Tourist Spot Recommendation by Clustering Using an SDP-Based Max-Cut

Yuichi Tanigawa
Graduate School of Data Science
Musashino University
Tokyo, Japan
g2450003@stu.musashino-u.ac.jp

Virach Sornlertlamvanich
Faculty of Data Science / Faculty of Engineering
Musashino University / Thammasat University
Tokyo, Japan / Pathum Thani, Thailand
virach@gmail.com

Thatsanee Charoenporn
Faculty of Data Science
Musashino University
Tokyo, Japan
thatsanee@ds.musashino-u.ac.jp

***Abstract*—Collaborative filtering, a widely used recommended algorithm, prioritizes the generation of highly similar item lists. However, this approach specializes in aggregating similar items, which poses challenges in ensuring diversity, user autonomy, and explainability. Based on the opposite idea of conventional recommendation methods, we propose a clustering-based recommendation approach that intentionally separates highly similar items to ensure diversity. A Bag-of-Words corpus was constructed by collecting 600 English reviews from 60 major tourist attractions in Kyoto and applying noise reduction techniques. Using the Latent Dirichlet Allocation Method (LDA), each tourist spot was transformed into a 15-dimensional vector and a similarity graph was created based on cosine similarity. Max-Cut was then used to maximize dissimilarity between clusters by placing nodes with higher edge weights in different partitions to ensure diverse group formation. Experimental results demonstrate that the SDP-based Max-Cut method successfully generates clusters with high variance and outperforms comparable methods using traditional k-means clustering in terms of intra-cluster diversity. In addition, we analyze the impact of partition granularity on the recommended system and investigate its applicability to a wider range of areas beyond the tourism sector.**

Keywords—Max-Cut algorithm, clustering, user autonomy, explainable recommendation, diversity

I. Introduction

Tourist destination recommendation systems traditionally use similarity-based collaborative filtering (CF). This recommends items based on the choices of users with similar interests or past preferences. CF-based systems are widely adopted in real-world applications due to their relative accuracy and scalability, but often have certain limitations. First, there may be a lack of intuitive explanations that clarify why a particular item is recommended. Second, presenting a final list of "predicted best options" without providing transparent reasoning tends to reduce user autonomy. Third, focusing on items that are already similar to those selected by the user or similar users may narrow the diversity of recommendations [1]. These challenges can lead to a "filter bubble" and limited coverage of the full range of user preferences.

Motivated by the need to balance diversity, explainability, and user engagement, we propose a heterogeneity-driven clustering framework. Rather than grouping only similar items, as is common in traditional clustering and CF-based approaches, we use a Max-Cut methodology that explicitly targets the separation of items (sights) with very similar potential semantic characteristics. This creates diverse partitions where each cluster covers a more diverse set of spots. This separation allows the recommendation system to offer tours and suggestions that cover a broader range of user interests.

The core technology used is the SDP-based Max-Cut algorithm. This algorithm guarantees an approximation ratio of 0.878 for the Max-Cut problem through semi-positive definite programming and randomization [2]. Unlike previous approaches that use a distance metric as an edge weight to group similar items, our approach utilizes similarity (cosine similarity) as an edge weight in the Max-Cut formulation. Because Max-Cut is designed to partition the graph to maximize the sum of cut edges, items with high similarity are intentionally placed in different clusters, increasing cluster diversity. To evaluate how well Max-Cut-based clustering promotes diversity, we compared it to k-means clustering across various numbers of partitions.

This paper is organized as follows. Methods describes the data collection, preprocessing, topic modeling, and vectorization steps, and finally the Max-Cut based clustering approach that uses cosine similarity as edge weights. Results report the hierarchical partitions identified by Max-Cut, show the average similarity within clusters, and compare them to k-means. In Discussion, we highlight the benefits, potential challenges, and explain how to leverage partitions in a meta approach to recommendations. We also discuss how such diversity-oriented groupings can be applied not only to tourist destinations, but also to staffing in schools, businesses, or any community setting where balanced diversity is desirable. Finally, Conclusion summarizes the contributions and future research directions.

