



e-ISSN: 2147-8228

www.dergipark.org.tr/ijamec

Volume 11
Issue 01

March, 2023

*Research Article***A Manhattan distance based hybrid recommendation system****Begüm Uyanık^{a,*} , Günce Keziban Orman^b** ^a*Galatasaray University, Graduate School of Science and Engineering, 34349 İstanbul/Turkey*^b*Galatasaray University, Faculty of Engineering and Technology, 34349 İstanbul/Turkey*

ARTICLE INFO

Article history:

Received 10 January 2023

Accepted 16 March 2023

Keywords:

Recommender system

Hybrid approach

RFM Analysis

Collaborative filtering

Content based filtering

ABSTRACT

Many online service providers use a recommendation system to assist their customers' decision-making by generating recommendations. Accordingly, this paper proposes a new recommendation system for tourism customers to make online reservations for hotels with the features they need, saving customers time and increasing the impact of personalized hotel recommendations. This new system combined collaborative and content-based filtering approaches and created a new hybrid recommendation system. Two datasets containing customer information and hotel features were analyzed by Recency, Frequency, Monetary (RFM) method in order to identify customers according to their purchasing nature. The main idea of the recommendation system is to establish correlations between users and products and make the decision to choose the most suitable product or information for a particular user. As a result of the exponential growth of online data, this vast amount of information for use in the tourism industry can be leveraged by decision-makers to make purchasing decisions[20]. Filtering, prioritizing, and beneficially presenting relevant information reduces this overload. There are following three main ways that recommendation systems can generate a recommendation list for a user; content-based, collaborative-based, and hybrid approaches[1]. This paper describes each category and its techniques in detail. RFM Analysis is used to identify customer segments by measuring customers' purchasing habits. It is the process of labeling customers by determining the Recency, Frequency, and Monetary values of their purchases and ranking them on a scoring model. Scoring is based on how recently they bought (Recency), how often they bought (Frequency), and purchase size (Monetary). Experimental results show that the accuracy of behavior analysis using Manhattan distance-based hybrid filtering is greatly improved compared to collaborative and content-based algorithms.

This is an open access article under the CC BY-SA 4.0 license.
(<https://creativecommons.org/licenses/by-sa/4.0/>)

1. Introduction

In the last 15 years, the Internet has evolved from a group work tool to support the work of scientists at CERN, but rather a global knowledge space with more than a billion users. With the spread of the internet, many opportunities have emerged, such as sharing information and ideas with other users [3].

Nevertheless, this time, users encountered a new problem with the internet. The amount of data and units has increased greatly, leading to data overload. Finding out what the user is actually looking for has become a big problem. With the need to filter and sort items and information, developers found recommendation systems as a solution.

With the rapid development of e-commerce, online tourism website has become a quite common way to book tourism services[17]. Nowadays, people widely use online reservation systems to plan their holidays. A large number of options makes it difficult for hotel customers to decide when and where to go. In addition, due to the wealth of information available in online reservation systems, customers may miss out on a more suitable option for them. In this sense, recommendation systems play a major role in customers' choices [22]. Recommendation systems are useful for service providers and users [5]. They decrease transaction costs for finding and choosing products in an online shopping environment [6]. Recommender systems also have some benefits for

* Corresponding author. E-mail address: uyanikbegum@gmail.com
DOI: 10.18100/ijamec.1232090

businesses. Firstly, Revenue Algorithm studies using recommendation systems to increase revenue by increasing the number of sales for companies with online customers such as Amazon sites. Secondly, Personalization - Data collected indirectly can be used to ensure that the services of the website are suitable for the user's preference. Thirdly, Discovery - giving recommendations to people like shopping, movies, songs, etc. increases the chances of revisiting a web page when they find it [4].

Recommender systems have the following different types of filtering to create an effective recommendation engine; content-based, collaborative-based, and hybrid [4].

In this study, real data from SeturTech, a technology company that develops systems that offer hotels and holidays to users, are used. These data are made with a hybrid approach to hotel recommendations for users. This hybrid recommendation system is developed by combining two recommendation systems based on content-based and collaborative filtering to increase the real-life performance of the hotel recommendation system. Unlike similar studies in the literature, a rich hotel attribute list was used in the content-based method in this study. In addition, in the collaborative filtering method, the interaction matrix is created with the preference amounts of the guests and resembles the user-product score matrix. There are different distance metrics to calculate the similarity of customers;

The Cosine distance & Cosine Similarity metric is mainly used to find similarities between two data points. As the cosine distance between the data points increases, the cosine similarity, or the amount of similarity decreases, and vice versa. Thus, points closer to each other are more similar than points that are far away from each other. Cosine similarity is given by $\cos \theta$, and cosine distance is $1 - \cos \theta$.

