Hybrid recommender system based on Yelp user reviews
Shanmugavel Gnanasekar, Ravi Shankar Subramanian
Masters in Business Analytics - Oklahoma State University

Introduction

• Typical recommendation engines in Amazon, Pandora or Netflix, analyze your historical buying behavior and make various recommendations that makes your shopping enjoyable.
• We have built a similar recommender system to suggest restaurants by combining the content from Yelp user’s profile, their reviews on various restaurants, the rating given on a scale of 1-5, the restaurant details and the tips provided by the user.
• Traditional systems provide recommendations based on only the user’s ratings. However, our system will take advantage of both the user’s reviews/content and the ratings to recommends.
• The content based system is modeled by analyzing their each user’s past reviews and ratings, to identify their preferences and associating them with key words such as a particular cuisine, hygiene, bar, valet parking.
• The collaborative-filtering system is modeled based on similarity measures of users preference from their reviews/ratings. The restaurants recommended to a user are those preferred by similar users.

Methods

Hybrid recommender system based on Yelp user reviews

Data Preparation

• For our study, around 2.2M reviews and 591K tips by 552,000 users were downloaded from the Yelp website. Due to the sheer volume of data, only the businesses in Las Vegas were considered for modeling.
• The user reviews dataset has variables such as user id, rating, review and business id corresponding to the restaurant. The Businesses dataset has variables such as business id, location and variables for facilities available.
• Yelp provided the data in the JSON format which was converted to CSV using a Python script. After importing the csv file into SAS system, PROC SQL and custom-built MACROs were used to clean the data and make it consistent.

Data Modeling

• The reviews with rating of 3 and above were considered as 1 and those below 3 are 0. A user profile based on his ratings were created. Similarly a profile was created for each business based on the ratings it received from the users.
• SAS Text Miner was used to analyze the reviews of the user and extract key terms from each review. Text parser node was used to remove stops words, perform stemming and identifying the parts of speech.
• Text filter node is used to identify the key words in each review. The concept links were generated to find out each key terms and its associated terms These terms were used to provide content based suggestion.
• To identify the sentiment of each review, Text Rule node is used with the default user topics. The users who have given similar kind of reviews for the business were aggregated using the Text cluster node.
• Concept links were derived from the text filter node which helps to understand the association of one term with other. This will help in making content-based recommendation to the user based on their reviews.
• The user reviews were classified as positive and negative based on the key words and the ratings they have given. For computation purpose, positive reviews were coded as ’1’ and negative reviews were coded as ’0’.
• In the similar fashion, cumulative positive and negative score for each business based on the user review was calculated.
In content based system we build profile for each user and business.

If an user prefer Chinese restaurants, restaurants with valet parking facilities, these preferences are saved as his profile.

Similar to user profile, we build profile for each business, based on the attributes provided by Yelp.

Inverse Document Frequency (IDF) for each attributes are calculated and is included in calculation of recommendations for each user, to prioritize rare preference over common preference.

We used the rating of the model to build the positive and negative profile for user. For instance if user has rated 5 for hotel that serves Chinese, then we increase Chinese attributes by 1. Larger the number for preference

Then the recommendation prediction is calculated as per the above formula

The performance of the model is evaluated by comparing the top five recommendations with their original rating and calculating the Root Mean Square Error(RMSE) and Mean Absolute Error(MAE).

\[
RMSE = \sqrt{\frac{\sum (pred - rating)^2}{N}}
\]

\[
MAE = \frac{\sum |pred - rating|}{N}
\]

The RMSE value of our content based recommender system is 0.447 and MAE value is 0.2

The system could provide near accurate recommendations based on the user profile.
Collaborative-filtering Based Recommender System

Method

- In the Collaborative-filter Based Recommender system we use ratings of neighbor (other users with similar taste) to make recommendations any user.
- Building collaborative recommender involves 2 steps:
  1. Neighborhood for each user is computed.
  2. After determining neighborhood, we make prediction as normalized similarity weighted rank.

Rating

- In this method neighborhood of each user is selected by computing correlation of the user with every other user. Then top 5 or 10 user is selected to be neighborhood.
- After deciding the neighborhood for given user, neighbors rating for different business is aggregated as normalized, similarity weighted rating as shown in above equation.
- Top 5 predictions are shown to user or used to calculate metric

Review

- In this model, reviews of users are mined and clustered into 10 clusters.
- Neighbors are then clustered based on the values in these clusters.
- Users belonging to same clusters have similar preferences in the restaurant selection.

References


Acknowledgement

- We thank Dr. Goutam Chakraborty, Director, Master of Science in Business Analytics – Oklahoma State University and Dr. Miriam Mcgaugh – Oklahoma State University for their continuous support throughout our research.