사용자 맞춤형 관광지 추천 시스템

Customized Tourist Destination Recommendation System

정원홍
2020320089
Korea University
thetwo0525@korea.ac.kr

박지원
2021160043
Korea University
studypjw0104@naver.com

우혜민
2021340035
Korea University
woohm3@naver.com

최윤재
2022320303
Korea University
jyunchoi0710@naver.com

ABSTRACT

본 연구는 여행 로그 데이터를 기반으로 한 개인 맞춤형 관광지 추천 시스템을 제안한다. 제안된 시스템은 다단계 분석 파이프라인을 중심으로 구되며, 각 단계는 연속적이고 통합된 방식으로 연결된다.

우선, 관련성이 낮은 장소 유형을 필터링하고, 분산된 다중 테이블을 통합하여 데이터 전처리를 수행한다. 사용자별 평균 만족도, 재방문 의도, 추천 의도 점수에 대해 사분위수 범위(IQR)를 기반으로 이상치를 탐지하고, 해당 범위를 벗어나는 극단값을 포함하는 기록은 분석의 정확도를 저해할 수 있으므로 제거하여 보다 정제된 데이터셋을 확보한다.

이후, 성별, 연령 등 인구통계학적 특성과 더불어 다양한 여행 스타일을 기준으로 비지도 학습 기반의 사용자 클러스터링을 수행한다. 사용자 클러스터링 이후에는 각 그룹의 대표적인 여행 성향을 분석하고, 이를 기반으로 해석 가능한 사용자 프로파일을 도출한다. 특히 성별 및 연령대에 따른 여행 스타일의 차이를 클러스터링 결과를 통해 확인하며, 이는 추천의 정밀도를 향상시키는 기초 자료로 활용된다.

다음 단계에서는 각 클러스터 내에서 빈발 패턴 마이닝을 적용하여 자주 함께 방문되는 장소 트랜잭션을 도출하고, 이를 바탕으로 추천 후보지를 선정한다.

마지막으로, 분류 모델을 활용하여 각 후보지에 대한 사용자의 만족도를 예측하고, 예측된 만족도에 따라 추천 대상을 정렬 및 정제함으로써 최종 추천 결과를 도출한다.

본 시스템은 각 분석 단계를 통합적으로 구성함으로써, 사용자 맞춤형이며 맥락을 반영한 관광지 추천의 가능성을 탐색하고자 하였다. 시스템 설계 과정을 통해, 사용자 특성과 행동 양식을 반영한 맞춤형 추천의 가능성과 그 한계를 함께 조명하였다. 서비스화 실질적인 성능 향상보다는, 추천 시스템의 구조와 데이터 활용 방식에 대한 탐색적 기여에 의의를 둔다. 향후 추가적인 데이터 보완과 방법론 개선을 통해 실질적인 추천 정확도를 높일 수 있을 것으로 기대된다.

Data Science Team ***잘 묵겠습니다****, June 2025, Korea University*

1 INTRODUCTION

1.1 Motivation and Problem Setting

오늘날 관광 산업은 개인화되고 데이터 기반의 경험을 중심으로 빠르게 진화하고 있다. 그러나 기존의 대부분의 추천 시스템은 여전히 인기순, 리뷰 점수와 같은 일반적인 지표에 의존하고 있으며(Figure 1), 이러한 시스템은 사용자 개인의 선호나 여행 맥락을 고려하지 않고 모든 이용자에게 동일한 목적지 목록을 제시한다. 이와 같은 획일화된 접근 방식은 개별 여행자의 다양한 요구를 충족시키지 못하는 한계를 지닌다.

이는 지역 관광 분야에서도 문제로 인식될 수 있다. 지역 관광에서는 방문객의 행동을 정확하게 이해하는 것이 효과적인 추천과 참여를 이끌어내는 핵심이기 때문이다. 여행 목적지를 보다 적절하게 추천하기 위해서는 사용자가 누구인지, 어디를 다녀왔는지, 무엇을 즐겼는지에 대한 정보가 반드시 고려되어야 하지만, 이러한 정보는 대부분의 기존 시스템에서는 반영되지 않는다.

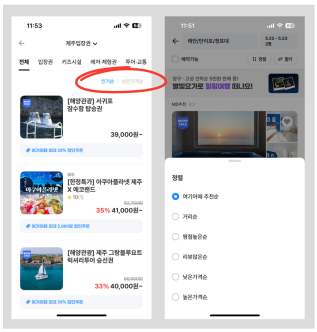

Figure 1: '여기어때' 앱 UI

1.2 Goal

이러한 문제의식에 기반하여, 본 프로젝트는 실제 여행 로그 데이터를 바탕으로 개인 맞춤형 관광지 추천 시스템을 설계하는 것을 목표로 한다. 본 연구에서 활용한 데이터셋은 성별과 연령을 포함한 인구통계 정보, 자연/도시, 새로운 장소/익숙한 장소, 인기 지역/한적한 지역, 선호 지역, 여행 목적(휴식 또는 체험), 여행 성향(계획형 또는 즉흥형), 사진 촬영의 중요도와 같은 여행 스타일 설문 응답, 방문 장소, 만족도 및 재방문 의사 점수 등을 포함한 12,000명 이상의 국내 여행자에 대한 정보를 담고 있다.

