JOURNAL OF CRITICAL REVIEWS

ISSN- 2394-5125 VOL 08, ISSUE 05, 2021

Machine Learning Based Hotel Recommendation System and Prediction

Pavankumar Thummeti , Rapolu Sai Kumar

Department of Computer Science and Engineering

Sree Dattha Group of Institutions, Hyderabad, Telangana, India.

ABSTARCT

Choosing a tourist destination from the information that is available on the Internet and through other sources is one of the most complex tasks for tourists when planning travel, both before and during travel. Previous Travel Recommendation Systems (TRSs) have attempted to solve this problem. However, some of the technical aspects such as system accuracy and the practical aspects such as usability and satisfaction have been neglected. To address this issue, it requires a full understanding of the tourists' decision-making and novel models for their information search process. This work proposes a novel human centric TRS that recommends destinations to tourists in an unfamiliar city. It considers both technical and practical aspects using a real-world data set we collected. The system is developed using two-steps feature selection method to reduce number of inputs to the system and recommendations are provided by recommendation system. The experimental results show that the proposed TRS can provide personalized recommendation on tourist destinations that satisfy the tourists.

Keywords: Travel Recommendation Systems (TRSs), supervised learning, feature selection.

1. INTRODUCTION

Recommender Systems (RS) is the information system that acquired information about the preferences of the users using the explicitly or implicitly methods. It then uses that information to predict what a user might like/dislike or recommend related items from a given set. The recommender system has been used in a different context such as the prediction of the movie in video streaming like YouTuber or NetFlix or the recommendation of the related book on Amazon.com. They are also one of the key fundamental architectures of most online services from shopping to newscasting to educational sites. One of the industries where the recommender system is generally considered necessary is in the travel/tourism sector. In the past decade, tourism is one of the largest industries in Thailand. According to the Ministry of Tourism and Sports of Thailand [1], there were about 305 million tourists visited Thailand in 2019 (including Thais and foreigner) which generated 2,781 billion baht. This gives Thailand rank eight among the countries with the highest income from this industry.

The main aim of using personalization techniques is to generate customized recommendation according to the user preferences and interests. The recommender system has an objective to filter unwanted information and to provide specific results for the user [2]. In the travel recommender systems [3], proposed model learns the user preferences and generates places of attractions according to the user interests. This paper focuses on the recommender systems and their application in tourism. To make this paper useful to all, including new readers of recommender systems, it covers topics from evolution to applications along with the challenges in it. Since more research is required to improve the effectiveness and efficiency of recommender systems, this paper will be more useful to the upcoming researchers to develop a user specific recommender system.

However, there is still a lag of the recommender system that recommends the tourist attractions in Thailand without the need for a user's effort. By considering the described aspects, this paper proposed a Tourism Recommender System (TRS) using the content-based filtering approach. It aims

to better assist users in locating the tourist attractions that suitable for them without significantly having to collaborate with the system. The system with getting the preferred attributes of tourist attractions from the user's Instagram photos. Machine Learning (ML) will then use to extract the terms from those photos. Those terms will be used to find the similarity index with the terms of the photos from tourist attraction using the vector space model. Lastly, the system will recommend the top 10 places with the highest similarity index with the user. The prototype of our TRS has been fully developed as a web application; a user study has been conducted to evaluate the effectiveness of the provided recommendations.

2. LITERATURE SURVEY

Thannimalai et al. proposed a new recommendation system for recommendation generation based on users' ratings and personal profiles. Motivated by existing studies, firstly this work proposed item-based collaborative filtering to recommend tourist spots based on users' rating. In addition, incorporated the content-based filtering algorithm with Naïve Bayes Classifier, for recommendation generation. Detailed analysis of these proposed methods is discussed which will give a clear view on how the core part of the recommendation systems has been implemented. The proposed TRS was evaluated using several data sets to indicate its efficiency.

