International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

Hostel Finder: Location-Based Recommendation System for Hostels and PGS with Transit Information

***Basavesh D*[1], *Laharishree S*[2], *Sthuthi S*[3], *Tejaswini N*[4], *Vidya R*[5]**

[1]Assistant Professor, Department of computer Science Jyothy Institute of Technology Bangalore, India basaveshd@jyothyit.ac.in
[2]Department of computer Science Jyothy Institute Of Technology Bangalore, India laharishree.s@gmail.com
[3]Department of computer Science Jyothy Institute Of Technology Bangalore, India sthuthi.cse@gmail.com
[4]Department of computer Science Jyothy Institute Of Technology Bangalore, India tejaswinin258@gmail.com
[5]Department of computer Science Jyothy Institute Of Technology Bangalore, India vidyaraju129@gmail.com

DOI - https://doi.org/10.55248/gengpi.4.523.42455

ABSTRACT—

In the recent years, there has been a rise in immigration. Most of these immigrants are students and employees. Most of these individuals require long-term, affordable accommodation based on their preference. When these individuals arrive at their destination, one of the major problems faced by them is accommodation. Due the language and cultural barrier they find it difficult to search a suitable Hostel or PG. This project involves identifying accommodation for these students/Employees according to their choices for amenities, affordability and proximity to the place in order to apply K-Means clustering to find the best accommodations in the chosen location. Recommendation system helps to estimate and predict user preference, our project uses content-based filtering as the recommendations are specific to user.

Keywords—K-means, Content-based filtering, recommendation system, clustering, geolocation, visualization.

I. Introduction

Migration is the process of moving from one destination to another. People migrate for a number of reasons, such as the search for work, economic opportunities, or higher learning. There has been a large amount of migration throughout the nation, which is primarily being done by students/Employees in order to receive a better education/Job in another state. This is one of the oldest activities, wherein students travel a great distance to enable professional study or even to enjoy the benefits of the university's increased opportunities.

This duty is becoming increasingly tough since the majority of students go alone to entirely new places they have never been before. These students typically don't speak the regional language well and don't have any friends or acquaintances to help them with any tasks or issues they might encounter in that state.

The absence of support makes it difficult to carry out routine daily tasks, some of which may be relatively basic. Finding a suitable place to stay is one the difficult tasks as it requires understanding of the local language and their culture. It is hard to gain such information without being native or staying in the locality for few days. This creates a challenge for international students to seek subsidized housing that fulfils their requirements.

This creates a number of obstacles for people who need to stay anywhere for a long period of time, failing which they may be forced to return to the outside. The creation of an accommodation is required, which leads to the creation of a recommendation system that can assist these persons in finding an accommodation.

As a result, an effective strategy is necessary to propose a proper alternative for either a hospitality option based on the student's preferences. This sort of recommender system is uncommon and has been demonstrated to achieve extremely low accuracies.

This research article defines an appropriate approach to obtaining residence recommendation with the help of machine learning methodologies.

As a result, an effective method is necessary for providing a reasonable solution for something like an hospitality option based on the student's preferences.

Our research article efficient approach to retrieve paying guest rooms and hostels for the users preferred location. This will be achieved using K-Means clustering algorithm to find the accommodation for the users location. We also use content-based filtering approach to find suitable hostel or PG rooms according to user's preference.

The purpose of this project is to provide users a recommendation system which allows them to choose an accommodation based on their budget, amenities and proximity of location.

II. LITERATURE SURVEY

K-means algorithm is one of the simple and efficient clustering algorithms in the field of machine learning. It is widely used technology among the existing algorithms, but due to noisy data and outlier the accuracy of the results may be reduced, therefore we are utilizing Improved K-means clustering proposed by Hui Xu 1[1] where clique grid is used to remove the noisy data.

The purpose of K-means algorithm is to find similarity between data by iteratively minimizing the measure between cluster center and observed data [2]. K-means is an unsupervised algorithm that are used in clustering dataset.