Manhattan Distance metric, also known as city block distance, or taxicab geometry is used if we need to calculate the distance between two data points in a grid-like path.

While determining customer profiles, the RFM method is used. Recency shows the time elapsed since the user's last consumption so far. The closer the consumption time, the higher the customer score. Time is measured in days.

Frequency indicates how many times the user visits in a given time period.

Monetary represents the amount of money a user has spent in a given period of time. One point is based on the transaction amount, the higher the transaction amount, the higher the score.

To present the details of the work done, the rest of this paper is structured as follows. In section II, the developed hotel recommendation system and the three types of recommendation system filtering methods are introduced.

In section III, RFM analysis of the hotel data is reviewed. In section IV, the results of the tests are given. Finally, we conclude in Section V.

2. Recommendation System Creation

The recommender system is the biggest subfield of data mining and there are two main approaches:

- 1) Non-customized
- 2) Customized.

Each approach has different techniques in machine learning. The non-customized recommender system gives the same item recommendation to all system users, rather than individual user data. Users' interests are not considered. In contrast, the personalized recommendation system considers the preference or interest of each user, thus recommending certain items to the user more effectively.

There are three basic approaches in the customized recommender system:

- 1) **Content-based filtering approach:** Characteristics are derived from information items.
- 2) **Collaborative filtering approach:** Characteristics derived from the user's environment.
- 3) **Hybrid filtering approach:** Each of all content-based filtering approaches and collaborative filtering approaches has its pros and cons. To cope with these disadvantages, the hybrid approach, a combination of both approaches, is used [1].

2.1. Content Based Filtering

Content-based filtering method, in other words, "cognitive filtering", works according to user profiles that were created in the beginning. The creation of such profiles is provided by the user creating an account for himself and logging into the system.

A profile contains information about a user and their tastes. During profile creation, it is necessary to provide initial information about the user, and for this reason, the recommendation systems prepare a survey.

The recommendation engine compares the items that users have rated positively with each other and determines their similarities [3]. The more a user interacts with the system, the stronger the user profile is created. In this content-based filter, only the recorded information of the user is sufficient instead to other similar users [1]. For instance, if a user liked a website containing the words "car", "engine" and "petrol", the pages proposed by CBF will be relevant to the automotive world[8].

Content-based recommendation systems are formed of three basic parts in terms of high-level architecture. Firstly, the system preprocesses units with a content analyzer. After that, a professional profile learner gets information concerning the users.

Eventually, the filtering component reveals several suitable suggestions [12]. The three sections mentioned are detailed below:

Content Analyzer: The task of the content analyzer is to prepare the available data for the next process and convert the information from its original state into more abstract and feasible. For instance, this converter has the ability to accept a web page as input and convert it to a keyword vector[9].

Profile Learner: Profile learner is a module specially designed for the user. It obtains preprocessed information from the content analyzer and generalizes them to structure the user's preferences[9].

Filtering Component: The last step of the content-based filtering method is the filtering component. It recommends the relevant items to the user based on the user profile[9].

Creating a model for the user's preference from user history is a classification learning style. It is divided into binary categories such as "user likes" and "user dislikes". For example, if a user buys a product, it is a sign that the user likes that product.

However, if the user buys the product and returns it, this is a sign that the user does not like the product. Generally, implicit methods can collect large amounts of data with some uncertainty as to whether the user likes the item [10].

In this study, the features of the hotels' customers visited before are taken as the basis to create the profiles of the users. For example, in Table I, user 1 has gone to hotels A and B. The capacities of the hotels that User 1 visits are 300 and 400 people, respectively, their distance to the sea is 200 and 100 meters, and breakfast is served in these two hotels. Therefore, when calculating User 1's profile, this content information is used and the average values of the relevant rows in Table I are taken. In this case, the capacity of the hotels where User 1 goes is 350 people on average, and their distance from the sea is 150 meters. It also seeks breakfast service at the hotels that User 1 visits. On the other hand, User 2 displayed a different profile by choosing hotels with less capacity and hotels farther from the sea. As a result, the profiles of these users will be as in Table II.

Table 1. User-Hotel Features Example

User	Hotel	Capacity	Distance to the Sea	Breakfast
1	A	300	200	1
1	B	400	100	1
2	C	100	600	0
2	D	150	1250	0
2	E	125	1500	1

Table 2. User Profiles with CB

User	Capacity	Distance to the Sea	Breakfast
1	350	150	1
2	125	1100	0.33

Using Microsoft Excel, the hotels selected by the customer and the features of the hotels are combined in a single table.