Nilashi et al. proposed a new recommendation method based on multi-criteria CF to enhance the predictive accuracy of recommender systems in tourism domain using clustering, dimensionality reduction and prediction methods. This work used Adaptive Neuro-Fuzzy Inference Systems (ANFIS) and Support Vector Regression (SVR) as prediction techniques, Principal Component Analysis (PCA) as a dimensionality reduction technique and Self-Organizing Map (SOM) and Expectation Maximization (EM) as two well-known clustering techniques. To improve the recommendation accuracy of proposed multi-criteria CF, a cluster ensembles approach, Hypergraph Partitioning Algorithm (HGPA), is applied on SOM and EM clustering results.

Roopesh et al. studied that recommender systems is one of the most useful application of machine learning. They are collection of simple algorithms which tend to provide most relevant and accurate data as per user's requirement. Travel and Tourism domain is one of the important economic areas of a nation and recommender systems in this domain would cater to not only the tourists but also to the governments. This paper is a study of the various recommender systems available in the field of travel and tourism.

Srisawatsakul et al. developed the prototype of a tourism recommender system that automatically understands the user's preferences of their favourite tourist attractions without asking them any question. It applied machine learning to extract the user's preferences from the user's Instagram photos. Those preferences then use to compute the similarity with the attributes from 23 example tourist attractions in Ubon Ratchathani Province. A user study was conducted with 42 participates to preliminary study the precision and the adoption of the prototype.

Rula et al. provided a state-of-the-art e-tourism data management classification taxonomy based on smart concepts and reviews works in different fields against that classification. The retrieved articles were filtered according to the defined inclusion criteria. Finally, 65 articles were selected and classified into two categories. The first category included smart-based TRS that accounts for 87.70% (n = 57/65) and classified into four approaches: collaborative filtering, content model, context model and hybrid model. The second category includes tourism marketing that accounts for 12.30% (n = 8/65). The reliability and acceptability of smart-based TRS approach from the implemented 12 smart key concepts show a significant difference.

ISSN- 2394-5125 VOL 08, ISSUE 05, 2021

Banerjee et al. devoted on a live project implementation and testing of a learning model prototype in tourist information system and service industry. The elaborated model is followed by result sessions, which demonstrate that artificial agents could mimic the collective service and product pattern effectively compared to other contemporary techniques. The cost optimization to address the service issues in tourism industry could also be achieved with the help of such prototype models.

Ramzan et al. proposed an intelligent approach which also deals with large-sized heterogeneous data to fulfill the needs of the potential customers. The collaborative filtering (CF) approach is one of the most popular techniques of the RS to generate recommendations. This work proposed a novel CF recommendation approach in which opinion-based sentiment analysis is used to achieve hotel feature matrix by polarity identification. This approach combined lexical analysis, syntax analysis, and semantic analysis to understand sentiment towards hotel features and the profiling of guest type (solo, family, couple etc). The proposed system recommends hotels based on the hotel features and guest type for personalized recommendation. The developed system not only could handle heterogeneous data using big data Hadoop platform, but it also recommends hotel class based on guest type using fuzzy rules.

Zheng et al. proposed a feature set with 56 dimensions from the tourist's historical traveling data. By utilizing the entropy from information theory, all features are ranked. Evaluation results showed that when selecting the most important subset of 20 features as this final input of Random Forests, we can get a 4% higher accuracy and a 70% reduction of computation complexity regarding using the full set of features.

3. PROPOSED SYSTEM

The below Fig. 1 shows the proposed block diagram

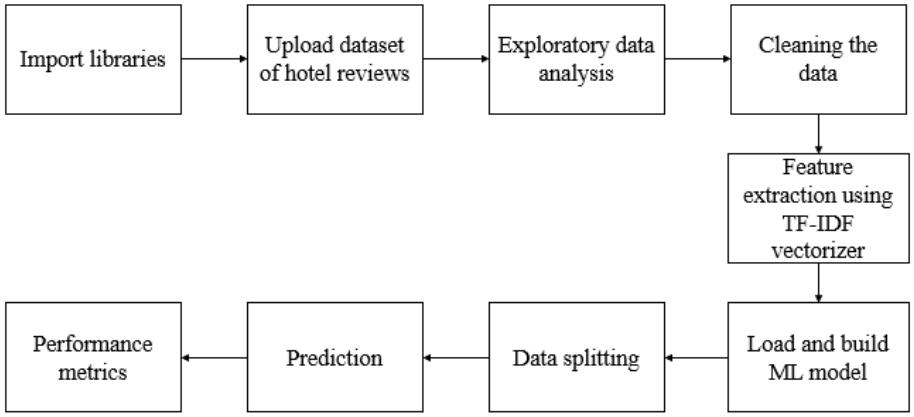


Fig. 1: Block diagram of proposed system.