K-means algorithm uses Euclidean distance to measure the similarity between the data points. The similarity between the objects is inversely proportional to the Euclidean distance between them. The greater the similarity, the smaller the distance [3].

Technological development has caused many businesses to move online. During pandemic most of the business converted from offline to online mode because of these users are confused to choose the suitable product. Therefore, usage of recommendation system provides users a better perspective on the product. User may choose the product using keywords [4].

Recommendation system gathers information and provides an algorithm which considers users diverse needs. The recommendation system is broadly classified into three approach Collaborative, Content-based and Hybrid recommendation approach.[5]

Recommendation system has evolved gradually over past years; however, many systems use user’s emotion as feedback based on this recommendation system suggests the product [6].

By analyzing user preferences, the recommender system provides customized recommendations. However, performance drops dramatically when it encounters sparse data, particularly when it confronts a cold start user. In order to overcome this issues Kai Zhang research, suggest to combine content-based filtering and collaborative filtering to achieve higher accuracy.[7]

Content based filtering is one of the common methods used in building recommendation system. Content-based retrieves data from content of the item, sometimes they are words or features which describes the item. Based on this data recommendation system suggests the product [8].

Thousands of websites provide many advice on choosing a accommodation which would suits the users preferences. Using machine learning algorithm provides best solution which is preferred by users.[10]

Conventional recommendation system recommends accommodation based on pricing and rating but does not consider surrounding environment like bus stop, metro station. Based on the user’s preferences about the surrounding environment, the recommendation system suggests suitable accommodation to the user.[11]

The user's location is critical information that can be linked to the existing user profile to deliver effective service recommendations. Furthermore, the widespread availability of GPS-enabled devices results in a significant number of GPS trajectories representing user mobility records. These GPS trajectories can be used to uncover interesting user patterns. We investigated the value and application of information acquired from GPS trajectory data of users in recommender systems.[12]

Recommendation system provides suggestion of items or services based on user's preference or previous purchases. But in recent advancements recommendation system recommends services based on geolocational data.

It also gives a more appropriate choices for the user from the large dataset.[13]

People can find what they want by using the recommendation system, which suggests potential products of interest to them. It frequently uses existing associations between users and/or things to predict people's preferences for certain items. The recommendation system is now piquing the interest of the social network engineering and academic research sectors.[14]

When looking for a place to stay, it is typical to see a list of accommodations that match a query, ordered in descending order of the average assessment value. Because such a list does not reflect user preferences, determining a hotel takes too long for many unskilled users. We offer a method for extracting review author preferences from a set of reviews.

The extracted preferences are used for hotel recommendations in such a way that a contributor's evaluation value for having preferences comparable to the user is given more weight. According on the results of questionnaire-based evaluations, our suggested system can recommend hotels that meet the user's preferences.[15]

III. Problem statement

Imagine a student or staff member who has just moved to a new location. He/she has specific preferences and it would be challenging to find the best accommodation based on his/her preferences in terms of facilities, budget and proximity to location.

The aim of Immigration geolocation analysis and recommendation system is to recommend the best accommodation facility based on the user's preferences.

It also provides information on nearby transportation options from the user's desired destination.

4. Objectives

In our immigrant geolocation analysis and recommendation system we have the following objectives:

1. To retrieve all the available Hostel/PG rooms in the user's city.
2. To filter PG/Hostel rooms using user's requirements in the given location by the user.
3. To present the nearest transportation option to the PG/Hostel room.
4. Implement recommendation system to retrieve affordable PG/Hostel rooms.
5. To estimate a price for PG/Hostel room based on user's preference.

5. Proposed methodology

Our project will collect the data that is user's city and accesses its data set which contains all the location details. The collected data will be further cleaned which removes all the outliers and noisy data.

This processed data will be clustered based on user's locality using K-means algorithm. Using foursquare API, we will access all the hostel and paying guest room in the locality. The obtained dataset will be filtered based on user's preferences by content-based filtering.

Lastly, all the resultant data will be visualized using folium or maps.