Hotel list data has been cleaned by preprocessing steps. Some hotels selected by customers are not in the hotel feature list, they were detected using excel and removed from the list.

These pre-processing steps consist of correcting missing data (Ex: Hotel names) and removing inconsistent data (Ex: Number of rooms). Currency code and foreign currency sales amounts columns, which will not be used in the setup of the recommendation system, are deleted and a comparison was made in Turkish Lira.

The names of districts and towns were removed, and the cities remained.

Branch name, code, and type; sales, entry, and exit dates, and how many people stayed overnight are removed. Then, the most important 8 features (pool, beach, breakfast, etc.) of the hotels are determined by subtracting the features with less than 600 selections of customers. Thus, by reducing the size of the data, it is ensured that the working time was shortened while calculating the similarity between the profiles and the hotels.

The average of the features in the hotels selected by the customers is calculated using the pivot table in excel. Customers with a binary value of 0 for the features of all selected hotels are removed from the study and the data set was simplified by removing them from the data set.

2.1.1. Content-Based Filtering Method Limitations

I. New user issue:

It is a problem caused by a lack of information. When a new user enters the system, the system does not have enough data about the user profile and preferences, so a suitable product profile cannot be created accordingly. As a result, the advice may not be good enough[1].

II. Excessive specialization:

It is caused by the recommendation system recommending similar types of products based on available historical data. When the user wants to try a new product, it may cause problems. For example, even if a user will like fiction movies if they are suggested; Since he only liked comedy movies in the past, the system will only suggest new comedy movies to him. Therefore, excessive specialization can narrow the scope of recommendation[1].

2.2. Collaborative Filtering

Collaborative filtering is a popular user recommendation approach. The collaborative filtering approach was first described and explained in 1997 by Paul Resnick and Hal Varian [7].

If two users have common items with similar ratings, it is assumed that they have similar tastes. Such users form a group or a so-called neighborhood.

A user receives recommendations for items that they have not previously rated but are already rated positively by users in their neighborhood [15].

In CF systems, a recommendation is made to a user collectively based on past ratings of all users. The Grundy system is the first recommendation system to propose the use of stereotypes as a mechanism to create user models based on limited information about each user.

The system creates the model of the individual user and relevant books are recommended for each user [1]. Video Recommender [13] and amazon.com [14] are some examples of collaborative filtering.

Collaborative filtering suggestions improve their accuracy as the amount of data on items increases [11].

Collaborative filtering method approaches can be divided into three subgroups:

I. User-based approach:

This approach was proposed by University of Minnesota Professor Jonathan L. Herlocker in the late 1990s[15]. In this approach, users take the main role. In this filtering, the subset of users is selected based on their similarity to active users. Customers who have the same taste construct a group. The user is given suggestions based on the items evaluated by other users in his group[3]. The weighted combination of their ratings is then used to estimate the rating for the user[1].

II. Item-based approach:

This approach was proposed by University of Minnesota researchers in 2001[16]. As a system grows, the number of users increases and so does the complexity of finding similar users. Therefore, a new approach to item-item collaborative filtering was proposed rather than finding similar users[1]. The system creates neighborhoods according to the tastes of the users. The system then generates suggestions with items found in a user's preferred neighborhood[3].

III. Model-based approach:

The system makes recommendations for users by estimating the parameters of statistical models for user ratings. In this approach, a pre-calculated model is a design based on available data. This model-based approach quickly responds to the user's preference when the user query appears. Thanks to this approach, the system can be visualized more accurately and can also reduce errors. The most commonly used methods are MF (Matrix factorization), and SVD (Singular value decomposition)[1].

In this study, while using the CF method, the number of times a customer went to each hotel is accepted as the score the customer gave to the hotel. In order to simplify the number of hotels, hotels with less than a total of 10 selection by the customers are removed from the dataset.

2.2.1 Limitation of collaborative filtering

I. Cold-start problem:

This problem refers to insufficient information to give the user a recommendation. Collaborative filtering is entirely dependent on the similar neighbor in the system, but these similar neighbors are not present in the system in the first stage known as the cold start issue and they are not known by the system. This reduces the performance of the recommendation system[7]. This problem can be avoided with the hybrid approach[1].

II. First-rater problem:

The system cannot recommend an item that has not been previously rated. As new items are entered into the system, many users did not refer to the items, so there are not enough ratings for these items. The problem can be solved with a hybrid approach.