II. Methods

A. Data Collection

We collected data from TripAdvisor, one of the world's most frequently used travel review websites, which provides a large volume of user-generated content reflecting real consumer perspectives [3]. Focusing on Kyoto City—a global tourist destination that attracted approximately 50 million visitors in 2023—we aimed to reduce external factors such as regional differences by restricting our scope to Kyoto's central area [4], [5].

From 756 potential spots, we selected the top 100 based on TripAdvisor's "Traveler Ranking" and filtered down to 60 spots that had at least 10 English reviews. We then acquired the 10 most recent English reviews for each of these 60 spots, yielding 600 reviews. Our dataset included attributes such as `location_id`, `location_name`, `review_id`, and the `review_text`.

979-8-3315-2040-3/25/$31.00 ©2025 IEEE

B. Text Processing

To preprocess the collected review texts, we employed Python's NLTK library with its "stopwords," "wordnet," and "vader_lexicon" components. We converted text to lowercase, removed non-alphabetic characters using regular expressions, filtered out English stopwords, and applied WordNetLemmatizer to unify word forms. This produced a tokenized list for each review. Subsequently, we created a Bag-of-Words (BoW) corpus by constructing a Gensim Dictionary and transforming each document into the BoW representation, an appropriate format for later topic modeling [6].

C. Topic Modeling and Latent Determination of Topic Number

We applied the Latent Dirichlet Allocation Method (LDA) to the BoW corpus, focusing on its ability to increase both intelligibility and robustness. Clearness is very important because reviews of attractions often encompass multiple aspects such as history, food, and cultural experiences, and LDA is particularly effective at capturing these potential topics [7]. Ruggedness is another advantage, as LDA is a probabilistic generative model that reduces over-fitting, maintains validity as new documents and attractions are introduced, and outperforms methods such as pLSI in dynamic settings [8].

To determine the optimal number of topics, we calculated perplexity and coherence scores. Perplexity measures how well the model predicts unseen data, while coherence captures human interpretability of the extracted topics [9][10]. We found that 15 topics provided a good trade-off between interpretability and predictive accuracy.

D. Vectorization of Topic Distributions for Each Tourist Spot

LDA posits that each document d is represented by a topic distribution θd, which is drawn from a Dirichlet distribution. Each topic k is associated with a word distribution ϕk. The probability of each word in a document is thus governed by these distributions [8]. Since each tourist spot contained multiple (10) reviews, we aggregated the topic distributions for each spot to obtain a single 15-dimensional vector. Concretely, let $\{\theta d1, \theta d2, \ldots, \theta dm\}$ be the topic vectors for a spot's mm reviews. The representative topic vector $\bar{\theta}location$ is:

$$\bar{\theta}location = \frac{\theta d1 + \theta d2 + \cdots + \theta dm}{m} \quad (1)$$

In our dataset, m=10 for all tourist spots.

E. Graph Construction and SDP-Based Maximum Cut with Cosine Similarity

1) Graph Construction

After obtaining the 15-dimensional vectors for each of the 60 locations, we constructed a weighted graph $G=(V,E)$, where each tourist spot is a node in V. Instead of using the distance between vectors as the weight (as done in some earlier studies), we defined the edge weight wij between nodes i and j to be the cosine similarity of their topic vectors:

$$wij = \cos(\bar{\theta}i, \bar{\theta}j) \quad (2)$$

Cosine similarity ranges from 0 to 1 for nonnegative vectors; higher values indicate more closely related topics. Our key insight is to leverage Max-Cut in a way that assigns items with high edge weights to opposite sides of the partition, thereby encouraging diversity within each subset.

2) SDP-Based Maximum Cut

The Max-Cut problem seeks a partition of the vertex set VV into two disjoint subsets A and B such that the sum of the weights of the edges crossing between A and B is maximized [2]. Formally, if S⊆V, then the cut value is:

$$\sum_{i \in S, j \in V \setminus S} wij \quad (3)$$

By using the Goemans-Williamson semidefinite programming approach, we can achieve a 0.878 approximation ratio for Max-Cut [2]. The randomization step from the SDP solution ensures a partition with a cut value guaranteed to be near-optimal in expectation.

