Personalized Tourist Attraction Recommendation System Using Collaborative Filtering on Tourist Preferences

Weeriya Supanich
Department of Information Technology
Faculty of Business Administration and Information Technology
Rajamangala University of Technology Tawan-Ok
Bangkok, Thailand, 10400
weeriya_su@rmutto.ac.th

Suwanee Kulkarineetham
Department of Information Technology
Faculty of Business Administration and Information Technology
Rajamangala University of Technology Tawan-Ok
Bangkok, Thailand, 10400
suwanee_ku@rmutto.ac.th

2022 19th International Joint Conference on Computer Science and Software Engineering (JCSSE) | 978-1-6654-8510-4/22/$31.00 ©2022 IEEE | DOI: 10.1109/JCSSE54890.2022.9836255

***Abstract*—A recommendation system becomes a good assistant in filtering various information from diverse sources to perform a matching result to users. These systems can provide a list of recommendations personalized to user preferences and needs. Almost any business can benefit from a recommendation system, including the tourism industry. In this paper, A personalized tourist attraction recommendation system (PTARS) based on a collaborative filtering technique is proposed. The research objective is to find the best model to recommend a customized destination to a new target user based on their preferences and behavior by using a user's travel-related data source acquired by an explicit approach. Our research result exhibits that the best similarity measure that yields the most accurate result is Euclidean distance; that calculation was from the top 25 k-neighbor values.**

Keywords— Travel recommendation system; Tourist attraction recommendation; Collaborative filtering;

I. Introduction

Nowadays, we buy products or use various services from websites or online systems. We will find that what comes with when choosing a product or using a service is recommending other related products or services by the recommendation system. It might be advisable to read some books or buy something or recommend watching certain movies or listening to some songs.

A recommendation System (RS) is a system that attempts to recommend the most suitable items (products or services) to particular users by predicting a user's interest in an item based on related information [1]. Almost any business can benefit from a recommendation system. We can identify industries that would benefit from recommendation systems, such as e-commerce, media companies, social media platforms for advertising, and the medical domain for diet, treatment, and pharmaceutical suggestions [2].

Being researchers in Thailand, which is a country where tourism is an economic contributor to the country. Estimates of tourism revenue directly contributing to GDP from 9% to 17.7% in 2016 [3]. Therefore, we believe in applying a recommendation system technique to the tourism industry would be a beneficial output to the sector.

One of the primary purposes of tourism is to visit various tourist attractions. The type of attraction, the time required at each spot, traveling time between the spots, users' interests, and transportation are some of the deciding factors. When planning for a trip, these factors are taken into account. However, searching through various travel blogs and websites to collect the required information for such places becomes tiring and time-consuming. Consequently, a recommendation system for tourist attractions may be desired [4].

Different recommendation approaches have been proposed by researchers to provide better results. Commonly used approaches include collaborative filtering (CF), content-based (CB), and hybrid approaches. Each approach has advantages and limitations; for example, CF has sparseness, scalability, and cold-start problems, while CB has overspecialized recommendations [1] and cannot represent user interests precisely in some cases such as when a user may refer to products on behalf of others, not for themselves [6].

One of the greatest challenges in developing a travel recommendation system (TRS) that provides personalized recommendations of tourist destinations is to enhance the tourist decision-making process. In order to achieve this, it requires a deep understanding of the tourists' decision-making and developing novel models for their information search process [7]. Various tourist attractions may serve different purposes for visitors. The selection of appropriate destinations is largely dependent on individuals' preferences and contexts. Many people want to have different kinds of experiences while traveling, such as exploring cultural tourism, going to adventure camps, or preferring leisure and relaxation. Such domain-specific contexts can be used to make recommendations for specific locations [4].

