CSCI 5622 Project: Recommender System based on Yelp Dataset

Team members : Xu Han, Yichen Wang, Yan Li, Juan Lin

20 Feb 2018

1 Background and Motivation

With the explosion of the web information, users are being presented with tremendous ranges of choices, at the same time, sellers are being faced with the challenge of personalizing their advertising efforts. Along with that, it has become common for enterprises to collect large volumes of transactional data that allows for further analysis of how a customer interacts with the space of available products. The motivation of a Recommender system(RS) is to provide meaningful recommendations to a collection of current users or potential users for items or products in which they might be interested in. In this study, reading all the reviews of a single business itself is time-consuming and requires more effort than the average user is willing to spend. Simplifying and automating this process using machine learning techniques make us believe an effective recommendation system would let the users benefit from it a lot.

2 Data

The data was pulled from the Yelp API which contains 5.2 million reviews, 174 thousand businesses, and over 1.2 million business attributes. The data is stored in an uncompressed SQL file containing 7.55 Gbs which is free for everyone to download and use. Yelp also provides open APIs for users and developers. They are easy to get what data they need with GraphQL, a query language. We plan to store data in an Amazon RDS in AWS which allows us to connect Amazon Web Server instead of our local server. It would help us to use same data at the same time and it is more safe and available than local SQL server. We may plan to try MongoDB if it is efficient.

3 Techniques

We plan to divide the dataset as training part and validation part. The ratio is 4:1. With the training dataset, we will build utility matrices[1] with users' rating data and use several techniques to predict recommendation results. The techniques we are going to use include machine learning techniques and other current RS techniques i.e. content-based system building[2] and collaborative filtering[3]. Furthermore, we will compare and evaluate the recommendation results made by different techniques on the validation set.

3.1 Machine Learning Techniques

Based on the utility matrices, we will first focus on feature engineering and then deploy several machine learning classifiers to predict the ratings of all items for each user and make recommendations based on the predicted ratings. For feature engineering, we will extract basic features(like user profile feature, merchant-related features), review features(like TF-IDF feature, LDA feature), user behavior features(like tip feature, check-in activity feature, time-related feature) and complex features(like aggregated features, PCA features, trend and similarity features). Since this rating prediction issue can be simplified as a binary classification problem[1], some classic classifiers like Logistic Regression, Random Forest, Gradient Boosting Machine, Support Vector Machine, AdaBoosting will be utilized. What's more, to

further improve the performance of our machine learning techniques, we plan to deploy a blending model to combine all these single classifiers.

3.2 Current Recommendation System Approaches

There are two main architectures for the recommendation system, content-based system and collaborative filtering system. The content-based system focuses on the profiles of entities. The profiles will be used to measure the similarities between entities. The preference of users on items will be evaluated by their profiles' similarities, based on which items will be recommended to users. Collaborative filtering focuses on the relationship between users and items. For example, the similarity between two users is measured by the items they liked, i.e. the rows in the utility matrix. Items liked by user A will be recommended to user B if they are the most similar users.

3.3 Evaluation

For machine learning techniques, we will evaluate the results based on the validation set. Precision, recall and AUC score will be used as the measurement. For recommendation system, the validation set will also be used as evaluation dataset.

4 Timeline

Mar.16 - Data preprocessing and initial feature engineering

Apr.6 - Feature engineering and model training

Apr.20 - Evaluation and website building

May.4 - Other RS algorithms implementation and report writing

References:

[1]. Ricci, Francesco, Lior Rokach, and Bracha Shapira. "Introduction to recommender systems handbook." *Recommender systems handbook*. springer US, 2011. 1-35.
[2]. Pazzani, Michael J., and Daniel Billsus. "Content-based recommendation systems." *The adaptive web*. Springer, Berlin, Heidelberg, 2007. 325-341.
[3]. Linden, Greg, Brent Smith, and Jeremy York. "Amazon. com recommendations: Item-to-item collaborative filtering." *IEEE Internet computing* 7.1 (2003): 76-80.