3.1 Random Forest Algorithm

Random Forest is a popular machine learning algorithm that belongs to the supervised learning technique. It can be used for both Classification and Regression problems in ML. It is based on the concept of ensemble learning, which is a process of combining multiple classifiers to solve a complex problem and to improve the performance of the model. As the name suggests, "Random Forest is a classifier that contains a number of decision trees on various subsets of the given dataset and takes the average to improve the predictive accuracy of that dataset." Instead of relying on one decision tree, the random forest takes the prediction from each tree and based on the majority votes of predictions, and it predicts the final output. The greater number of trees in the forest leads to higher accuracy and prevents the problem of overfitting.

ISSN- 2394-5125 VOL 08, ISSUE 05, 2021

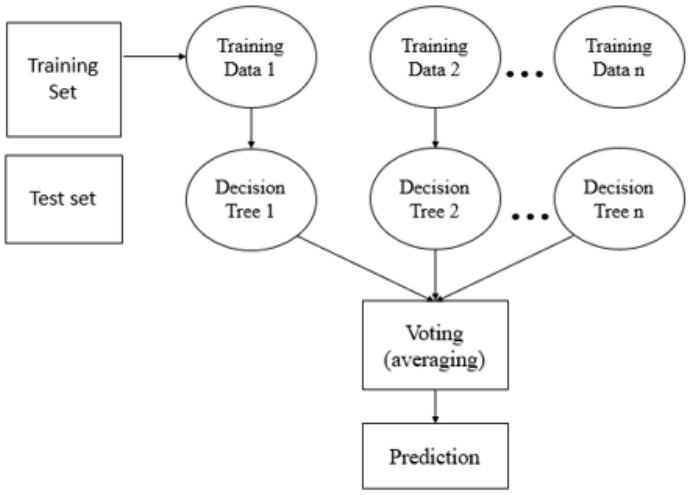


Fig. 2: Random Forest algorithm.

Random Forest algorithm

Step 1: In Random Forest n number of random records are taken from the data set having k number of records.

Step 2: Individual decision trees are constructed for each sample.

Step 3: Each decision tree will generate an output.

Step 4: Final output is considered based on Majority Voting or Averaging for Classification and regression respectively.

Important Features of Random Forest

- Diversity- Not all attributes/variables/features are considered while making an individual tree, each tree is different.
- Immune to the curse of dimensionality- Since each tree does not consider all the features, the feature space is reduced.
- Parallelization-Each tree is created independently out of different data and attributes. This means that we can make full use of the CPU to build random forests.
- Train-Test split- In a random forest we don't have to segregate the data for train and test as there will always be 30% of the data which is not seen by the decision tree.
- Stability- Stability arises because the result is based on majority voting/ averaging.

3.1.1 Assumptions for Random Forest

Since the random forest combines multiple trees to predict the class of the dataset, it is possible that some decision trees may predict the correct output, while others may not. But together, all the trees predict the correct output. Therefore, below are two assumptions for a better Random forest classifier:

- There should be some actual values in the feature variable of the dataset so that the classifier can predict accurate results rather than a guessed result.
- The predictions from each tree must have very low correlations.
- Below are some points that explain why we should use the Random Forest algorithm
- It takes less training time as compared to other algorithms.
- It predicts output with high accuracy, even for the large dataset it runs efficiently.
- It can also maintain accuracy when a large proportion of data is missing.

3.1.2 Types of Ensembles

Before understanding the working of the random forest, we must look into the ensemble technique. Ensemble simply means combining multiple models. Thus, a collection of models is used to make predictions rather than an individual model. Ensemble uses two types of methods:

Bagging– It creates a different training subset from sample training data with replacement & the final output is based on majority voting. For example, Random Forest. Bagging, also known as Bootstrap Aggregation, is the ensemble technique used by random forest. Bagging chooses a random sample from the data set. Hence each model is generated from the samples (Bootstrap Samples) provided by the Original Data with replacement known as row sampling. This step of row sampling with replacement is called bootstrap. Now each model is trained independently which generates results. The final output is based on majority voting after combining the results of all models. This step which involves combining all the results and generating output based on majority voting is known as aggregation.