Most of the previous TRSs that recommend tourist destinations have focused on different schemes for acquiring users' travel-related data from an implicit method by gathering user information from various online sources without interrupting the user's online activities. For instance, using a user checked-in data history [6], using geotagged-photo social media data [7], using user location with the global positioning system (GPS) [8], user sentiment and emotion recognition [9], the visual information extracted from user photos [10], user and friend's travel-related data from social media post [11]. With regard to technical aspects, However, the accuracy of user preferences extrapolation is one of the significant issues of the implicit method [6]. While the explicit method works directly by asking the user for individual information using a questionnaire can elicit more accurate data from users. Thus, the second method is preferred and deployed by some intelligent RSs.

Therefore, in this paper, we propose a personalized tourist attraction recommendation system (PTARS) using the

978-1-6654-3831-5/22/$31.00 ©2022 IEEEE

CF-based technique integrated with a user's travel-related data source acquired by an explicit approach. The research objective is to find the best model to recommend personalized attractions to new target users based on their preferences and interests.

The paper is organized into the following sections. Section 2 provides recommendation techniques in the tourism domain. Section 3 describes related works. Section 4 presents the proposed PTARS research methodology. The experimental results and analysis of this study are illustrated in Section 5. Finally, we present some tentative conclusions and our future works in the last section.

II. Recommendation techniques

A. *Collaborative filtering*

The collaborative filtering (CF) technique recommends items to a particular user based on the rating/opinions of the other users who share similar interests [1]. CF system performs recommendations by building a database of preferences for items by the user. The system then finds the user with similar interests and preferences by calculating similarities between the user profiles and building a group of similar users called a neighborhood. A user gets the recommendation to those products that have not been rated/purchased, but his neighbors are rated. Collaborative filtering performs predictions or recommendations; the prediction is a numerical value, and the recommendation is a list of top N items that the user will like the most [1, 5].

Collaborative filtering techniques can be classified into two broad categories (1) user-based CF and (2) item-based CF. The aim behind user-based CF is to suggest the items based on the ratings given by other members who have already examined the items. If a member finds an item interesting, it will be automatically recommended to other users who have expressed similar views in the past. To achieve this goal, this system will create a Users × Users matrix to hold the similarity ratings between users. As a result, the active user's potential rating of an item will be estimated based on his similar neighbors. In item-based CF, uses an Items × Items matrix to store the similarity scores between items. In practice, the system will suggest items that are the most similar to a group of items that the active user has already given a high rating. Indeed, the anticipated rating is based on the similarity value between the item and its neighbor. The main CF process is divided into three steps: similarity computation, neighborhood selection, and rating prediction [12].

B. *Content-based filtering*

Content-based (CB) technique recommends items that are similar in characteristic to the item that was previously preferred by a user [1, 5]. CB approach performs more analysis on the attribute of the item in order to produce recommendations. CB technique is most successful in web pages, publications, and news recommendations. The CB system automatically creates personalized profiles for users based on the user's feedback and item preferences. To generate appropriate recommendations, collected user information is compared to the item's characteristic [5].

C. *Hybrid filtering*

The hybrid filtering (HF) technique has been proposed to eliminate the limitations of content-based and collaborative filtering approaches [6] by a combination of two or more recommendation systems in order to get better performance over CF and CB approaches. It is possible to combine CF and CB techniques in a different way to obtain a hybrid filtering system, which may produce several outputs. There are seven different types of hybridization processes: (1) weighted (2) switching (3) mixed approach (4) feature combination (5) feature augmentation (6) cascade and (7) meta-level [5].

III. Related works

The CF approach was widely applied in tourism since it discovers the user's preferences by mining the user's historical behavior data, classifies the users based on different preferences, and recommends similar products with similar tastes [13]. One crucial basis aspect of this approach is the user travel behavior datasets. Some research acquired the user travel behavior datasets from an explicit approach by directly asking the user for individual information by using a questionnaire either online or offline. Pree T. et al. [7] proposed a novel human-centric TRS that recommends destinations to tourists in an unfamiliar city. In this work, a decision tree-based model was presented and the model was pre-processed and built offline. Another study by Joseph C. et al. [11] employed volunteer users' Twitter travel behavior datasets to develop TRS, and the tweets were then used to personalize recommendations about the place of interest for an active user. Historical buildings, museums, parks, and restaurants are examples of places of interest. In a survey, volunteers were asked to provide their Twitter handles as well as rank their preferred travel categories.

