Paper Draft: Hotel Recommendation System Using Content-Based Filtering

Group 2: Qian Li, Xin Ding, Wenyu Yang, Zhengyi Xu

Introduction

With the growth of online travel platforms, travellers today face an unprecedented number of accommodation options. While this abundance provides great flexibility in travel planning, it also means that the more options available, the harder it is for users to decide.

Traditional hotel recommendation systems need help with three major limitations: the cold-start problem for new users, outdated historical data, and information overload during hotel selection. These challenges often result in irrelevant recommendations. Our content-based filtering system addresses these issues by providing instant, personalised recommendations based on current hotel characteristics rather than user history.

To address these limitations, we propose an intelligent hotel recommendation system specifically designed for Seattle visitors, utilising natural language processing (NLP) and content-based filtering techniques. Our system analyses hotel descriptions to generate recommendations through textual similarity, solving the cold-start problem while maintaining recommendation relevance. The system leverages text processing techniques including text preprocessing, TF-IDF vectorization, and cosine similarity calculations. Our content-based filtering approach provides instant recommendations based on users' preferences without requiring booking history. This ensures that recommendations stay relevant regardless of seasonal changes.

Related Work

Several approaches have been studied carefully and popularly used in this field, including content-based filtering, collaborative filtering, and context-aware recommendations.

Content-based Filtering

This is the approach we attempt to focus on in this project. Instead of utilising existing users' data, this approach focuses more on the product's (in our case, hotels) features and properties. It is effective for a cold start or the case of insufficient user data.

Collaborative Filtering

This approach uses past user behaviour for a personalised recommendation result. There are mainly 3 directions to implement:

- User based. It recommends products that similar users liked. The assumption is that users could have similar consumer behaviour and product preference inside a group. After clustering users with similar interests, we could assume that if user A made a purchase on product P, then user B from the same group would probably agree that product P is a great fit for him.
- Item based. It recommends products similar to ones the user previously purchased. This is based on the assumption that people are likely to have the same taste for similar products.

- Matrix factorization techniques like SVD effectively combine aspects of both user-based and item-based collaborative filtering. It uses both user and item factors in prediction simultaneously for better accuracy.

Context-Aware Recommendations

This method considers situational and environmental factors instead of traditional 'user-item' relationship and it adapts recommendations based on current circumstances. For example, considering the seasonal context, people may care more about ski access, heated pools and indoor facilities in Winter while focusing more on beach access, outdoor pools and gardens in Summer. Other contextual factors include travel purpose (business, leisure, family, etc), travel group size, length of stay and special events (festival, wedding, conference, sports, etc).

The system calculates individual scores for each context and combines them using weighted averages to provide final recommendations. The weights can be adjusted based on the relative importance of each factor.

In real-world projects, we often combine those approaches for better coverage and accuracy. For example, we could use content-based filtering for new users as an effective cold start, and then switch to collaborative as more user data becomes available. Another option is that for each user we could weight results from both approaches based on data availability.

Dataset Description

Our original plan was to collect descriptions of hotels in Seattle downtown and surrounding neighbourhoods from third party hotel booking websites. However, after some initial search, we found out that those descriptions on third party websites appear too brief for our content-based filtering analysis. Therefore, we changed our data collection strategy, and directly collected data from the hotel's official website based on a list of hotels in Seattle found on third party hotel booking websites. In this way, we managed to generate a dataset of 152 hotels containing 3 main fields: hotel name, hotel address and a comprehensive hotel description, stored in the form of a csv file convenient for later training data parsing and model training.

In locating target hotels to be included in our dataset, we set the geographic coverage to be hotels in Seattle downtown and nearby areas such as Bellevue, Redmond, and SeaTac Airport area to imitate typical traveller's search range when travelling to Seattle. We included hotels of different types from Luxury hotels and Business hotels to Hostels and Bed & Breakfasts so that our data is diversified enough for similarity analysis and easy result validation. For fields of each hotel, we use the basic information of hotel name, full street address with zip code, as well as detailed hotel description. Specifically for the hotel description, we tried our best to include information of location details and nearby attractions, room types and amenities, property facilities, services offered, target audience (e.g. business travellers, tourists, etc.), transportation access, and nearby points of interest. By using such rich texts for hotel description, our content-based filtering and analysis is made possible, although requiring lots of preprocessing work to make the textual data more structured to apply similarity calculation.

To sum up for our dataset, it is a csv file containing 152 hotels with name, address and detailed description ready for rich text preprocessing the following text preprocessing, feature extracting and similarity analysis. Its diversified range of hotel types and rich text makes it a good fit for our content-based filtering analysis. We decided not to include structured quantitative fields such as price, rating and distance to key tourist interests, so as to differentiate from other more quantitative based hotel recommendation implementations and to focus our recommendation solution on content-based filtering technique.

Results

The start of this analysis focused on understanding the dataset. We performed data cleaning and text preprocessing mainly to remove special characters and filter out stop words. We then identified the top 20 most frequent words and phrases appeared in the hotel descriptions are: *seattle, hotel, centre, downtown, free, located, rooms, stay, place, airport, market, space, enjoy, pike, inn, business, pike place, pike place market, place market and just*. Additionally, we plotted a histogram graph to visualise the distribution of description lengths to have a deeper insight into the data structure.

Moving to the similarity calculation, we used TfidfVectorizer to compute the TF-IDF scores. This generated a matrix with the size of (152, 27001). Using this matrix, we calculated the cosine similarity that represents the relationship between each hotel and others with the size of (152, 152). As expected, the diagonal values of this matrix are always 1, reflecting the perfect similarity of each hotel with itself.

For the feature of hotel recommendation, currently our model identifies the top 10 most relevant hotels based on cosine similarity. To enhance accuracy and improve user experience, we plan to optimise this feature by including real-world user ratings later. We will ask users to rate the similarity of hotels based on their personal experience and preferences. By combining the qualitative rating results with quantitative computing results, we aim to create a more user-centric and realistic hotel recommendation model.