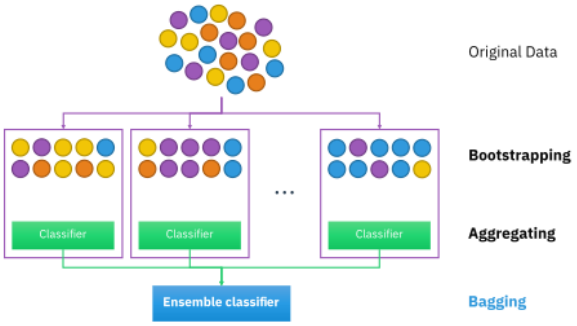


Fig. 3: RF Classifier analysis.

Boosting– It combines weak learners into strong learners by creating sequential models such that the final model has the highest accuracy. For example, ADA BOOST, XG BOOST.

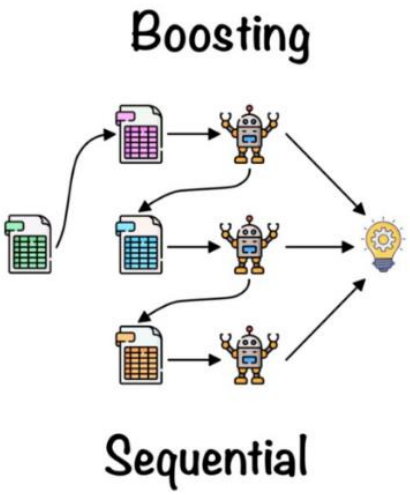


Fig. 4: Boosting RF Classifier.

3.1.3 Advantages of Random Forest

- It can be used in classification and regression problems.
- It solves the problem of overfitting as output is based on majority voting or averaging.

- It performs well even if the data contains null/missing values.
- Each decision tree created is independent of the other thus it shows the property of parallelization.
- It is highly stable as the average answers given by a large number of trees are taken.
- It maintains diversity as all the attributes are not considered while making each decision tree though it is not true in all cases.
- It is immune to the curse of dimensionality. Since each tree does not consider all the attributes, feature space is reduced.

3.1.4 Applications of Random Forest: There are mainly four sectors where Random Forest mostly used:

- Banking: Banking sector mostly uses this algorithm for the identification of loan risk.
- Medicine: With the help of this algorithm, disease trends and risks of the disease scan be identified.
- Land Use: We can identify the areas of similar land use by this algorithm.
- Marketing: Marketing trends can be identified using this algorithm.

3.2 TF-IDF

TF-IDF which stands for Term Frequency – Inverse Document Frequency. It is one of the most important techniques used for information retrieval to represent how important a specific word or phrase is to a given document. Let's take an example, we have a string or Bag of Words (BOW) and we have to extract information from it, then we can use this approach.

The tf-idf value increases in proportion to the number of times a word appears in the document but is often offset by the frequency of the word in the corpus, which helps to adjust with respect to the fact that some words appear more frequently in general. TF-IDF use two statistical methods, first is Term Frequency and the other is Inverse Document Frequency. Term frequency refers to the total number of times a given term t appears in the document doc against (per) the total number of all words in the document and The inverse document frequency measure of how much information the word provides. It measures the weight of a given word in the entire document. IDF show how common or rare a given word is across all documents. TF-IDF can be computed as tf * idf

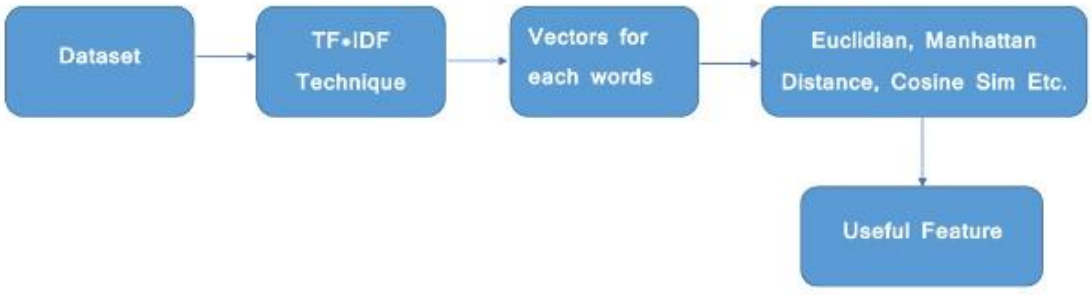


Fig. 5: TF-IDF block diagram.