Whereas conventional clustering might group nodes with high similarity together, our Max-Cut approach explicitly places nodes with high similarity in different groups. The rationale is that each cluster thus becomes more internally diverse in terms of textual content or thematic emphasis. We further extended this approach by applying it hierarchically, repeatedly performing Max-Cut on each of the newly formed subgroups to yield additional layers of diversity.

F. Diversity Evaluation: Comparing Intra-Cluster Cosine Similarity

To quantify the diversity of the resulting clusters, we calculated the average intra-cluster cosine similarity for each group formed by Max-Cut. A lower average intra-cluster similarity implies higher internal variance and thus, more diverse grouping. We compared the Max-Cut partitions with those generated by k-means clustering (with identical numbers of clusters) across multiple divisions (e.g., 2, 4, and 8). In practice, we computed:

$$AvgSim(G) = \frac{1}{|G|} \sum_{i \in G} \left(\frac{1}{|G| - 1} \sum_{\substack{j \in G \\ j \neq 1}} \cos(\theta i, \theta j) \right)$$

where G denotes a specific cluster, and $\bar{\theta}_i$ is the topic vector of spot i. We also reported the overall average for all 60 spots combined for reference.

These measures allow us to see how strongly the Max-Cut-based partition strategy decreases within-group similarity, thereby potentially expanding the variety of spots a user is recommended.

III. Results

A. LDA Findings

The LDA model with 15 topics produced interpretable themes. Topics often pertained to cultural/historical elements, food or beverage references, experiential classes (e.g., cooking or crafts), shrines and temples, and modern entertainment. We visualized each topic using word clouds as shown in Fig. 1, identifying high-frequency words such as "temple," "tour," "class," "traditional," and "cultural." These diverse latent topics reflect the multifaceted nature of tourist experiences in Kyoto.

B. Max-Cut Clustering Results and Diversity

1) Overall Observations

Applying the SDP-based Max-Cut algorithm to split the graph into two larger groups resulted in clusters of 32 nodes (Group A) and 28 nodes (Group B). It is noteworthy that the average intra-cluster similarity for Group A (0.3638) and Group B (0.4124) is considerably lower than grouping spots based purely on similarity. This indicates that the Max-Cut approach effectively increases clustering diversity.

To further refine the partition, Max-Cut was applied recursively within each group to create subgroups such as GroupA1 and GroupA2. This hierarchical partitioning allowed us to generate finer clusters while emphasizing diversity. Table I shows the average similarity of each cluster. For example, in one iteration, GroupA1, which included 16 locations, had an average similarity of 0.3287, while GroupA2, which also included 16 locations, had an average similarity of 0.3593. Each subgroup was further divided, for example GroupA1 was divided into GroupA1α and GroupA1b, with an average similarity of 0.2915 and 0.2900, respectively.

Detailed analysis of the clustering results revealed that Max-Cut consistently formed diverse groups across multiple hierarchical levels. First, the overall average similarity across all 60 destinations was 0.40. When the graph was split into two groups, the average similarity was 0.36 for Group A and 0.41 for Group B. And with four partitions, the average similarity for each cluster ranged from 0.32 to 0.42, which on average was lower than the clustering with two partitions. When the number of divisions was further increased to eight, the average similarity of each cluster ranged from 0.29 to 0.38, again, on average, lower than the clustering with four divisions. The overall strategy encouraged a wider separation of highly similar nodes.

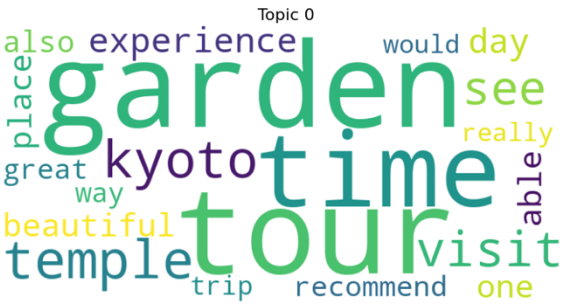

Figure 1. Example of a word cloud

TABLE I. Average Similarity Within Hierarchically Divided Groups

A				B			
0.36				0.41			
A1		A2		B1		B2	
0.33		0.36		0.36		0.42	
A1a	A1b	A2a	A2b	B1a	B1b	B2a	B2b
0.29	0.29	0.30	0.35	0.29	0.37	0.38	0.38

As subdivision deepened, the average similarity became lower, and the segmentation strategy of severing the edges of strongest similarity yielded better results. Based on these results, we can say that this approach could foster globally diverse clusters and ultimately increase the diversity of recommendations in the selection of attractions.