In our project we have the following steps to recommend the hostel/PG rooms to the user:

A. Data Collection

In this stage, we will collect the data of given city by the user. This will be done by accessing the data set of the city. In here data means the information each location in the city. The data set is the set of information of the whole city.

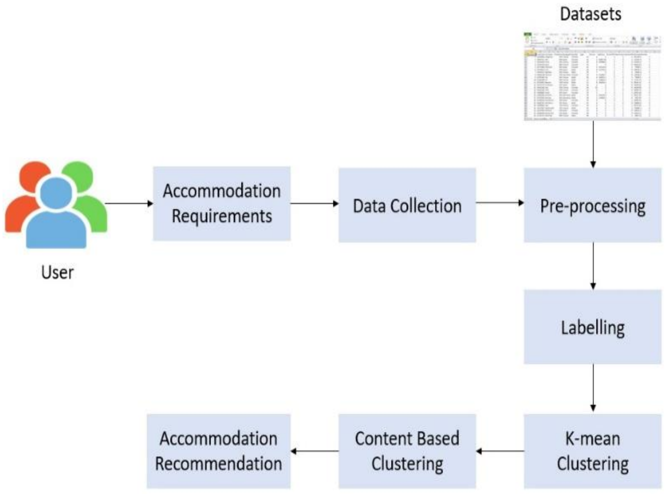


Figure IV.I: Immigrant geolocation analysis and recommendation system overview

B. Data Cleaning

In this stage, we will take the collected data from the first stage, and clean the data, that is eliminating the noisy data and outliers which are not needed for our objective. This will be done using Pandas. This makes the data's more city centric, as all the unwanted data is eliminated. And will be prepared for analysis.

C. Data Visualisation

This step, will visualise the gathered data in boxplots on the graph. All of the city cleaned data will be visualised, which can be achieved using Seaborn or Matplotlib or Pandas.

D. Fetching Geolocational Data

The location given by the user as input will be taken and using the Geopy API, which gives the access for an updated data of the given location.

Geopy API is a python library, used by many developers, to access the Geolocation data of the user input, it will take the user's location in longitude and latitude format with comma separated, and access that location to gather more information of the location and the user experience of that location, other venues, etc.

E. Clustering the data

In this step, the gathered geolocational data will be clustered and the hostels and paying guest data nearest to the given location will be returned as a set, of all the hostels and paying guest rooms. To achieve this, we will be using K-Means algorithm, it is a clustering algorithm, based on similarity and we can use ScikitLearn.

ScikitLearn is a machine learning tool, in this there are different prediction algorithms like Classification, Clustering, Regression. In which we'll be using Clustering, to cluster all the similar data objects, In our project we will be clustering all the Hostels and Paying Guest Rooms data whose geolocation is similar to the location given by the user.

By this we can access the Hostel and Paying Guest Room data which are closer to the given user location.

F. Visualising data on Map

Now the data which we have gathered during the clustering process, we will present them on the map, we will use the world widely used Google Maps, to present our findings.

To present the finding on the map, we can use either the Folium or Seaborn.

Both Folium and Seaborn are used to visualise the data. Folium is more concentrated on the maps visualisation, where the data is presented on the map. In Folium, it will take the clustered set which was the output of K-Means and will present the output on the map, by taking the longitude and latitude of each item from the set, with its name.

G. Recommendation based on User Preference

In this, we will be suing Content-Based Filtering, this is an application of Natural language processing, where the recommendation is based on what the user has given the input. This will be achieved as, there will be a filtering option for the user, and he/she can select their preferences, like Food, Price, accommodation choices.

The preferences of the user is given in the form of list, and the Content-Based Filtering will view the Accommodation, based on the user's preference. There has been many applications of the Content-based filtering like in Netflix, movie recommendation.