III. Sparsity

The sparsity problem is an important issue. Collaborative recommendation systems often build users' neighborhoods using their profiles. Sparsity occurs when the user does not rank these items[3]. If a user has only rated a few items, it is quite difficult to determine their taste, and may be in the wrong neighborhood[1]. Sparsity is a problem of lack of information.

IV. Popularity Bias

The system cannot recommend products to someone with unique tastes. Sometimes the user has a unique taste compared to all other users on the system. This problem is known as the "popularity bias" issue. This problem can be solved with a hybrid approach[1].

2.3. Hybrid Approach

For better results, some recommendation systems use a combination of collaborative and content-based approaches to take advantage of each of them. By using the hybrid approach, the limitations of the content-based and collaborative approaches such as cold-started problems can be avoided. The combination of these two approaches can be achieved in different ways:

- Applying both methods separately and combining the results.
- Combining some content-based features with a collaborative approach.
- To include some collaborative features in the content-based approach.

Babodilla, et al have classified collaborative filtering and content-based filtering combined into four different groups as in figure 1 to make a hybrid method [9].

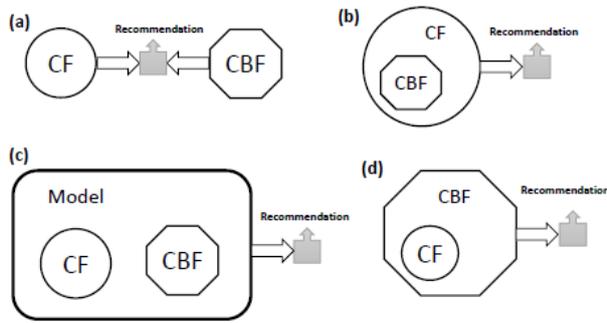


Figure 1. Different methods of combining CF and CBF [9]

Figure 1. a shows the hybrid method combining CF and CBF with a weighting method.

Figure 1. b shows the methods using CBF methods to extract the features and send the recommendation to CF.

Figure 1. c shows a combined model using CF and CBF to obtain the outputs of another classifier such as the probability model.

Figure 1. d shows a model for CBF using output from CF. For example, user ratings can help CBF better identify users [9].

Probability methods are used in hybrid filtering. Examples such as genetic algorithms, neural networks, and Bayesian networks can be given [8].

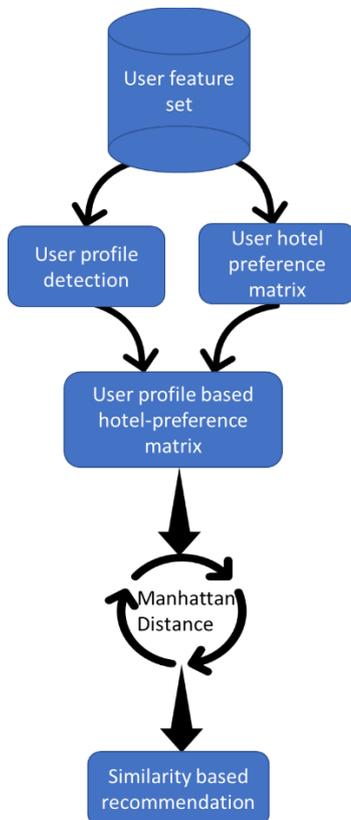


Figure 2. Hybrid approach to SeturTech Data Set

SeturTech dataset includes user properties and user-hotel preferences. First of all, the user profiles are analyzed with RFM in a single data with an engine. A user-hotel preference matrix is created from the same data. Then, a

user profile-based hotel preference matrix is created. While using the Manhattan distance method, a similarity-based recommendation system is created. For example, calculating user-2 that is most similar to user-1, the hotel that user-1 did not go to but user-2 went to is recommended to user-1.

2.3.1 Manhattan Distance

Manhattan distance is a metric in which the distance between two points is calculated as the sum of the absolute differences between their Cartesian coordinates. In a simple way of saying it is the total sum of the difference between the x-coordinates and y-coordinates.

ManhattanDistance [{a, b, c}, {x, y, z}]

$$\text{Abs} [a - x] + \text{Abs} [b - y] + \text{Abs} [c - z] \tag{1}$$

Using Manhattan Distance given in equation (1), it is intended to calculate the similarity between customers. The python program is used for this calculation. Each group of 8 groups is saved as a CSV file and the Manhattan distance is calculated. Since customers whose distance from Manhattan is closer to each other are considered similar to each other, the distance of each customer from Manhattan to the other is obtained. Since customers whose Manhattan distance is "0" will have gone to the same hotels, 0's have been eliminated from the code, and outputs are saved.