2) Comparison with k-means

To understand how Max-Cut clustering differs from a more conventional approach, we used k-means with the same feature vectors and the same numbers of clusters (2, 4, and 8) as shown in Table II. For 2 divisions, the k-means clusters displayed average similarities of 0.45 and 0.49—substantially higher (i.e., less diverse) than the Max-Cut results. Increasing the number of k-means clusters to 4 and then to 8 still generally yielded clusters with higher intra-cluster similarities than the Max-Cut approach.

A particularly striking observation is that some k-means clusters showed average similarities above 0.70, indicating that spots within a single cluster were very much alike, narrowing the variety of experiences if one were to use those clusters as "itinerary segments" for a tourist. In contrast, our Max-Cut approach ensured that many high-similarity edges were cut, more evenly distributing similar spots across different groups.

C. Implications for Tourism

Given these results, Max-Cut-based clustering encourages the construction of itineraries or tours that include a richer array of experiences (e.g., modern entertainment mixed with historical or spiritual visits). For tourists seeking novelty and a well-rounded trip, this method can systematically push them toward exploring places they might overlook in a similarity-driven approach.

Moreover, the overall average similarity for all 60 places was 0.3957. By strategically cutting edges with high similarity, we can maintain or even reduce the average similarity within each group compared to that global baseline, thus ensuring each group remains broad and distinct.

TABLE II. Comparison of Average Similarity Between Max-Cut and K-means

Clusters	Max-Cut Avg. Similarity	K-means Avg. Similarity
2	0.36, 0.41	0.45, 0.49
4	0.32–0.42	0.45–0.62
8	0.29–0.38	0.52–0.73

IV. Discussion

A. Advantages of the Max-Cut Approach for Diversity

1) High-Dispersion Clustering: Unlike clustering approaches (e.g., k-means) that often minimize variance within clusters, the Max-Cut approach intentionally cuts edges with high similarity. This results in partitions that have lower intra-cluster similarity, offering a high degree of internal variety.

2) Better Coverage of Different Themes: By splitting apart strongly similar spots, each group ends up covering multiple thematic dimensions gleaned from the LDA. This is particularly beneficial for tourist applications, where travelers frequently desire a balanced itinerary combining, for instance, historical temples, contemporary art venues, hands-on cooking classes, and nature sites.

3) Potential Alleviation of Filter Bubbles: If a traditional CF-based system were to predominantly recommend items close to a user's existing preferences, it might perpetuate a narrow scope. Integrating Max-Cut-based partitions can periodically broaden users' exposure to different categories, mitigating echo-chamber effects.

B. Challenges and Future Research

1) Hierarchical Submission and User Fatigue

While our approach can keep subdividing clusters to reveal increasingly fine-grained diversity, we risk creating clusters that are too small or too numerous, potentially overwhelming end-users. A practical stopping criterion—such as halting splits when a cluster size is below a user-defined threshold—could mitigate this issue [11], [12].

2) Computational Costs

Max-Cut is NP-hard, and while the Goemans-Williamson SDP algorithm provides an approximation ratio, it can still be computationally expensive for large datasets [2]. Future work might explore more efficient or parallelizable versions of the SDP-based method, or even emerging paradigms such as quantum approximate optimization [13].

3) Interpretability and Labeling

Automatic labeling of each diverse cluster remains nontrivial. While LDA topic-word distributions and Large Language Models (LLMs) can offer label suggestions, domain experts or dedicated labeling protocols could refine these labels for practical deployment.

4) Generalizability Beyond Tourism

Although our empirical study focuses on tourist spots, the concept of "forcing diversity" by cutting high-similarity edges is broadly applicable. For example, in personnel management or academic group allocations, the approach could help form teams with diverse skill sets, fostering innovative problem solving [14].

