Customized Tourist Destination Recommendation System

WonHong Jeong
2020320089
Korea University
thetwo0525@korea.ac.kr

JiWon Park
2021160043
Korea University
studypjw0104@naver.com

HyeMin Woo
2021340035
Korea University
woohm3@naver.com

YunJae Choi
2022320303
Korea University
jyunchoi0710@naver.com

ABSTRACT

This study presents a personalized tourist destination recommendation system driven by travel log data and structured through a multi-stage analytical pipeline. First, we pre-process the data by filtering irrelevant location types and integrating scattered tables. Based on demographic factors such as gender and age, users are clustered by travel styles—including companion type, destination region, trip duration, season, and purpose—using unsupervised learning. Within each cluster, frequent pattern mining reveals co-visited location pairs, forming the basis for candidate recommendations. Finally, a classification model predicts the satisfaction level of users toward each candidate destination, enabling ranked and refined suggestions based on expected user satisfaction. By aligning each analytical stage into a cohesive framework, this system aims to deliver personalized, data-driven recommendations that are both accurate and context-aware.

1 INTRODUCTION

Tourism is evolving rapidly toward personalized and data-driven experiences. Yet, most existing recommendation systems rely heavily on generic metrics such as popularity or review scores, offering the same list of destinations to all users regardless of their personal preferences or travel context. This one-size-fits-all approach fails to address the diverse needs of individual travelers.

This issue is especially critical in local tourism, where effective engagement depends on accurately understanding visitor behavior. To make destination recommendations truly relevant, the system must account for who the user is, where they've been, and what they've enjoyed—information that is typically missing in conventional systems.

In response, our project aims to design a personalized tourist destination and accommodation recommendation system based on real-world travel log data. The dataset, provided by AI Hub Korea, includes detailed demographic information, travel style, location visit patterns, satisfaction scores, and revisit intentions from over 12,000 travelers across Korea.

The significance of this work lies in its potential to offer travelers more meaningful suggestions while simultaneously contributing to regional tourism development. By leveraging behavioral data and applying clustering, pattern mining, and satisfaction prediction techniques, our system seeks to bridge the gap between user individuality and recommendation accuracy.

2 RELATED WORKS

2.1 Hodu

your content here.

2.2 Happiness

your content here.

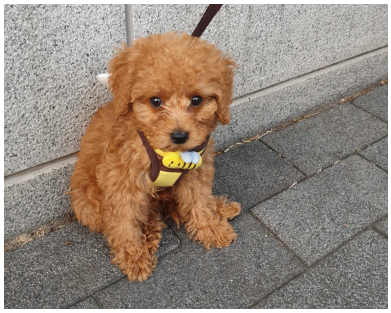

Figure 1: Very cute puddle

3 METHODOLOGY

Our recommendation system is designed as a multi-stage analytical pipeline that integrates data preprocessing, user clustering, pattern mining, and satisfaction prediction.

3.1 Data Preprocessing

To prepare the dataset for clustering and recommendation, we performed extensive preprocessing on raw travel log data. Multiple source tables containing traveler information, trip records, visited locations, and companion types were merged into a unified format. Only essential columns were retained to remove noise and irrelevant details.

Data Science Team 19, Korea University,

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Several derived features were added to capture trip-level characteristics, such as total travel duration and the dominant month of travel when a trip spanned multiple months. We also categorized companion types into broader groups (e.g., solo, family, friends, others) to simplify behavioral analysis.

To ensure data quality, we applied strict filtering. Only records associated with location types relevant to tourism were preserved, while entries involving residential, transit, or dining facilities were excluded. Additionally, records with abnormal durations or missing satisfaction-related variables were removed.

Outliers were detected and excluded using the interquartile range (IQR) method, based on users' average satisfaction and intention scores. The cleaned and enriched dataset was finally partitioned by region and saved for downstream analysis.

3.2 User Clustering

To group users with similar travel preferences, we applied clustering on the preprocessed dataset. Users were first grouped into subpopulations based on gender and age. Within each subgroup, clustering was performed using selected travel behavior features, such as travel style and companion type.

We adopted a stratified clustering strategy, generating three distinct behavioral clusters within each gender-age subgroup. This approach resulted in a total of 30 clusters across the entire user base. The choice of three clusters per group was empirically supported through visual inspection using principal component analysis (PCA), which showed well-separated boundaries at this setting.

Each user record was annotated with a cluster label, which served as the foundation for downstream pattern mining and satisfaction prediction. This approach allowed us to tailor recommendations to the behavioral tendencies of each user segment.