3. RFM Analysis of Customer Features

RFM analysis is based on the following assumptions:

1. Customers who have made recent purchases are more likely to purchase again than customers who have not purchased recently.
2. Customers who make more frequent purchases are more likely to repurchase the company's products.
3. Customers with a higher total purchase amount are more likely to purchase again.

For RFM analysis in the SeturTech dataset, each customer's Invoice number, Sales Date, and Sales Amount information is used in this study. A customer may have made more than one purchase. Since RFM is about customers, sales amounts are aggregated by customer ID.

The data is grouped with the pivot table feature in the Excel program. With the pivot table, operations such as sorting, summing, and averaging can be performed.

In the pivot table, customers are added to the "rows" column. Thus, each customer appears only once.

For the "Recency" calculation; customers need to know when the last time they visited the site is.

The "maximum" of the data in the "Sales Date" column is selected from the pivot table, and the last time they visited is shown in the column.

For "Frequency", due to the need for information about how often he visits the website, it is calculated how many invoices belong to the same customer to find out how many times the customer has visited the website.

“Monetary” requires information on how much the customer has made a total purchase. The total amount spent by the same customer is calculated with the pivot table.

To calculate RFM scores, from 1 to 5 points are distributed (eg frequency value of a frequent customer should be 5). Recency, frequency, and monetary matrices are created with formulas in Excel. For recency, for example, the customer who visited the website the longest time takes the value 1.

The time elapsed since the last visit is divided into groups according to percentiles. The last visitor was given after the 80% slice and took the value of 5. The customer who spends the least money should get 1 as its monetary value. According to these calculations, the RFM scores of all customers are determined.

4. Customer Segmentation

To achieve its business goals, a company can use customer segmentation to target its marketing efforts and resources to valuable and loyal customers[18].

To prepare a recommendation system using collaborative filtering, customers are divided into specific groups. Since applying this system to those in the lost customer class would not yield efficient results, it is needed to determine focus groups. For this purpose, primarily customer profiles are determined. These focus groups are selected according to the RFM scores assigned by RFM analysis, using frequency, monetary, and recency scores. Thus, 6 different customer groups emerged.

Customers in SeturTech data are divided into the following six different segmentations:

- 1) Loyal
- 2) Potential Loyal
- 3) Promising
- 4) Hesitant
- 5) Need Attention
- 6) Detractors

This customer segmentation is determined according to RFM scores given in Table 3.

Table 3. RFM Scores of Customer Profiles

Customer Profile	RFM SCORES									
Loyalists	444	434	443	344	442	244	424	441		
Potential Loyalists	332	333	342	343	334	412	413	414	431	
	421	422	423	424	433	441	432			
Promising	233	234	241	311	312	313	314	321	322	
	331	341	324	323						
Hesitant	124	133	134	142	143	144	214	224	234	
	232	243	242							
Need attention	122	123	131	132	141	212	213	221	222	
	223	231								
Detractors	111	112	113	114	121	131	211	311	411	

Loyal, potential loyal, and promising groups consist of active customers who visited the company last recently, shop from the company frequently, and spend much when shopping.

Hotel scores created by collaborative filtering are added to customers in each group.

5. Results

In this section, the outputs obtained after performing the above-mentioned analyzes are shared.

5.1. Data Analysis Results by Using Content-Based Filtering Method

Two different data sets are used in this study. One of them is the “Hotel Features” data set which includes which of the 30 features (rows of the first data set) such as bed & breakfast, spa-thermal hotel, casino hotel, all-inclusive, half board, pool, sandy beach of 2562 different hotels meet.

Table 4. A Brief Example from the Hotel Features Data Set

HotelID	Child-Baby Friendly	Pool	Pool-Summer	Beach	Spa-Thermal Otel
370	1	1	1	1	1
371	1	1	1	1	1
374	1	1	1	1	0
377	1	1	1	1	1
378	1	1	1	1	1
379	1	1	1	0	1
380	1	1	1	0	1
381	1	1	1	1	0
383	1	1	1	1	1
384	0	1	1	1	1
385	1	1	0	0	1
386	1	1	1	1	1
388	1	1	1	1	1
389	1	1	1	0	1

Another is a “Sales Data” data set containing 40599 sales data (rows of data set) of a company and 32 sales information (columns of the data set) such as Hotel ID, the code of the customer from whom the sale was made, the gender of the customer, the date of sale, the name of the branch where the sale was made.