TF-IDF do not convert directly raw data into useful features. Firstly, it converts raw strings or dataset into vectors and each word has its own vector. Then we'll use a particular technique for retrieving the feature like Cosine Similarity which works on vectors, etc.

Terminology

t — term (word)

d — document (set of words)

N — count of corpus

corpus — the total document set

Step 1: Term Frequency (TF): Suppose we have a set of English text documents and wish to rank which document is most relevant to the query, "Data Science is awesome!" A simple way to start out is by eliminating documents that do not contain all three words "Data" is", "Science", and "awesome", but this still leaves many documents. To further distinguish them, we might count the number of times each term occurs in each document; the number of times a term occurs in a document is called its term frequency. The weight of a term that occurs in a document is simply proportional to the term frequency.

$$tf(t,d) = count\ of\ t\ in\ d\ /\ number\ of\ words\ in\ d$$

Step 2: Document Frequency: This measures the importance of document in whole set of corpora, this is very similar to TF. The only difference is that TF is frequency counter for a term t in document d, whereas DF is the count of occurrences of term t in the document set N. In other words, DF is the number of documents in which the word is present. We consider one occurrence if the term consists in the document at least once, we do not need to know the number of times the term is present.

$$df(t) = occurrence\ of\ t\ in\ documents$$

Step 3: Inverse Document Frequency (IDF): While computing TF, all terms are considered equally important. However, it is known that certain terms, such as "is", "of", and "that", may appear a lot of times but have little importance. Thus, we need to weigh down the frequent terms while scale up the rare ones, by computing IDF, an inverse document frequency factor is incorporated which diminishes the weight of terms that occur very frequently in the document set and increases the weight of terms that occur rarely. The IDF is the inverse of the document frequency which measures the informativeness of term t. When we calculate IDF, it will be very low for the most occurring words such as stop words (because stop words such as "is" is present in almost all of the documents, and N/df will give a very low value to that word). This finally gives what we want, a relative weightage.

$$idf(t) = N/df$$

Now there are few other problems with the IDF, in case of a large corpus, say 100,000,000, the IDF value explodes, to avoid the effect we take the log of idf . During the query time, when a word which is not in vocab occurs, the df will be 0. As we cannot divide by 0, we smoothen the value by adding 1 to the denominator.

$$idf(t) = log(N/(df + 1))$$

The TF-IDF now is at the right measure to evaluate how important a word is to a document in a collection or corpus. Here are many different variations of TF-IDF but for now let us concentrate on this basic version.

$$tf - idf(t,d) = tf(t,d) * log(N/(df + 1))$$

Step 4: Implementing TF-IDF: To make TF-IDF from scratch in python, let's imagine those two sentences from different document:

first sentence: "Data Science is the sexiest job of the 21st century".

second sentence: "machine learning is the key for data science".

4. RESULTS AND DISCUSSION

Sample dataset

	Hotel_Address	Additional_Number_of_Scoring	Review_Date	Average_Score	Hotel_Name	Reviewer_Nationality	Negative_Review
0	s Gravesandestraat 55 Oost 1092 AA Amsterdam ...	194	8/3/2017	7.7	Hotel Arena	Russia	I am so angry that i made this post available...
1	s Gravesandestraat 55 Oost 1092 AA Amsterdam ...	194	8/3/2017	7.7	Hotel Arena	Ireland	No Negative
2	s Gravesandestraat 55 Oost 1092 AA Amsterdam ...	194	7/31/2017	7.7	Hotel Arena	Australia	Rooms are nice but for elderly a bit difficul...
3	s Gravesandestraat 55 Oost 1092 AA Amsterdam ...	194	7/31/2017	7.7	Hotel Arena	United Kingdom	My room was dirty and I was afraid to walk ba...
4	s Gravesandestraat 55 Oost 1092 AA Amsterdam ...	194	7/24/2017	7.7	Hotel Arena	New Zealand	You When I booked with your company on line y...