Some work acquired the user's travel-related data from an implicit approach, by gathering user information from various online sources without interrupting the user's online activities. M Al-Ghobari et al. [8] integrated the CF approach with user location data. They proposed LAPTA, which combines user preferences and the global positioning system (GPS) to provide tailored and location-aware recommendations. The suggested system uses the K-Nearest algorithm to match the name and category tags with the user's input in order to build personalized recommendations. In another work, K. Kesorn et al. [6] presented a technique that uses Facebook data to extrapolate user interests in tourism attractions using Facebook check-in data to recommend attractions for users.

IV. Research methodology

In this section, we present a framework for the personalized tourist attraction recommendation system (PTARS) as illustrated in Figure. 1. The overview system architecture consists of 4 phases including (1) data acquisition, (2) data pre-processing and transformation (3) modeling a PTARS, and lastly (4) model deployment through API.

A. *Data acquisition*

To understand tourists' behavior in assessing travel information and decision-making processing for destination choice, we use an explicit method by making an online questionnaire through Twitter as a data collection tool due to its effective mechanism for collecting information from tourists.

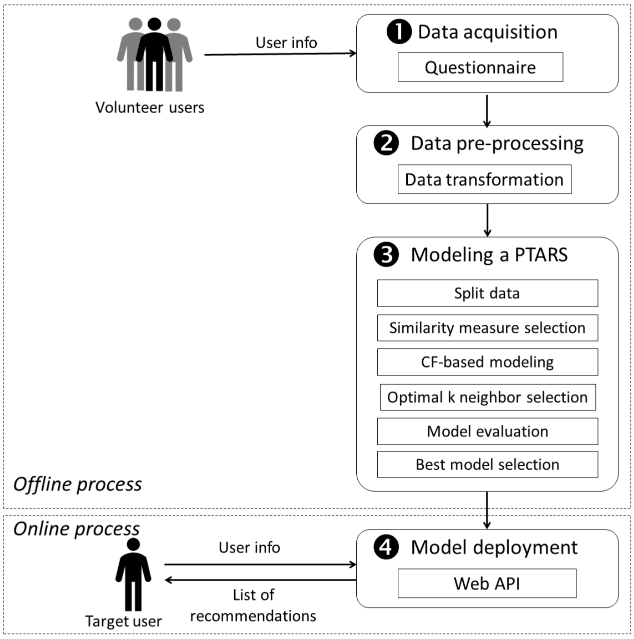

Fig. 1. The overview system architecture

We scope our research using tourist destinations in Chonburi province, Thailand. Due to Chonburi is a famous tourist destination in Thailand with beautiful beaches, cultural activities, nightlife etc. We ask volunteers who have traveled to Chonburi to fill in the online questionnaire. The question is divided into two parts containing a set of factors related to tourist's preferred destinations as follows:

1) Tourist sociodemographic and behavior information

Individual demographics may influence information-seeking behavior. Tourism behavior is one of the critical factors we have found from literature reviews when tourists select their destinations [7]. There are 11 questions as shown in Table 1.

2) Attraction rating information

The attraction rating information is used to provide a list of top N places recommendations to target users. The respondents rate their preferences for each attraction. The rating has 6 levels: 5 = strongly liked, 4 = liked, 3 = neutral, 2 = unknown, 1 = disliked, and 0 = strongly disliked. Table 1 depicts the list of 22 preferred attractions that was selected from the Trip-advisor website and can be categorized into 5 groups.