Table 5. A Brief Example from the Sales Data Set

Customer ID	Hotel ID	City Name	Tkl Master ID	Number of Rooms	Adult Count
13	466	Antalya	267826	1	3
13	466	Antalya	267826	1	2
31	3235	Amasya	34737	1	2
31	418	Antalya	34737	1	3
31	2019	Konya	34737	1	3
141	9831	İzmir	34805	1	2
166	492	Girne	34820	1	2
166	492	Girne	34820	1	2
171	512	Antalya	34823	1	2
171	475	Antalya	34823	1	2
171	786	Antalya	34823	1	2
179	1291	Muğla	60499	1	2
179	1291	Muğla	60499	1	2
179	1291	Muğla	60499	1	2
191	373	Muğla	34834	1	2
217	396	Bursa	77242	1	2
254	1539	Ankara	34878	1	2
294	1023	İzmir	326046	1	2
296	1909	Girne	34893	1	1
300	454	Bolu	166854	1	2
318	786	Antalya	140287	1	2

Hotel features and sales data features that are not needed for this study are cleared from these two data sets.

The following features, which will not affect the results of the analyzes in the sales dataset, are not considered in this study:

- 1) Currency code and foreign currency sales amount columns were deleted, and a comparison was made in TL.
- 2) The names of districts and towns were removed, and the cities remained.
- 3) Branch name, code, and type are removed.
- 4) Sales, entry and exit dates, and how many people stay overnight are excluded because they will not be used in classification and RFM methods.

The 19 features, which contain relatively little data in the hotel features dataset, are not considered in this study, and the following 8 features are used:

- 1) Child-Baby Friendly
- 2) Pool
- 3) Pool-Summer
- 4) Other Beach
- 5) Spa-Thermal Hotel
- 6) Sauna-Turkish Bath
- 7) Fitness
- 8) Room Breakfast

In these two data sets, only the data to be used in this study is obtained, the sales data are sorted on the basis of the customers' code, and the hotel features data are kept to be integrated with the customers' choices.

After determining the hotels selected by the customer codes, the value of 1 or 0 assigned to the hotel selected by the customer from the data set with the characteristics of the hotels positioned with "=VLOOKUP()" in Microsoft Excel Program.

It is determined by the IFERROR formula that some of the hotels selected by the customers are not in the hotel feature list and are deleted from the list.

In order to obtain content-based filtering, the average value of the features selected by the customers in the hotels is taken with the pivot table, and the values given by the customers to each hotel feature are assigned as this average value. The brief section of the result matrix is given in Table 6.

Table 6. A Brief Example of Result Matrix of Content Based Filtering

Customer ID	Average-Child-Baby-Friendly	Average-Pool	Average-Spa-Thermal-Hotel	Average-Sauna-Hamam	Average-Fitness
13	1	1	1	1	1
31	1	1	1	1	1
141	0	0	0	0	0
166	1	1	1	1	1
300	1	1	1	1	1
318	1	1	1	1	1
329	1	1	1	1	1
430	0,6667	1	1	1	1
446	1	1	1	1	0
458	1	1	1	1	0
644	1	1	0,5	0,5	0
732	1	1	1	1	1

825	1	1	1	1	0
982	1	1	1	1	1
1032	1	1	1	1	0,667
1080	0,75	0,75	0,75	0,75	0,5
1090	1	0	0	0	0

16381								
16383								2
16385					1			
16440	1							
16465								5
16467								1
16468								
16469								1
16516			1					
16546								1
16559								1
16583								1
16584								10

5.2. Data Analysis Results by Using Collaborative Filtering Method

For collaborative filtering, due to the need to calculate the points given by each customer to each hotel, the number of visits by the customer to the hotels is calculated as the score given by that customer to the relevant hotel, and the result matrix is obtained. Additionally, it is aimed to increase the reliability by removing the customers who are included in the data set and whose features of the selected hotels are not included in the "hotel features" dataset (whole row is 0).

While applying this collaborative filtering method, the total number of selections of the hotels is calculated with the pivot table and hotels with less than 10 visits by customers are excluded from the matrix in order to obtain clearer and more reliable results.

In the end, a matrix is obtained with customers in the rows and hotels in the columns, and the scores given by the customers to each hotel as seen in Table 7.