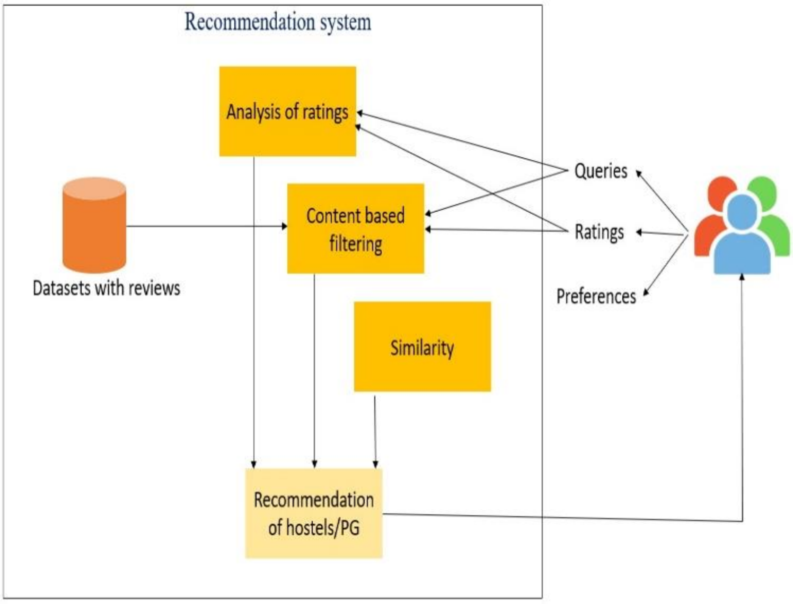


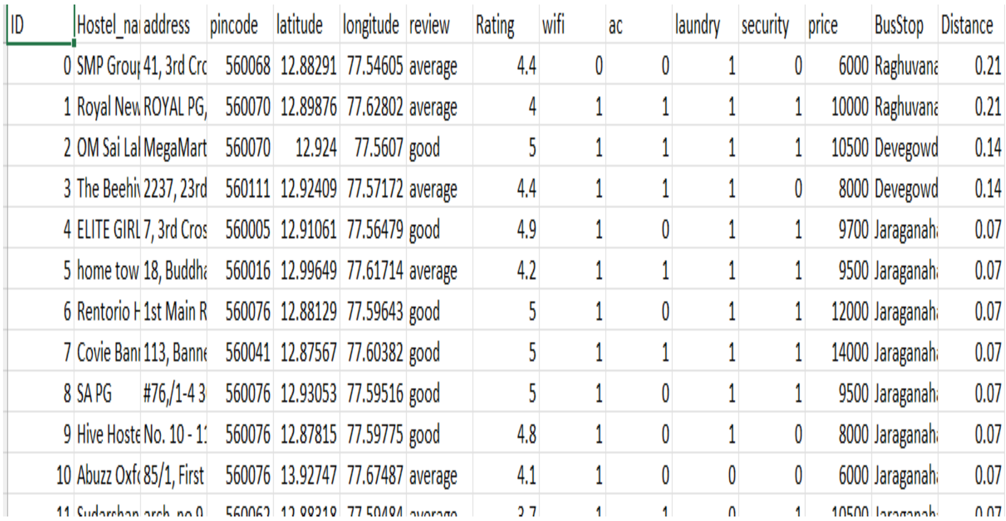
Figure IV.II: Overview of recommendation system

6. Implementation

The proposed methodology consists of two approaches, a clustering model using Machine learning and natural language processing is carried out.

Cluster-Based Recommendation

In the first approach, a clustering-based K-means algorithm is employed which takes several parameters such as location, price, rating, Wi-Fi, laundry, security as input. In fig- VI. dataset sample is shown. Data pre-processing is performed to eliminate the noisy and outlier data.