1.3 Significance

본 연구 설계 목표의 의의는 개별 여행자에게 보다 의미 있는 추천을 제공함과 동시에 지역 관광 활성화에 기여할 수 있는 가능성에 있다. 본 시스템은 행동 기반 데이터를 활용하고, 클러스터링, 패턴 마이닝, 만족도 예측 기법을 통합함으로써, 사용자 개별성과 추천 정확도 간의 간극을 해소하고자 한다.

2 DATASET

본 프로젝트에서는 AI Hub Korea에서 제공한 실제 국내 여행 로그 데이터를 활용하였다. 원본 데이터는 여행객 정보, 여행 기록, 방문지 정보, 동행자 정보로 구성된 총 네 개의 테이블로 분산되어 있었으며, 이를 분석 목적에 맞게 하나의 통합 테이블로 구성하였다.

이 과정에서 다음과 같은 주요 컬럼을 선별하였다.

- **여행객 정보**: 여행자 고유 ID, 성별 (GENDER), 연령대 (AGE_GRP), 여행 스타일 관련 항목:
  - 자연/도시: TRAVEL_STYL_1
  - 익숙한 지역/낯선 지역: TRAVEL_STYL_3
  - 인기 관광지/한적한 관광지: TRAVEL_STYL_5
  - 휴양 여행/체험활동 위주 여행: TRAVEL_STYL_6
  - 계획적 여행/즉흥적 여행: TRAVEL_STYL_7
  - 사진 촬영 중요도: TRAVEL_STYL_8

  동행 여부 (TRAVEL_STATUS_ACCOMPANY)
- **여행 기록**: 여행 고유 ID, 여행 시작일 (TRAVEL_START_YMD), 여행 종료일 (TRAVEL_END_YMD), 이동 수단 (MVMN_NM), 총 여행일 수 (TRAVEL_DAYS), 여행 시기 (MAIN_TRAVEL_MONTH)
- **방문지 정보**: 방문지명 (VISIT_AREA_NM), 방문지 유형 (VISIT_AREA_TYPE_CD), 방문지 만족도 (DGSTFN), 재방문 의사 (REVISIT_INTENTION), 추천 의사 (RCMDTN_INTENTION)

데이터 품질 확보를 위해 다단계 전처리를 수행하였다. 먼저, 방문지 유형 코드가 관광과 직접적으로 관련된 값인 1, 2, 3, 4, 5, 6, 7, 8, 12, 13에 해당하는 레코드만을 유지하고, 주거지, 교통시설, 식당 등 일상적 장소는 제외하였다. 또한 여행 기간이 비정상적으로 짧거나 긴 경우, 만족도나 재방문/추천 의향과 같은 핵심 변수들이 누락된 경우 역시 제거하였다.

이와 함께, 사용자별 평균 만족도 (DGSTFN), 재방문 의사 (REVISIT_INTENTION), 추천 의사 (RCMDTN_INTENTION) 점수를 기준으로 사분위수 범위(IQR)를 활용한 이상치 탐지를 적용하였다. 각 변수별 IQR을 계산하여 $[Q1 - 1.5 \times IQR, Q3 + 1.5 \times IQR]$ 범위를 벗어나는 사용자에 대해서는 극단값으로 판단하고 전체 레코드를 제거함으로써, 분석의 신뢰도를 높였다.

최종적으로 정제된 데이터셋은 분석의 효율성을 높이기 위해 권역 단위(수도권, 서부권, 동부권, 제주권)로 분할 저장하였으며, 이후 단계에서 클러스터링, 패턴 마이닝, 분류 모델 입력값으로 사용되었다.

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 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 preferred destinations.

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.

3.2 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.

3.3 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.

4 EXPERIMENTS AND EVALUATION

your content here.

5 DISCUSSION AND CONCLUSION

5.1 Insight

5.2 Feedback

5.3 Limitations

5.4 Meaning

6 ROLES AND RESPONSIBILITIES

정원홍 프로젝트 구조 기획 및 발표 자료 제작, Pattern Mining, 실험 및 결과 분석, 최종 발표

박지원 프로젝트 구조 기획 및 발표 자료 제작, 데이터 전처리, Clustering, 실험 및 결과 분석, 최종 발표

최윤재 프로젝트 구조 기획 및 발표 자료 제작, Classification, 실험 및 결과 분석, 최종 발표

우혜민 프로젝트 구조 기획 및 발표 자료 제작, Introduction, Discussion and Conclusion 작성, 프로젝트 워크로드 조정, LaTeX 문서 작업, 중간/최종 발표

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:

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

Table 1: Table ef

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

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).

(a) Before a walk

(b) After a walk

Figure 2: Comparison of emotional well-being