Four hundred fifty-three volunteers from Twitter took an online questionnaire about tourism behavior and tourist attractions in Chonburi province, Thailand. The proposed framework uses variables extracted from the questionnaire as inputs for the CF approach to the tourist's preferred destination

TABLE I. EXAMPLE LIST OF QUESTIONS

Questionnaire type	Example list of questions
Tourist sociodemographic and behavior information	1. Gender 2. Age 3. Education 4. Occupation 5. Income 6. Marital status 7. How often do you travel? 8. Which group of people do you travel with most often? 9. When do you choose to travel? 10. Which type of activity or attraction you like to do or want to visit? 11. What do you focus your money for the best value?
Attraction rating information	1. Koh Lan 2. Nong Nooch Tropical Garden 3. Cartoon Network Amazone 4. Pattaya Floating Market 5. Ripley's Believe It or Not Museum 6. Marine Science Museum 7. Takhian Tia Community 8. Sanctuary of Truth 9. Khao Chi Chan 10. The Tiffany Show

B. Data pre-processing and transformation

Using an online questionnaires tool via google forms, it is possible to create and design good questionnaires that can support user input according to the desired data format, such as defining required fields, making checkbox selections. This allows users to fill in the information correctly and completely. With the use of this tool, it reduces time spent and greatly reduces the burden of the data cleaning process.

From the questionnaire data, it was found that all topics were categorical data; therefore data transformation was required before using this dataset in further modeling steps. We use a one-hot encoding technique to convert categorical data variables to numerical data. We convert each categorical value into a new categorical column with one-hot encoding and assign a binary value of 1 or 0 to those columns. Each integer value is represented as a binary vector.

After doing one-hot encoding of the categorical data of all 11 tourist information questions with multiple choices of each question, a total of 62 choices, a Users × Items matrix size of 453 rows × 424 columns were generated.

C. Modeling a PTARS

The personalize tourist attraction recommendation system modeling process consists of six steps:

1) Split data

We separate questionnaire data into two datasets. The first dataset is the *user's sociodemographic and behavior dataset* which will be used to compute the user similarity matrix. The second dataset is the *attraction rating dataset* which will be used to recommend the top-N rated places to a target user.

One of the key aspects of supervised machine learning is model evaluation. In this research, the train_test_split() from the data science library scikit-learn will be used to split the dataset into two subsets: the training set (80%) and the test set (20%).

2) *Similarity measure algorithm selection*

Creating a user similarity matrix in a CF-based recommendation is quite a challenging task since the CF technique is extremely sensitive to the similarity measure used to evaluate the dependency strength between users [14]. There are many similarity measures to compute the distance between two users to determine the neighborhood. In this research, Cosine similarity, Pearson correlation, and Euclidean distance will be applied to the experiment. The system has to build a *Users* × *Users* matrix to store the similarity scores between users [15].

3) *Collaborative filtering based modeling*

The algorithm behind this is a user-based CF scheme. This method aims to predict the rating of places recommended to a target user on the non-rated items based on the user's interests and the neighborhood interests [15]. The algorithm in table 2. summarizes the user-based CF method. To suggest the rating of tourist destinations recommended, a common average measure is used by calculating the mean rating of the *i*-tourist destination.

TABLE II. THE USER-BASED CF ALGORITHM

```
Algorithm 1  The used-based CF of PTARS
Input     Users × Items matrix
Output    Three sets of top-N rated places

Begin
    For each three different similarity measures:
        1. From the Users × Items matrix, compute the user similarity
           matrix by applying each similarity measure between a target
           user (from a test dataset) and all other users (from the training
           dataset), using the users' sociodemographic and travel
           behavior data, the output would be a Users × Users matrix.
        2. Find the top-k most similar users.
        3. Calculate a mean value of the tourist attraction rating of the
           top-k most similar users.
        4. Recommend the top-N rated places to a target user.
    End For
End
```

4) *Optimal k-neighbor selection*

To select the set of nearest neighbors (in terms of similarity), the top-k technique is used, where k denotes the number of users. We set up an experiment of selecting the optimal top k-nearest neighbor value when using different similarity measures by using the concept of the Elbow method. The k-neighbor experiment was started at 10 neighbors to 100 neighbors, increasing for every 5 neighbors.