ID	Hostel_na	address	pincode	latitude	longitude	review	Rating	wifi	ac	laundry	security	price	BusStop	Distance
0	SMP Grou	41, 3rd Crc	560068	12.88291	77.54605	average	4.4	0	0	1	0	6000	Raghuvana	0.21
1	Royal New	ROYAL PG,	560070	12.89876	77.62802	average	4	1	1	1	1	10000	Raghuvana	0.21
2	OM Sai Lal	MegaMart	560070	12.924	77.5607	good	5	1	1	1	1	10500	Devegowd	0.14
3	The Beehi	2237, 23rd	560111	12.92409	77.57172	average	4.4	1	1	1	0	8000	Devegowd	0.14
4	ELITE GIRL	7, 3rd Cros	560005	12.91061	77.56479	good	4.9	1	0	1	1	9700	Jaraganah	0.07
5	home tow	18, Buddha	560016	12.99649	77.61714	average	4.2	1	1	1	1	9500	Jaraganah	0.07
6	Rentorio H	1st Main R	560076	12.88129	77.59643	good	5	1	0	1	1	12000	Jaraganah	0.07
7	Covie Ban	113, Banne	560041	12.87567	77.60382	good	5	1	1	1	1	14000	Jaraganah	0.07
8	SA PG	#76,/1-4 3	560076	12.93053	77.59516	good	5	1	0	1	1	9500	Jaraganah	0.07
9	Hive Hoste	No. 10 - 1	560076	12.87815	77.59775	good	4.8	1	0	1	0	8000	Jaraganah	0.07
10	Abuzz Oxf	85/1, First	560076	13.92747	77.67487	average	4.1	1	0	0	0	6000	Jaraganah	0.07
11	Sudarshan	arch, no 0	560062	12.88318	77.50484	average	3.7	1	1	0	1	10500	Jaraganah	0.07

Figure VI.I: Dataset

The dataset used contains information about hostels, including their geographical coordinates (latitude and longitude), price, rating, and amenities such as Wi-Fi, air conditioning, laundry, and security.

The model begins by importing the necessary libraries, including pandas for data manipulation, NumPy for numerical operations, scikit-learn for machine learning algorithms, and Folium for visualizing geographical data.

Next, the dataset is read from a CSV file and preprocessed. The features of interest are selected, and standard scaling is applied to normalize the feature values. The optimal number of clusters is determined by calculating the silhouette score for different numbers of clusters.

The K-means algorithm is then applied with the optimal number of clusters to perform clustering on the preprocessed data. The K-means clustering algorithm is applied for different values of k ranging from 2 to 10. The within-cluster sum of squares (WCSS) is calculated as the inertia_ attribute of the K-Means object, which represents the sum of squared distances between each sample and its nearest cluster center.

Finally, the elbow curve is plotted, showing the relationship between the number of clusters (k) and the WCSS. The optimal number of clusters can be determined by identifying the "elbow" point in the curve, which is the point of diminishing returns in terms of reducing WCSS.

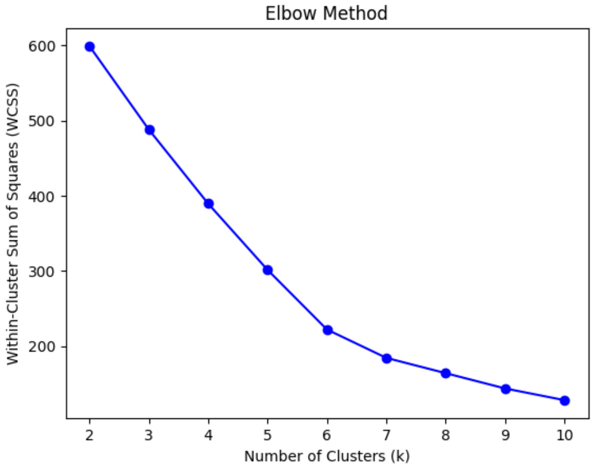


Figure VI.II: Elbow Method to optimal number of clusters

By visually inspecting the elbow curve, we can identify the appropriate k value for our clustering task is 6.

In the second approach, we have proposed a natural language processing which uses content-based filtering. This is a hotel recommendation system based on textual similarity using NLTK (Natural Language Toolkit) in Python. The dataset used contains information about hotels, including their addresses, ratings, names, tags, and localities.