Table 7. A Brief Example of Result Matrix of Collaborative Filtering

CUSTOMER ID	HOTEL ID SCORES							
	370	383	384	385	386	419	445	466
430	2							
732					1			
821								1
825						1		
1080		1						
1134							1	
16230								1
16234							1	
16247								1
16278					1			
16378								3
16379								1

5.3. Data Analysis Results by Using RFM Method

Since the values we will calculate with RFM analysis are recency, frequency, and monetary, the values we will use in our 2 separate data sets are as follows:

- 1) Customer code
- 2) Billing information
- 3) Sales date
- 4) Sales Price

While analyzing these values, the pivot table is used and the customer code is selected as the line, the date of the last sale, how many invoices are issued, and the total sales price are determined.

Recency Scores:

A recency matrix is created to find customers' recency scores. The recency matrix is divided into percentiles with the PERCENTILE.INC formula. The customer who visited us the longest time is determined and scored 1, then the matrix is divided into the following percentages, in order: 20, 40, 60, and 80. The recency scores are assigned as in Table 8.

Table 8. Recency Matrix

RECENCY MATRIX	
Recency	Recency Score
1.01.2019	1
12.04.2019	2
18.07.2019	3
18.12.2019	4
30.07.2020	5

According to the created recency matrix (Table 8), R-Scores are assigned to customers by the VLOOKUP formula.

Examples of the result of the R-Score processed into the RFM matrix are shown in table 11.

Frequency Scores:

While assigning frequency values to customers, the PERCENTILE.INC formula is used. The customer who visits the company at least should get 1 as the frequency value. After calculating the customer with the least visits to the company with the pivot table, the most recent visit dates are divided into the following percentages, in order: 20, 40, 60, 80, 90, and 95. The frequency matrix is given in table 9.

Table 9. Frequency Matrix

FREQUENCY MATRIX	
Frequency	Frequency Score
1	3
2	4
3	5

As seen in Table 9, the frequency value of the customer is 1 up to the 80% slice. This value means that at least 80% of customers shopped from the company only once.

According to the created frequency matrix (Table 9), F-Scores are assigned to customers by the VLOOKUP formula.

Examples of the result of the F-Score processed into the RFM matrix are shown in table 11.

Monetary Scores:

According to the total sales price amount created with the pivot table in the RFM table, the monetary matrix is created with the PERCENTILE.INC formula and monetary scores are assigned. The results are given in Table 10.

Table 10. Monetary Matrix

MONETARY MATRIX	
Monetary	Monetary Score
0	1
5572800	2
11696000	3
18920000	4
30100000	5

According to the created monetary matrix (Table 10), M-Scores are assigned to customers by the VLOOKUP formula. Examples of the result of the M-Score processed into the RFM matrix are shown in table 11.

Table 11. Recency, Frequency, and Monetary Scores of the Customers

Customer ID	The Last Sale Date	Number of Master ID	Total Selling Price	R Score	F Score	M Score
13	15.07.2020	2	84753000	4	4	5
31	17.07.2020	3	19541300	4	5	4
141	16.07.2020	1	13557900	4	3	3
166	19.12.2019	2	31471700	4	4	5
171	29.09.2020	3	38562100	5	5	5
179	5.08.2020	3	10294200	5	5	2
191	29.08.2020	1	64396800	5	3	5
217	8.10.2019	1	1654400	3	3	1
254	10.02.2020	1	2285900	4	3	1
300	28.12.2019	1	9967400	4	3	2
318	5.08.2019	1	17028000	3	3	3

5.4. Customer Similarities

According to focus groups, Manhattan distance is calculated using python and the closest customers to each other are determined. Here, customers who have made the same choices are removed, and outputs are obtained.

Table 12 shows an example of the customers who most closely resemble each customer.

Table 12. Output of Manhattan Distances

Customer	Compatible Customer		
31	4066		
13	2887		
166	171	2385	4066
171	166	2554	2887
1080	13	1285	2021

6. Conclusions

The recommendation system is an area in the industrial system that has multiple content-based, collaborative and hybrid approaches to increase companies' growth and productivity.

Recommendation systems provide access to personalized information on the web and have progressed in the last 10 years.

Recommendation systems created new options for information search and filtering. For instance, online shopping stores have increased their profits and music lovers discovered new songs.

Besides the positive effect of the recommendation system on customers, there are some limitations and deficiencies. This paper has reviewed three different recommendation approaches in detail. Additionally, this paper has reviewed Recency, Frequency, and Monetary analysis in detail.

Acknowledgment

This article is partially supported by Galatasaray University Research Fund (BAP) within the scope of project number fba-2021-1063, and titled "Niteliklendirilmiş çift yönlü ağlarda bağlantı tahmini ile öneri sistemleri geliştirilmesi".

Author's Note

Previous version of this paper was presented at 6th International Conference on Engineering Technologies (ICENTE'22), 17-19 November 2022, Konya, Turkey.