C. Applications to Recommendation Systems

Conventional recommender systems are predominantly designed to sequentially present items that are highly like one another. While this approach has proven effective in various domains, it can be particularly limiting in the context of tourism. Algorithmically driven or search-engine-based suggestions often result in biased travel routes, concentrating recommendations within a narrow thematic or geographic scope. However, travel is inherently a multifaceted experience—encompassing culture, nature, food, and hands-on activities—and is most fulfilling when these diverse elements are interwoven. From this perspective, we propose a recommender framework that intentionally separates highly similar items to foster diversity in the resulting recommendations.

1) Rethinking Diversity in Tourism Recommendations

Traditional methods such as collaborative filtering and content-based recommendation rely on analyzing a user's past behavior and suggesting items with similar characteristics [15]. While these techniques are proficient in predicting preferences, they risk reinforcing what has been termed the "filter bubble" effect, wherein user exposure is progressively narrowed. In tourism, this can manifest as itineraries that include only temples, only shopping districts, or other overly homogeneous sequences—thereby missing the richness that stems from combining distinct experiential categories.

In contrast, intentionally mixing items with low semantic similarity may better reflect the organic and serendipitous nature of real travel experiences. From a psychological perspective, such diversity also aligns with the principles of Self-Determination Theory (SDT) [16], which posits that autonomy, competence, and relatedness are central to intrinsic motivation. Providing users with a broader, more heterogeneous set of recommendations not only enhances their sense of agency and competence but also fosters a deeper engagement with the recommendation system.

2) Structuring for Intentional Separation of Similar Items

The Max-Cut-based clustering framework introduced in this study is designed to enhance diversity by explicitly separating items with high topical similarity. Rather than aggregating similar points of interest into a single group, as is common in traditional clustering approaches, our method partitions such items into distinct clusters. This deliberate decoupling ensures that each cluster captures a broad spectrum of themes, thereby increasing internal heterogeneity.

This structure allows the recommender to surface content that spans multiple domains, enabling, for example, the combination of historical architecture with contemporary art spaces or traditional craft workshops with culinary experiences. In doing so, the system encourages exploration and prevents the user from being confined to a narrow set of familiar or predictable categories.

3) Designing a Diversity-Oriented Recommender Architecture

In contrast to systems that rely on historical user interactions or matrix-based similarity scores, our proposed architecture emphasizes diversity from the outset. The pipeline begins by extracting textual features from item descriptions, user reviews, or metadata using topic modeling techniques such as Latent Dirichlet Allocation (LDA). Each item is thus represented as a topic distribution vector that encapsulates its latent semantic structure.

Subsequently, a similarity graph is constructed by calculating cosine similarity between these topic vectors, with edge weights representing degrees of content similarity. Applying the Max-Cut algorithm to this graph produces clusters where highly similar items are intentionally placed in

separate groups. This clustering strategy, which prioritizes dissimilarity within each subset, forms the foundation for diversity-aware recommendation.

Recommendations are then generated by sampling items from each cluster, resulting in a final list that intentionally traverses thematic boundaries. To enhance explainability, the system can provide short textual descriptions based on the dominant topics in each cluster—for example, indicating that a recommendation set represents a blend of "traditional culture" and "modern entertainment." This transparency not only clarifies the logic behind the recommendations but also supports user exploration.

4) Supporting User Autonomy and Intrinsic Motivation

From the perspective of Self-Determination Theory, enabling users to make meaningful choices and understand the rationale behind those choices is essential to sustaining intrinsic motivation [16]. By offering recommendations that span a wide range of topics and allowing users to navigate across clusters, the Max-Cut-based system increases the sense of autonomy. Furthermore, the exposure to novel items enhances users' perceived competence, while thematic breadth fosters a sense of connection to broader experiences and communities.

In sum, our approach reconceptualizes the structure of recommendations—not as a sequence of similar items, but as a curated mixture of intentionally diverse experiences. This aligns more closely with the nature of tourism and other exploratory domains, and provides a pathway toward systems that are not only accurate but also engaging, inclusive, and psychologically satisfying.