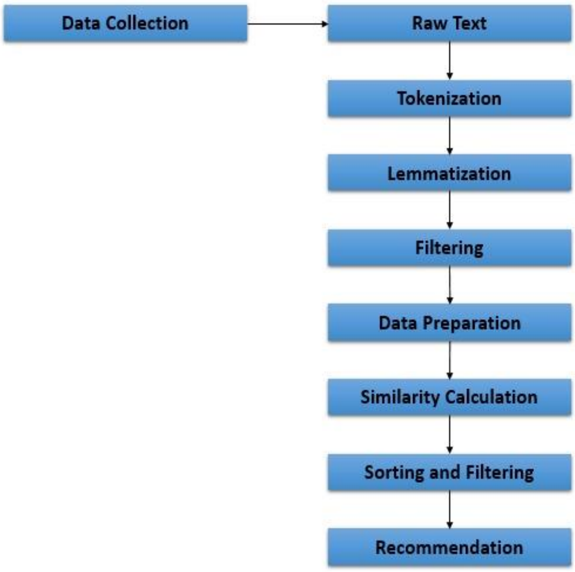


Figure VI.III: Overview of Content-based filtering

The model begins by importing the necessary libraries, including NLTK for natural language processing tasks and pandas for data manipulation.

The dataset is read from a CSV file, and some preprocessing steps are applied. The 'Locality' and 'Tags' columns are converted to lowercase for consistency. It takes a location and a description as inputs. The description is tokenized, and common stopwords are removed. The lemmatization process is applied to reduce words to their base form. A set of filtered words is created based on the description.

The model then searches for hotels in the specified location and calculates the similarity between each hotel's tags and the filtered set of words from the description. The similarity is calculated by finding the intersection of the two sets and counting the number of common elements.

The hotels are ranked based on their similarity scores, and in case of ties, the ranking is further sorted based on the hotel's rating.

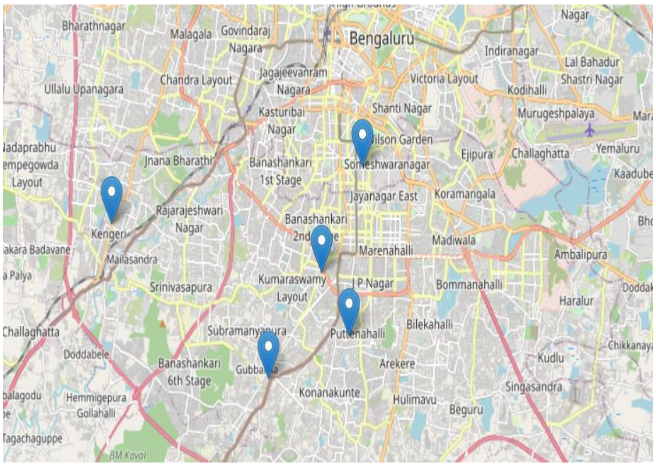


Figure VI.III: Data Visualization using Folium

The proposed methodology consists of price prediction model linear regression model for price prediction using the scikit-learn library in Python.

The dataset contains information about different hostels, including features such as the number of beds, whether the hostel has air conditioning (Ac), security measures, food type, rating, and laundry facilities. The target variable to be predicted is the price of the hostel.

The independent variables (features) are selected and stored in the variable X, while the target variable is stored in the variable y. The features include number of beds, Ac, security, food type, rating and Laundry. These features will be used to train the linear regression model to predict the price.

Linear regression model with an accuracy of 70.275% was implemented.

7. Result and Analysis

This model performs clustering on a dataset of hostels based on their features such as location, price, amenities, and ratings. It uses the K-means algorithm to group hostels into clusters. The optimal number of clusters is determined using the silhouette score and elbow method.

Given user preferences and location, the model recommends hostels that belong to the same cluster as the user's input. The recommendations are based on ratings and proximity to the user's location, calculated using the Haversine formula.

The model's results provide a list of recommended hostels with their names, addresses, prices, ratings, and distances from the user's location. This information helps users make informed decisions when choosing a hostel based on their preferences and location.

