CS591 Data Mining Mid-project Report
Yelp Dataset Challenge

Team Members:
Renqing Gao - gladius@bu.edu
Amal Kadi - akadi@bu.edu

Abstract

In this project we're going to analyse a dataset from Yelp for multiple cities around the world. We will concentrate on analysing the business data (to get information on each store\restaurant) and also the check-in data for those businesses. We are expecting to infer cultural and seasonal trends. We will also try to get relations between users (reviewers) to give recommendations based on that.

Introduction

Our main motivation is to analyse the businesses and from there get cultural\seasonal trends in addition to popular times and patterns. We will also be showing users that are similar by creating a graph network for each city and provide recommendations based on that.

Our dataset contains data on businesses\check-in times\reviews in multiple cities of the world (cities in: UK, Germany, Canada & US) so here is where the cultural trends will show.

We will be using clustering and classification to get some regulation about both users and businesses. We might also implement prediction and recommendation system for users to recommend hot restaurant near them.

Technique

Clustering:

We will be using k-means for clustering since we need a partitioning relocation clustering instead of hierarchical clustering technique or density-based clustering. We will cluster the businesses from all the included cities according to the check-in routine of users and also according to their attributes.(e.g., if they both accept credit card or if they provide delivery) Also, this clustering result will be helpful for our prediction & recommending application.

Creating a network graph to find highest degree between reviewers:

With network graph we can get the information about two users' similarity and this could also be used in our recommending application, for example, we could recommend users nearby according to which business they've checked-in previously.

Prediction & Recommendation:

Similar to homework 4, we can predict the user's rating for some restaurant in another city and recommend restaurants that a user might like no matter in his city or in other city. This function could be integrated with some travel info website.(e.g., Yelp could co-op with company like StudentUniverse, when someone buy a flight ticket from StudentUniverse, he can sign in with his yelp account and we can recommend restaurant in his destination according to his reviews).

Datasets and experiments

As a pre-processing step, we have collected the relevant information that we will be needing from each of the json files (yelp_academic_dataset_checkin.json - yelp_academic_dataset_business.json - yelp_academic_dataset_review.json). From the review file, we have made a dataframe including the review_id, user_id, business_id, stars, and date of each review. From the business file, we have made a dataframe including the business_id, review_count, stars, type, address, attributes(71 attributes), categories, hours, and name of each business. From the checkin file, we have made a dataframe including business_id and number of checkins at each hour of the day of each business.

Then we cluster those dataframes we got after pre-processing. We look into their check-in time and attributes and divide them into different group and compute their similarity using jaccard distance (In this process we will ignore their location).   In this way when we find that a user is giving a high rating to a business, we'll be able to look into the group that the business belongs to and find the most similar businesses. Before returning the result list we will filter using the location that our user provided and return those business that he/she might be interested in the specified city.

The third experiment will be creating a network graph about users data, in this way our users can find more friends that they share some common interests. No need to mention more friends in contact will make users tend to keep using the same application in the long term. We would also try recommend users from different countries that share some common interests, which we believe might promote users planning a trip.

Our final experiment will be implementing our recommending application which is mentioned above. We will using the information about a user and our previous result.

Results and Discussion

The first result should be a result clustering business, we're going to find out what is the most popular time for a restaurant, and this could go up to a higher level like what's the popular time for this district.

Our second result we're going to create a network graph about users and cluster similar users based on their reviews and ratings.

At last, we will also implement a recommending system based on the first two results. With this system, users can input their user ID and city (city could be optional), then the user will get an output list of business/users that he or she might be interested in.

Conclusion

We will provide a user based recommending application that's recommending friends and business that current user might be interested.