, date and comment as shown in Table 2.

### *4.2 - Application of clustering technique to improve hotel recommendation*

In Figure 3 below, hotels in each cluster are represented with the same color. We applied the clustering technique as explained above when a new hotel is added to the recommender system. The k- means algorithm is used to assign the new hotel to a cluster. Since the hotel attributes, price and no of reviews are numeric, we use Euclidean distance to find the similarity between two hotels. Based on this similarity calculation the new hotel is assigned to a cluster to which it is most similar. The prediction for the new hotel is the average of the user ratings of the hotels in the cluster to which the new hotel is assigned.

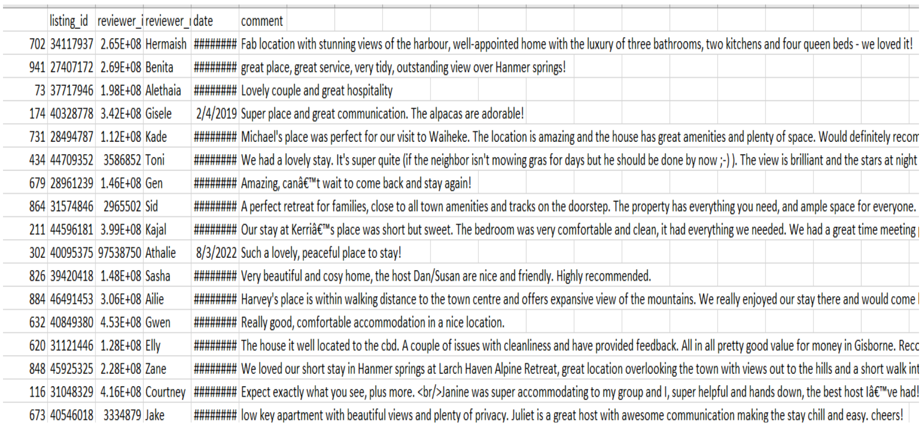
### *4.3 - Creation of taxonomy-based user profiles*

Now user profiles have to be created based on the information obtained from the users. The information is about the hotel in which the user is interested in. Taxonomy tree for hotel is created manually. Using this taxonomy driven user profile (P) we can construct a transactional dataset and mine the frequent patterns. Here the frequent patterns imply the clusters which appear together.



| id | name | neighbourhood | price | minimum_nights | number_of_revie | last_review | reviews_per_mo |
|---|---|---|---|---|---|---|---|
| 6113 | Queen Bedroom | Ōtaki Ward | 107 | 1 | 0 | | |
| 46071 | Kamahi | Turangi-Tongarir | 229 | 2 | 13 | 2022-07-23 | 0.34 |
| 48443 | room for the nigh | Kere Kere Ward | 117 | 1 | 0 | | |
| 48445 | room for the nigh | Kere Kere Ward | 117 | 1 | 0 | | |
| 49823 | Residential | Papanui Ward | 84 | 1 | 11 | 2016-01-21 | 0.1 |
| 50494 | Lakeview Karapi | Maungatautari W | 180 | 1 | 8 | 2022-02-26 | 0.08 |
| 51362 | PORTAGE HEIG | Marlborough Sou | 245 | 2 | 45 | 2022-03-28 | 0.43 |
| 52133 | Room to let | Howick Ward | 120 | 1 | 0 | | |
| 54680 | By Waipuna Par | Te Papa-Welcon | 93 | 1 | 24 | 2022-01-03 | 0.21 |
| 55332 | WarmFamily Ho | Halswell Ward | 70 | 2 | 38 | 2013-04-27 | 0.27 |
| 56214 | Whole Top floor | Paekākāriki-Rau | 120 | 1 | 38 | 2021-06-26 | 0.27 |
| 63342 | 16 Havelock Bec | New Plymouth C | 93 | 1 | 59 | 2022-01-06 | 0.43 |
| 69620 | Great Sea Views | Area Outside Wa | 169 | 2 | 248 | 2022-08-03 | 2.12 |
| 84393 | Abby's Holiday H | Geraldine Ward | 126 | 2 | 36 | 2022-06-25 | 0.39 |