Content Based Recommendation takes a location and a description as input and recommends hostels based on their similarity to the input description and their rating.

First, the input description is preprocessed by converting it to lowercase, tokenizing it into words, removing stop words, and lemmatizing the remaining words.

Then, the model filters the dataset based on the provided location and calculates the similarity between each hostel's tags (keywords) and the preprocessed input description. The similarity is measured by the number of common words between the hostel's tags and the preprocessed description.

The hostels are ranked based on their similarity score and rating, and the final recommendations are returned as a sorted list of hostel names, ratings, and addresses.

The model provides recommendations based on the assumption that hostels with similar tags and higher ratings are more likely to match the preferences of the user. The result and analysis can be further enhanced by considering additional factors such as user reviews, amenities, and pricing.

8. Conclusion

In this research paper, we have implemented a Hostel-Finder a recommendation system, which takes user input of a location to find a suitable hostel/paying guest room according to user's preference. We firstly, access the user's city dataset, and cluster the hostel and paying guest room nearest to the user's location using K-means algorithm. We have used Geopy API to access hostel and paying guest room location. Later we find the accommodation based on user preference using content-based filtering.

Lastly, we present the findings on map using data visualization tool(folium).

References

[1]. Hui Xu, Shunyu Yao, Qianyun Li, Zhiwei Ye. An Improved K-means Clustering Algorithm. IEEE International Symposium September 2020

[2]. Haesik Kim. Performance Analysis of K Means Clustering Algorithms for mMTC Systems. December 2020.

[3]. Chen Jioe, Zhang Jiyue, Wu Junhui, Wu Yusheng, Si Huiping, Lin Kaiyan. Review on the Research of K-means Clustering Algorithm in Big Data. 2020 IEEE the 3rd International Conference

[4]. Tessy Badriyah, Sefryan Azvy, Wiratmoko Yuwono, Iwan Syarif

[5]. Recommendation System for Property Search Using Content Based Filtering Method. IEEE April 2018

[6]. Nishigandha Karbhari, Asmita Deshmukh, Dr. Vinayak D. Shinde. Recommendation System using Content Filtering. 2017 IEEE.

[7]. Khamael Raqim Raheem, Israa Hadi Ali. Content-based Recommender System Improvement using Hybrid Technique. 2020 IEEE.

[8]. Kai Zhang, Keqiang Wang, Xiaoling Wang, Cheqing Jin, Aoying Zhou. Hotel Recommendation based on User Preference Analysis. 2015 IEEE.

[9]. Kristian Wahyudi ,Johanes Latupapua, Ritchie Chandra, Abba Suganda Girsang. Hotel Content-Based Recommendation System. IOP Conference 2017.

[10]. Yuval Shavitt, and Noa Zilberman. A Geolocation Databases Study. IEEE 2021.

[11]. Rakesh Verma, Prince Verma, Abhishek Bhardwaja. Hotel Recommendation System Using Machine Learning. IJSR 2022

[12]. Zhichao Chang, Mohammad Shamsul Arefin, Yasuhiko Morimoto. Hotel Recommendation Based On Surrounding Environments. 2013 IEEE.

[13]. Sunita Tiwari, Saroj Kaushik, Shivendra Tiwari, Priti Jagwani. Location Based Recommender Systems: Architecture, Trends and Research Areas. IEEE 2013.

[14]. Priya Naik, Palak V. Desai, Supriya Pati. Location Based Place Recommendation using Social Network. IEEE 2019.

[15]. Jitechana, Lubna Shaikh, Shefali Bhattacharjee, Vaishali Yenolge, Dr. N. P. Kulkarni. An Accommodation Recommendation System for Immigrants using Exploratory Data Analysis on Geolocation Data. IJARSCT 2022.

[16]. Koji Takuma, Junya Yamamoto, Sayaka Kamei, Satoshi Fujita. A hotel recommendation based on reviews: What do you attach important to? IEEE 2016.