5) *Model evaluation*

We employ two well-known assessment measures in evaluating regression models to evaluate the performance of our PTARS [15].

Mean Absolute Error (MAE): It assesses the disparity between RS's predicted ratings and the ratings given by users. It returns a positive value.

Root Mean Square Error (RMSE): it represents the square root of the mean of the square of all of the differences between predicted values and observed values. It returns a positive value.

6) *Best model selection*

After experimenting with the performance of the three similarity measures and finding the optimal top-k nearest neighbor value, the similarity measure with the lowest MAE and RMSE values will be used in our model, which will present the result in detail next section

D. *Model deployment through API*

We built an API with Flask RESTful API and coding in Python language and deployed the API on the Heroku platform.

V. EXPERIMENTAL RESULTS

In this section, we present the experimental results of our method in generating rating predictions.

A. *Dataset analytics*

Figure 2, displayed some user sociodemographic features distribution.

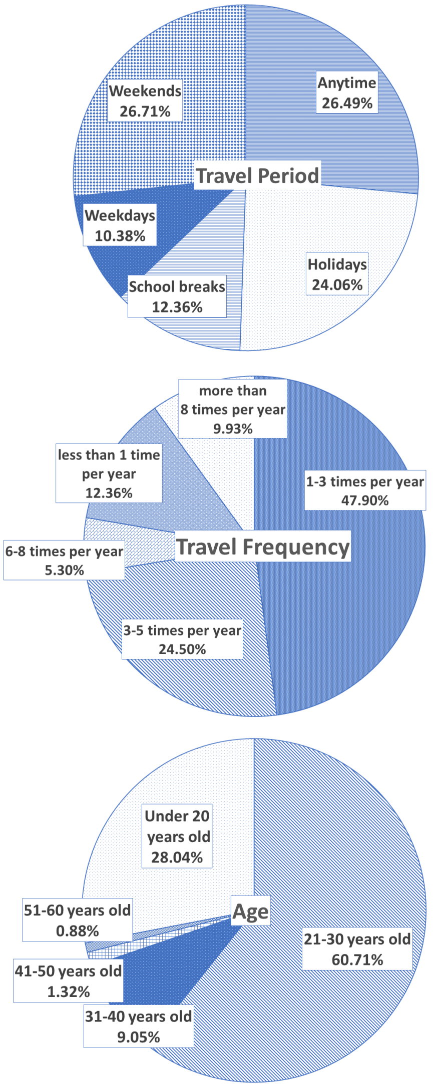

Fig. 2. Example of some sociodemographic feature distribution

For instance, the age feature was mainly distributed to 21-30 years old (60.7%), followed by under 20 years old (28.04%). The travel frequency feature illustrates nearly half of tourists often travel 1-3 times per year (47.90%), follow by 3-5 times per year (24.50%). The travel period survey shows the majority of people prefer to travel on both weekends (26.71%), and anytime (26.49%).

B. *PTARS modeling results*

In the PTARS evaluation, we set up an experiment of selecting the optimal top k-nearest neighbor value with the use of Cosine similarity, Pearson correlation, and Euclidean distance. The evaluation results of MAE and RMSE are illustrated in Figure 3.

Applying the concept of the Elbow method [16], we can pick the Elbow of the curve as the number of *k* neighbor values in our data-driven model. From our perspective the higher *k* neighbor value, the lower the two evaluation metrics values. Therefore, the value of *k* equal to 25 is chosen because it is the point after which the distortion/inertia starts decreasing in a linear fashion.

In order to determine the appropriate similarity measure to use in the system, both (a) and (b) in Figure 3 show that the Euclidean distance gives the best accuracy in both MAE and RMSE assessment metrics.