Table 1. Airbnb hotel listing data



| | listing_id | reviewer_i | reviewer_i | date | comment |
|---|---|---|---|---|---|
| 702 | 34117937 | 2.65E+08 | Hermaish | ######## | Fab location with stunning views of the harbour, well-appointed home with the luxury of three bathrooms, two kitchens and four queen beds - we loved it! |
| 941 | 27407172 | 2.69E+08 | Benita | ######## | great place, great service, very tidy, outstanding view over Hanmer springs! |
| 73 | 37717946 | 1.98E+08 | Alethaia | ######## | Lovely couple and great hospitality |
| 174 | 40328778 | 3.42E+08 | Gisele | 2/4/2019 | Super place and great communication. The alpacas are adorable! |
| 731 | 28494787 | 1.12E+08 | Kade | ######## | Michael's place was perfect for our visit to Waiheke. The location is amazing and the house has great amenities and plenty of space. Would definitely recom |
| 434 | 44709352 | 3586852 | Toni | ######## | We had a lovely stay. It's super quite (if the neighbor isn't mowing gras for days but he should be done by now ;-) ). The view is brilliant and the stars at night |
| 679 | 28961239 | 1.46E+08 | Gen | ######## | Amazing, canâ€™t wait to come back and stay again! |
| 864 | 31574846 | 2965502 | Sid | ######## | A perfect retreat for families, close to all town amenities and tracks on the doorstep. The property has everything you need, and ample space for everyone. |
| 211 | 44596181 | 3.99E+08 | Kajal | ######## | Our stay at Kerriâ€™s place was short but sweet. The bedroom was very comfortable and clean, it had everything we needed. We had a great time meeting |
| 302 | 40095375 | 97538750 | Athalie | 8/3/2022 | Such a lovely, peaceful place to stay! |
| 826 | 39420418 | 1.48E+08 | Sasha | ######## | Very beautiful and cosy home, the host Dan/Susan are nice and friendly. Highly recommended. |
| 884 | 46491453 | 3.06E+08 | Ailie | ######## | Harvey's place is within walking distance to the town centre and offers expansive view of the mountains. We really enjoyed our stay there and would come |
| 632 | 40849380 | 4.53E+08 | Gwen | ######## | Really good, comfortable accommodation in a nice location. |
| 620 | 31121446 | 1.28E+08 | Elly | ######## | The house it well located to the cbd. A couple of issues with cleanliness and have provided feedback. All in all pretty good value for money in Gisborne. Recc |
| 848 | 45925325 | 2.28E+08 | Zane | ######## | We loved our short stay in Hanmer springs at Larch Haven Alpine Retreat, great location overlooking the town with views out to the hills and a short walk int |
| 116 | 31048329 | 4.16E+08 | Courtney | ######## | Expect exactly what you see, plus more. <br/>Janine was super accommodating to my group and I, super helpful and hands down, the best host Iâ€™ve had! |
| 673 | 40546018 | 3334879 | Jake | ######## | low key apartment with beautiful views and plenty of privacy. Juliet is a great host with awesome communication making the stay chill and easy. cheers! |

Table 2. User review data

### *4.4 - Application of association rule technique to enrich the user profiles*

From the user reviews dataset, we constructed a transactional dataset and mine the frequent patterns using the FP-growth algorithm. First, we added one more column cluster_id to this dataset that we mined in the previous step. After that, we ran the FP-growth algorithm that mined the frequent clusters where the user has booked the hotels. We listed all the frequent itemset in descending order to select only the top 10 itemsets. When a new user is added in the system, we enriched user highly liked clusters with mined frequent patterns.

### *4.5 - Project the Top-N recommendations to the user*

Recommendations generated by the recommender application are to be displayed to the user. The recommendations are displayed through the user interface of the application. In the implemented recommender system, the hotels which are not rated by him but come under his clusters of interests are given as recommendations. Also, randomly selected new hotels are also recommended. Thus, a total of 10 hotels (out of which 2 are new hotels) are recommended to the user.