Review_Total_Negative_Word_Counts	Total_Number_of_Reviews	Positive_Review	Review_Total_Positive_Word_Counts	T
397	1403	Only the park outside of the hotel was beauti...	11	
0	1403	No real complaints the hotel was great great ...	105	
42	1403	Location was good and staff were ok It is cut...	21	
210	1403	Great location in nice surroundings the bar a...	26	
140	1403	Amazing location and building Romantic setting	8	

Total_Number_of_Reviews_Reviewer_Has_Given	Reviewer_Score	Tags	days_since_review	lat	lng
7	2.9	[' Leisure trip ', ' Couple ', ' Duplex Double...	0 days	52.360576	4.915968
7	7.5	[' Leisure trip ', ' Couple ', ' Duplex Double...	0 days	52.360576	4.915968
9	7.1	[' Leisure trip ', ' Family with young childre...	3 days	52.360576	4.915968
1	3.8	[' Leisure trip ', ' Solo traveler ', ' Duplex...	3 days	52.360576	4.915968
3	6.7	[' Leisure trip ', ' Couple ', ' Suite ', ' St...	10 days	52.360576	4.915968

Exploratory data analysis

In [7]:
```
reviews_df = reviews_df[["review", "is_bad_review"]]
reviews_df.head()
```

Out[7]:

	review	is_bad_review
0	I am so angry that i made this post available...	1
1	No Negative No real complaints the hotel was g...	0
2	Rooms are nice but for elderly a bit difficul...	0
3	My room was dirty and I was afraid to walk ba...	1
4	You When I booked with your company on line y...	0

	review	is_bad_review
0	Nothing Everything The location and ease of ...	0
1	No Negative Very spacious and clean rooms Plea...	0
2	No Negative Nice central location lovely moder...	0
3	No Negative Made my mom s birthday special by ...	0
4	Small rooms Great location	0
...	...	...
9995	It s far away from the city centre Hotel look...	1
9996	Everything No what was advertised Nothing	1
9997	Rooms are roach infested you will see green r...	1
9998	Too far off the beaten track too much buildin...	1
9999	I believe the problem started with Booking co...	1

10000 rows × 2 columns

Data preprocessing

```
0       Nothing  Everything The location and ease of ...
1       Very spacious and clean rooms Pleasant person...
2       Nice central location lovely modern rooms Ove...
3       Made my mom s birthday special by arranging f...
4                           Small rooms Great location
                              ...
9995    It s far away from the city centre Hotel look...
9996          Everything No what was advertised Nothing
9997    Rooms are roach infested you will see green r...
9998    Too far off the beaten track too much buildin...
9999    I believe the problem started with Booking co...
Name: review, Length: 10000, dtype: object
```

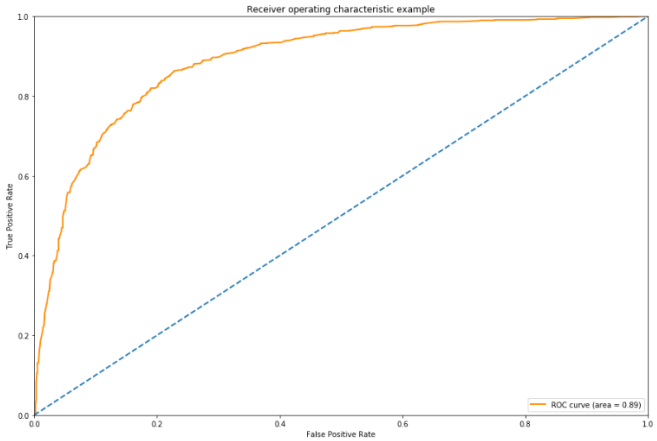
```
0       nothing everything location ease travel around...
1       spacious clean room pleasant personnel notice ...
2       nice central location lovely modern room overa...
3         make mom birthday special arrange special treat
4                           small room great location
                              ...
9995    far away city centre hotel look like quiete ol...
9996                        everything advertise nothing
9997    room roach infest see green roach fly nite bed...
9998    far beaten track much building work go room sm...
9999    believe problem start book com inform hotel up...
Name: review_clean, Length: 10000, dtype: object
```