CONCLUSION

This study presented a new approach to tourist spot recommendation—leveraging an SDP-based Max-Cut algorithm with *cosine similarity* as edge weights to achieve clustering that emphasizes diversity rather than similarity. We demonstrated its effectiveness using 600 TripAdvisor reviews from 60 major tourist attractions in Kyoto. By specifically targeting the cutting of edges representing high similarities, we formed clusters that displayed lower average intra-cluster similarity compared to conventional k-means. Hierarchical and multi-level partitions further allowed for both broad and fine-grained groupings, suitable for generating eclectic travel itineraries or for balancing user fatigue with a desire for variety.

Beyond tourism, the concept of maximizing cuts on high-similarity edges opens the door for forming diverse teams in academic, professional, or community settings, enabling better group composition and innovation potential. Future research should focus on computational optimizations for large-scale data, refining stopping criteria for hierarchical splits, and developing improved labeling and explanation mechanisms for each diverse cluster. By combining diversity-driven partitioning with targeted personalization, recommendation systems can achieve a richer balance of coverage, relevance, and user engagement.

ACKNOWLEDGMENT

We would like to express our gratitude to all the people at Musashino University who collaborated on this research, especially the Human Behavior Engineering and Social Innovation Laboratory, and to Professor Virach Sornlertlamvanich and Professor Thatsanee Charoenporn for their valuable support.

REFERENCES

[1] Y. Zhang and X. Chen, "Explainable recommendation: a survey and new perspectives," *Found. Trends Inf. Retr.*, vol. 14, no. 1, pp. 1–101, 2020.

[2] M. X. Goemans and D. P. Williamson, "Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming," J. ACM, vol. 42, no. 6, pp. 1115–1145, 1995.

[3] J. Miguéns, R. Baggio, and C. Costa, "Social media and tourism destinations: TripAdvisor case study," *Advances in Tourism Research*, 2008.

[4] Kyoto City Tourism and Industry Bureau, "Comprehensive Survey of Kyoto Tourism," 2024.

[5] Kato, "Community, connection and conservation: Intangible cultural values in natural heritage—the case of Shirakami-Sanchi World Heritage Area," *Tourism Hosp. Res.*, vol. 6, no. 2, pp. 93–104, 2006.

[6] P. Rehurek, P. Sojka, and J. Sojka, "Software framework for topic modelling with large corpora," *Proc. 7th Int. Conf. Language Resources and Evaluation (LREC'10)*, Valletta, Malta, pp. 45–50, May 2010.

[7] L. Griffiths and M. Steyvers, "Finding scientific topics," *Proc. Natl. Acad. Sci. U.S.A.*, vol. 101, suppl. 1, pp. 5228–5235, 2004.

[8] M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[9] A. Aletras and M. Stevenson, "Automatic evaluation of topic coherence," *Proc. 10th Int. Conf. Computational Semantics (IWCS)*, Potsdam, Germany, pp. 13–22, Mar. 2013.

[10] D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," *Proc. 2011 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, Edinburgh, Scotland, pp. 262–272, Jul. 2011.

[11] S. Iyengar and M. R. Lepper, "When choice is demotivating: Can one desire too much of a good thing?," *J. Pers. Soc. Psychol.*, vol. 79, no. 6, pp. 995–1006, 2000.

[12] S. Higashino and Y. Hamasuna, "A study on automatic estimation of the number of clusters based on hierarchical clustering," *Proc. 38th Fuzzy Systems Symposium (FSS2022)*, online, pp. 679–684, 2022.

[13] R. McClean, S. Boixo, V. N. Smelyanskiy, R. Babbush, and H. Neven, "Quantum approximate optimization of non-planar graph problems on a planar superconducting processor," *Nature Phys.*, vol. 16, no. 6, pp. 1061–1065, 2020.

[14] S. E. Page, *The Difference: How the Power of Diversity Creates Better Groups, Firms, Schools, and Societies*, Princeton University Press, 2007.

[15] Y. Koren, R. Bell, and C. Volinsky, "Matrix factorization techniques for recommender systems," *IEEE Comput.*, vol. 42, no. 8, pp. 30–37, 2009.

[16] M. Ryan and E. L. Deci, "Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being," *Am. Psychol.*, vol. 55, no. 1, pp. 68–78, 2000.