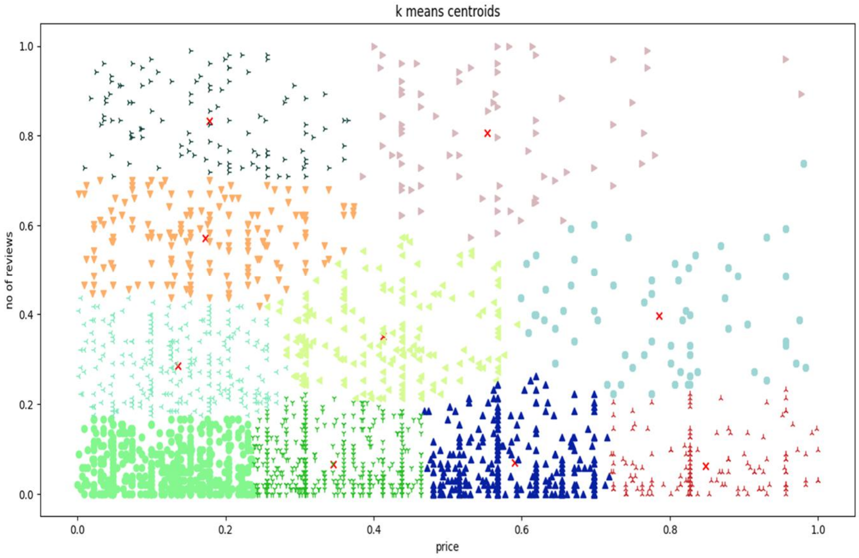


Figure 3. Hotel clusters based on price and reviews

## 5. Results

The quality of the recommendations can be checked using the software evaluation metrics such as precision, accuracy, recall and F1 measure. Here the evaluation of the recommendation accuracy is calculated using the accuracy metrics. Since only the Top-10 recommendations are displayed, the accuracy of the recommendations can be calculated as accuracy of the percentage of the number of items that are relevant to user u and which made it to top-10 items based on recommender system predictions. Relevancy is a term associated with human perception. A hotel which is relevant to one user may not be relevant to another user. In the proposed system, it recommended hotels with rating>2, to be considered as relevant. The precision is calculated for the recommendations generated by combination of association rules and clustering. The work is based on the suggestion that the use of a taxonomy driven profile will improve the recommendations since the system will be able to cover more topics in which the user is interested in. The enriched profile thus generated will be an added advantage in clustering the new items and produce quality recommendations for the users. As shown in Figure 4, the research shows that the proposed approach will be better than applying the association rules and clustering techniques separately.

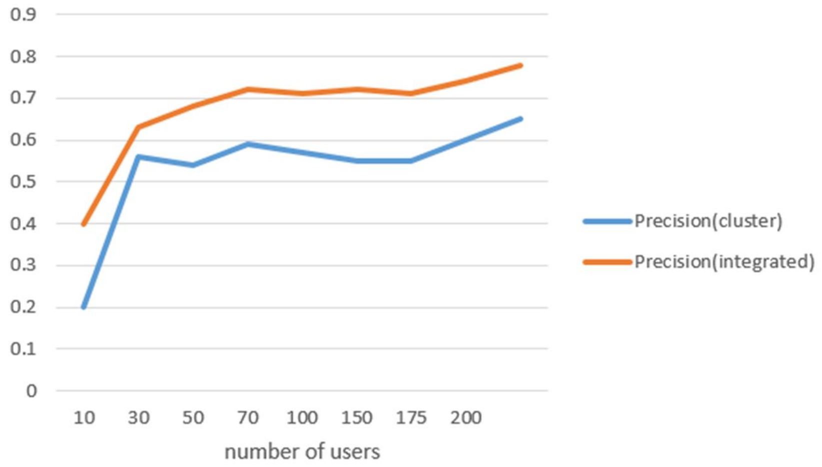


Figure 4. Accuracy improvement on using the combination technique over clustering technique

## 6. Conclusion & Future Works

In this work, the assumption was that out of the hotels given as recommendations only those which are having a prediction rating of more than 2 are relevant to the user. The accuracy values were obtained for 200 users. It contains the number of users who used the recommender service, the number of users registered in the system and the accuracy values for implemented techniques. The result clearly proves that the combination technique is better than applying clustering alone. An accuracy of 79% and an overall improvement of 28% was obtained on using the combination technique over clustering technique.

For this research we used only one site hotel listing and review data. In the future, multiple site data and multiple location data can be combined to get better accuracy. Further, user click activity data and web cookies can be tracked to mine important features. Additional advancements may be used to drastically enhance the versatility of recommender frameworks.