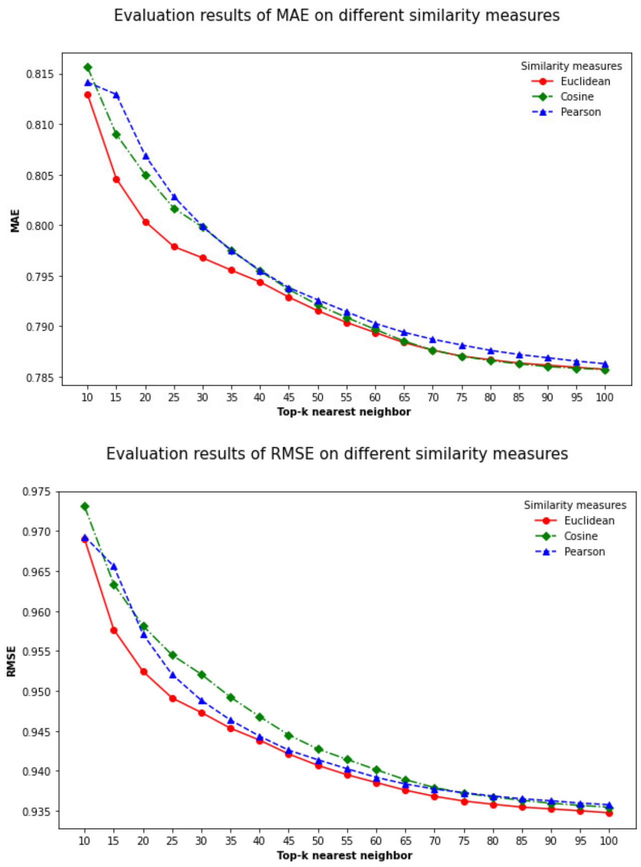

Fig. 3. Evaluation results on different similarity measures

From our Chonburi province tourism dataset and the experimental results, it can be concluded that the k-neighbor value best suited to our PTARS is k equal to 25. Evaluation metrics that yield the most accurate results are Euclidean distance, so we use Euclidean distance for similarity measure, and select a k-neighbor value equal to 25 to deploy the model.

Despite the fact that Euclidean distance is a simple similarity measure, the findings of this study reveal that it performs best in determining user similarity since the distance obtained using this formula represents the shortest distance between each pair of locations. While Cosine similarity showed differences in direction but not in position. It is appropriate for measuring document similarity and is widely used in text mining, natural language processing, and information retrieval systems. Pearson Correlation measures the strength of two variables in a linear relation. As the first computational step, the covariance value is used. However, the covariance is difficult to interpret because it does not show how far or near the data are to the line representing the pattern between the measurements [17].

The final model can be visited through a Web API at http://recommendermodel.herokuapp.com. Figure 4. displays an example output list of tourist attractions in Chonburi province of our API in a JSON format after a PTARS received an active user's preference information from an online questionnaire.

{'Ripley's Believe It or Not Museum': 4.2,
'Koh Samasarn': 4.0, 'Marine Science Museum': 4.0,
'Cartoon Network Amazone': 3.8, 'Papa Beach Pattaya':
3.8, ' Pattaya Floating Market ': 3.8}

Fig. 4. An example output list of tourist attractions in Chonburi province in a JSON format

VI. Conclusion and Future Works

In this paper, we offer an intelligent tourism recommender system that provides tailored suggestions of tourist attractions in a personalized way. The dataset has been decomposed into two sub-datasets. The first dataset is the user's sociodemographic and travel behavior dataset. The first dataset is used to find the top-k most similar users using the user-based CF technique. The second dataset is the tourist attraction rating dataset which is used to predict a list of tourist destinations to a target user. In our study, the Cosine similarity, Pearson correlation, and Euclidean distance were applied to the experiment in order to find the user similarity matrix. We study the optimal top k-nearest neighbor value selection using the Elbow method in our data-driven model. The study result shows the evaluation metric that yields the most accurate result is Euclidean distance, and the top-k most similar user value is 25.