Prediction results

```
Confusion Matrix :
[[803 209]
 [160 828]]
Accuracy Score : 0.8155
Report :
              precision    recall  f1-score   support

           0       0.83      0.79      0.81      1012
           1       0.80      0.84      0.82       988

    accuracy                           0.82      2000
   macro avg       0.82      0.82      0.82      2000
weighted avg       0.82      0.82      0.82      2000
```



5. CONCLUSION AND FUTURE SCOPE

This paper presented a novel human centric TRS that recommends destinations to tourists in an unfamiliar city. It considered both technical and practical aspects using a real-world data set we collected. The system is developed using two-steps feature selection method to reduce number of inputs to the system and recommendations are provided by recommendation system. The experimental results showed that the proposed TRS can provide personalized recommendation on tourist destinations that satisfy the tourists. In the future, it is planned to work on a real-time dataset and extend the system to be international containing hotels from all around the world.

REFERENCES

[1] Ministry of Tourism and Sports of Thailand tourism industry in Thailand, “The situation of Tourism Industry in Thailand.,” 2020.

[2] G. Adomavicius and A. Tuzhilin, “Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions,” IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 6, pp. 734–749, 2005.

[3] F. Ricci, “Travel recommender systems,” IEEE Intelligent Systems, vol. 17, pp. 55–57, 2002.

[4] V. Thannimalai and L. Zhang, "A Content Based and Collaborative Filtering Recommender System," 2021 International Conference on Machine Learning and Cybernetics (ICMLC), 2021, pp. 1-7, doi: 10.1109/ICMLC54886.2021.9737238.

[5] Nilashi, Mehrbakhsh & Fard, Karamollah & Rahmani, Mohsen & Rafe, Vahid. (2017). A Recommender System for Tourism Industry Using Cluster Ensemble and Prediction Machine Learning Techniques. Computers & Industrial Engineering. 109. 10.1016/j.cie.2017.05.016.

[6] L R, Roopesh & Bomatpalli, Tulasi. (2019). A Survey of Travel Recommender System. 10.13140/RG.2.2.34775.32168.

[7] C. Srisawatsakul and W. Boontarig, "Tourism Recommender System using Machine Learning Based on User's Public Instagram Photos," 2020 - 5th International Conference on Information Technology (InCIT), 2020, pp. 276-281, doi: 10.1109/InCIT50588.2020.9310777.

[8] Rula A. Hamid, A.S. Albahri, Jwan K. Alwan, Z.T. Al-qaysi, O.S. Albahri, A.A. Zaidan, Alhamzah Alnoor, A.H. Alamoodi, B.B. Zaidan, “How smart is e-tourism? A systematic review of smart tourism recommendation system applying data management”, Computer Science Review, Volume 39, 2021, 100337, ISSN 1574-0137.

[9] Banerjee, S., Chis, M., Dangayach, G.S. (2010). Developing an Adaptive Learning Based Tourism Information System Using Ant Colony Metaphor. In: Xhafa, F., Caballé, S., Abraham, A., Daradoumis, T., Juan Perez, A.A. (eds) Computational Intelligence for Technology Enhanced Learning. Studies in Computational Intelligence, vol 273. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-11224-9_3.

[10] B. Ramzan, I. S. Bajwa, N. Jamil, R. U. Amin, S. Ramzan, F. Mirza, N. Sarwar, "An Intelligent Data Analysis for Recommendation Systems Using Machine Learning", Scientific Programming, vol. 2019, Article ID 5941096, 20 pages, 2019. https://doi.org/10.1155/2019/5941096.

[11] S. Zheng, Y. Liu and Z. Ouyang, "A machine learning-based tourist path prediction," 2016 4th International Conference on Cloud Computing and Intelligence Systems (CCIS), 2016, pp. 38-42, doi: 10.1109/CCIS.2016.7790221.