## References

[1] Buket Kaya, "A hotel recommendation system based on customer location: a link prediction approach",*Multimedia Tools and Applications, 2020.pages 3-7.*

[2] Kai Zhang, Keqiang Wang, Xiaoling Wang, Cheqing Jin, Aoying Zhou, "Hotel Recommendation based on User Preference Analysis", *ICDE Workshops, 2015.*

[3] Marco Rossetti, Fabio Stella, Longbing Cao &amp; Markus Zanker, "Analysing User Reviews in Tourism with Topic Models.", *Inf Technol Tourism, 2016.*

[4] Michael P. O'Mahony and Barry Smyth, "Learning to Recommend Helpful Hotel Reviews", RecSys, 2009.

[5] Marie Al-Ghossein, Talel Abdessalem, Anthony Barr´e., "Exploiting Contextual and External Data for Hotel Recommendation", *UMAP, 2018.*

[6] Ruihai Dong, Barry Smyth, "From More-Like-This to Better-Than-This: Hotel Recommendations from User Generated Reviews", *UMAP, 2016.*

[7] Md. Shafiul Alam Forhad, Mohammad Shamsul Arefin, A. S. M. Kayes, Khandakar Ahmed, Mohammad Jabed Morshed Chowdhury, and Indika Kumara, "An Effective Hotel Recommendation System through Processing Heterogeneous Data", *Electronics, 2021.*

[8] Marie Al-Ghossein, Talel Abdessalem & Anthony Barré, "Open data in the hotel industry: leveraging forthcoming events for hotel recommendation", *Information Technology & Tourism volume 20, 2018*, pages191–216.

[9] Mónica Méndez Díaz, and Clara Martín Duque, "Open Innovation through Customer Satisfaction: A Logit Model to Explain Customer Recommendations in the Hotel Sector", *J. Open Innov. Technol. Mark. Complex, 2021.*

[10] Mohanad Al-Ghobari, Amgad Muneer, Suliman Mohamed Fati, "Location-Aware Personalized Traveler Recommender System (LAPTA) Using Collaborative Filtering KNN", *Computers, Materials & Continua, 2021.*

[11] Saman Forouzandeh, Kamal Berahmand, Elahe Nasiri and Mehrdad Rostami, " A Hotel Recommender System for Tourists Using the Artificial Bee Colony Algorithm and Fuzzy TOPSIS Model: A Case Study of TripAdvisor",*International Journal of Information Technology & Decision Making Vol. 20, 2021*,pp. 399-429.

[12] Aziz Khan, Shougi S. Abosuliman, Saleem Abdullah, and Muhammad Ayaz, "A Decision Support Model for Hotel Recommendation Based on the Online Consumer Reviews Using Logarithmic Spherical Hesitant Fuzzy Information", *Entropy, 2021.*

[13] Yuanyuan Zhuang and Jaekyeong Kim, "A BERT-Based Multi-Criteria Recommender System for Hotel Promotion Management", *Sustainability 2021.*

[14] Marie Al-Ghossein, Talel Abdessalem, Anthony Barré, "Exploiting Contextual and External Data for Hotel Recommendation", *UMAP, 2018* Pages 323–328.

[15] Zhe Wang, Yangbo Gao, Huan Chen, Peng Yan, "Session-based item recommendation with pairwise features", *RecSys Challenge, 2019*, Pages 1–5.

[16] Malte Ludewig, Dietmar Jannach, "Learning to rank hotels for search and recommendation from session-based interaction logs and meta data", *RecSys Challenge & Proceedings of the Workshop on ACM Recommender Systems Challenge, 2019.*

[17] Marco Rossetti, Fabio Stella & Markus Zank, "Analyzing user reviews in tourism with topic models", *Information Technology & Tourism volume 16, pages 5–21, 2016.*

[18] Saumya Bhadani, "Biases in Recommendation System*", In Fifthteenth ACM Conference on Recommender Systems, 2021.*

[19] Xuebo Sun, Feida Shi, "Research and optimization of Apriori algorithm based on Hadoop", *Computer engineering and design volume 39, pages 126-133, 2018.*

[20] Wang B, Dan C, Shi B, "Comprehensive Associations Rules Mining of Health Examination Data with an Extended FP-Growth Method", *Mobile Networks & Applications volume 22, pages 1-8, 2017.*