3.3 Frequent Pattern Mining

To extract destination recommendation rules tailored to each user group, we applied the FP-Growth algorithm within each cluster. This method was chosen over Apriori due to its superior computational efficiency, especially when applying multiple support and confidence filters.

Prior to pattern mining, we filtered out locations that appeared fewer than 30 times in the dataset to eliminate sparse or outlier regions. Each transaction was constructed as a 2-item set containing a cluster identifier and a visited area name, representing the co-occurrence of a cluster group and a specific destination.

We set a minimum support threshold of 0.8% and a confidence threshold of 0.35. These values were selected to ensure statistical significance while filtering out weak or infrequent association rules. The resulting rules followed the form: "Users in Cluster X often visited Destination Y with a confidence of Z%."

Example rules include:

- Users in Cluster_0 who visited *Everland* had a 38.6% probability of also visiting *National Museum of Korea*.
- Users in Cluster_2 who visited *Gyeongbokgung Palace* had a 46.9% probability of also visiting *Caribbean Bay*.

While the initial expectation was to discover stronger patterns with high confidence values (e.g., over 60%), most actual confidence values remained below 0.5. This indicates that although some weak tendencies exist, strongly polarized travel patterns were rare across clusters.

To visualize and interpret the discovered rules effectively, we utilized NetworkX to map co-visit patterns graphically. These visualizations proved more intuitive than raw textual tables, especially for observing cluster-to-location relationships at scale.

As a future direction, we plan to refine cluster segmentation or relax filter thresholds to discover more actionable association rules. Moreover, incorporating additional attributes (e.g., travel purpose or duration) may enable multidimensional rule mining with greater predictive power.

3.4 Satisfaction Prediction

To estimate how likely a user is to be satisfied with each recommended location, we formulate a binary classification task. A new binary label is defined where a satisfaction score (`DGSTFN`) of 4.0 or higher is considered positive (1), and lower scores are negative (0). Additional features include the name of the visited area (`VISIT_AREA_NM`), the month of the visit extracted from the date, and user feedback indicators such as revisit and recommendation intentions.

We preprocess the categorical feature (`VISIT_AREA_NM`) using one-hot encoding, and build a pipeline that combines feature transformation with a logistic regression classifier. The model is trained on a stratified 70/30 train-test split, and class weights are balanced to address label imbalance.

- **Model:** Logistic Regression (max_iter = 1000, class_weight = 'balanced')
- **Features:** Visit Area, Visit Month, Revisit Intention, Recommendation Intention
- **Target:** Binary satisfaction label (1 if `DGSTFN` ≥ 4.0)

The model is evaluated using accuracy, confusion matrix, and a full classification report including precision, recall, and F1-score. The results show the model's ability to distinguish between likely and unlikely satisfaction, which supports the ranking and filtering of recommendations.

!!!!!!!!Formula instruction cf!!!!!!!!!! We can define Hodu's happiness level as a function of snack count $H(s) = \log(s + 1)$.
To prevent overfeeding, we use a capped scoring model:

$$H(s) = \begin{cases} \log(s+1), & \text{if } s \leq 5 \\ \log(6) - \frac{1}{2}(s-5), & \text{if } s > 5 \end{cases} \tag{1}$$

This ensures that after five snacks, Hodu's happiness increase slows down — mimicking diminishing returns.

This log-based modeling approach is inspired by earlier work on attention and saturation dynamics [1].

REFERENCES

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems* 30 (2017).

4 EXPERIMENTS

your content here.

(a) Before a walk

(b) After a walk

Figure 2: Comparison of emotional well-being

Table 1: Table cf

Model	Hodu		Maru	
	Reaction	Well-being	Reaction	Well-being
Baseline1	0.4224	0.5757	0.5621	0.5932
Baseline2	0.2324	0.3789	0.2624	0.3996
Baseline3	0.4321	0.5678	0.4421	0.5987
YOURS	**0.9923**	**0.7123**	**0.9942**	**0.7271**
-w/o Snack	0.5642	0.6998	0.5830	0.7192
-w/o Walk	0.9877	0.7012	0.9922	0.7188

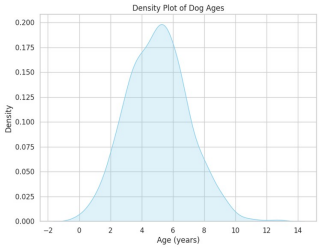

Figure 3: Enter Caption

5 CONCLUSION

Your content here.