References

- [1] H. Poor, *An Introduction to Signal Detection and Estimation*. New York: Springer-Verlag, 1985.
- [2] S. Chen, B. Mulgrew, and P. M. Grant, "A clustering technique for digital communications channel equalization using radial basis function networks," *IEEE Trans. Neural Networks*, vol. 4, pp. 570–578, July 1993.
- [3] G. R. Faulhaber, "Design of service systems with priority reservation," in *Conf. Rec. 1995 IEEE Int. Conf. Communications*, pp. 3–8.
- [4] J. Williams, "Narrow-band analyzer (Thesis or Dissertation style)," Ph.D. dissertation, Dept. Elect. Eng., Harvard Univ.,

- Cambridge, MA, 1993.
- [5] Motorola Semiconductor Data Manual, Motorola Semiconductor Products Inc., Phoenix, AZ, 1989.
- [6] R. J. Vidmar. (1992, August). On the use of atmospheric plasmas as electromagnetic reflectors. *IEEE Trans. Plasma Sci.* [Online]. 21(3). pp. 876–880. Available: <http://www.halcyon.com/pub/journals/21ps03-vidmar>
- [7] Isinkaye, F. O., Folajimi, Y. O., & Ojokoh, B. A. (2015). Recommendation systems: Principles, methods and evaluation. *Egyptian Informatics Journal*, 16(3), 261-273.
- [8] Bobadilla, J., Ortega, F., Hernando, A., & Gutiérrez, A. (2013). Recommender systems survey. *Knowledge-based systems*, 46, 109-132.
- [9] Seyednezhad, S. M., Cozart, K. N., Bowllan, J. A., & Smith, A. O. (2018). A review on recommendation systems: Context-aware to social-based. *arXiv preprint arXiv:1811.11866*.
- [10] Pazzani, M. J., & Billsus, D. (2007). Content-based recommendation systems. In *The adaptive web* (pp. 325-341). Springer, Berlin, Heidelberg.
- [11] Shani, G., & Gunawardana, A. (2011). Evaluating recommendation systems. In *Recommender systems handbook* (pp. 257-297). Springer, Boston, MA.
- [12] Seyednezhad, S. M., Cozart, K. N., Bowllan, J. A., & Smith, A. O. (2018). A review on recommendation systems: Context-aware to social-based. *arXiv preprint arXiv:1811.11866*.
- [13] Hill, W., Stead, L., Rosenstein, M., & Furnas, G. (1995, May). Recommending and evaluating choices in a virtual community of use. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 194-201).
- [14] Chien, Y. H., & George, E. I. (1999, January). A bayesian model for collaborative filtering. In *AISTATS*.
- [15] Sun, Z., & Luo, N. (2010, August). A new user-based collaborative filtering algorithm combining data-distribution. In *2010 International Conference of Information Science and Management Engineering* (Vol. 2, pp. 19-23). IEEE.
- [16] Kaya, T., & Kaleli, C. (2022). A novel top-n recommendation method for multi-criteria collaborative filtering. *Expert Systems with Applications*, 198, 116695.
- [17] Lv, X. (2021, March). Analysis and Optimization Strategy of Travel Hotel Website Reservation Behavior Based on Collaborative Filtering. In *2021 International Conference on Intelligent Transportation, Big Data & Smart City (ICITBS)* (pp. 362-365). IEEE.
- [18] Alsayat, A. (2022). Customer decision-making analysis based on big social data using machine learning: a case study of hotels in Mecca. *Neural Computing and Applications*, 1-22.
- [19] Wu, J., Liu, C., Wu, Y., Cao, M., & Liu, Y. (2022). A novel hotel selection decision support model based on the online reviews from opinion leaders by best worst method. *International Journal of Computational Intelligence Systems*, 15(1), 1-20.
- [20] Bueno, I., Carrasco, R. A., Porcel, C., Kou, G., & Herrera-Viedma, E. (2021). A linguistic multi-criteria decision making methodology for the evaluation of tourist services considering customer opinion value. *Applied Soft Computing*, 101, 107045.
- [21] Chang, J. L., Li, H., & Bi, J. W. (2022). Personalized travel recommendation: a hybrid method with collaborative filtering and social network analysis. *Current Issues in Tourism*, 25(14), 2338-2356.
- [22] Berker Türker, B., Tugay, R., Öğüdücü, Ş., & Kızıllı, İ. (2020). Hotel Recommendation System Based on User Profiles and Collaborative Filtering. *arXiv e-prints*, arXiv-2009.