For future work, we plan to add a new feature to our systems, such as collecting user usage data and the comments and rating of our suggested result list to develop an adaptive real-time PTARS

References

[1] J. Lu, D. Wu, M. Mao, W. Wang, and G. Zhang, "Recommender system application developments: A survey," Decision Support System, vol. 74, pp. 12–32, June 2015.

[2] P. Thiengburanathum, S. Cang, and H. Yu, "A Decision Tree based Recommendation System for Tourists," In Proceedings of the 21th International Conference on Automation and Computing (ICAC), University of Stratchclyde, Glasgow, UK, 11-12 September 2015.

[3] Tourism in Thailand https://en.wikipedia.org/wiki/Tourism_in_Thailand.

[4] K. Chaudhari, and A. Thakkar, "A Comprehensive Survey on Travel Recommender Systems," Archives of Computational Methods in Engineering, vol. 27, pp. 1545–1571, November 2020.

[5] P. Kumar, and R. S. Thakur, "Recommendation system techniques and related issues: a survey," International Journal of Information Technology, vol. 10, pp. 495-501, April 2018.

[6] K. Kesorn, W. Juraphanthong, and A. Salaiwarakul, "Personalized Attraction Recommendation System for Tourists Through Check-In Data," IEEE Access, vol. 5, pp. 26703-26721, November 2017. Available: 10.1109/ACCESS.2017.2778293.

[7] X. Sun, Z. Huang, X. Peng, Y. Chen, and Y. Liu, "Building a model-based personalized recommendation approach for tourist attractions from geotagged social media data," International Journal of Digital Earth, vol. 12, pp. 661–678, 2019. Available: 10.1080/17538947.2018.1471104

[8] M. AI-Ghobari, A. Muneer, and S. M. Fati, "Location-Aware Personalized Traveler Recommender System (LAPTA) Using Collaborative Filtering KNN," Computers, Materials and Continua, vol. 69, pp. 1553–1570, July 2021.

[9] L. Santamaria-Granados, J. F. Mendoza-Moreno, and G. Ramirez-Gonzalez, "Tourist Recommender Systems Based on Emotion Recognition—A Scientometric Review," Future Internet, vol. 13, pp. 1–37, January 2021. Available: 10.3390/fi13010002.

[10] WT. Chu, and YL. Tsai, "A hybrid recommendation system considering visual information for predicting favorite restaurants," World Wide Web, vol. 20, pp. 1313–1331, 2017. Available: 10.1007/s11280-017-0437-1.

[11] J. Coelho, P. Nitu, and P. Madiraju, "A Personalized Travel Recommendation System Using Social Media Analysis," In Proceedings of 2018 IEEE International Congress on Big Data (BigData Congress), San Francisco, CA, USA, 2-7 July 2018. Available: 10.1109/BigDataCongress.2018.00046.

[12] R. Hassannia, A. V. Barenji, Z. Li, and H. Alipour, "Web-Based Recommendation System for Smart Tourism: Multiagent Technology," Sustainability, vol. 11, pp. 1–18, January 2019. Available: 10.3390/su11020323.

[13] Z. Wang, and B. Liu, "Tourism recommendation system based on data mining," Journal of Physics: Conference Series, vol. 1345, November 2019. Available: 10.1088/1742-6596/1345/2/022027.

[14] K. Patel, and H. B. Patel, "A state-of-the-art survey on recommendation system and prospective extensions," Computers and Electronics in Agriculture, vol. 178, pp. 1–10, November 2020. Available: 10.1016/j.compag.2020.105779.

[15] F. Fkih, "Similarity measures for Collaborative Filtering-based Recommender Systems: Review and experimental comparison," Journal of King Saud University – Computer and Information Sciences, September 2021. Available: 10.1016/j.jksuci.2021.09.014.

[16] Elbow method (clustering) https://en.wikipedia.org/wiki/Elbow_method_(clustering)

[17] 17 types of similarity and dissimilarity measures used in data science. https://towardsdatascience.com/17-types-of-similarity-and-dissimilarity-measures-used-in-data-science-3eb914d2681.