YELP BUSINESS RECOMMENDER SYSTEM

Yuan Jiang, Young H.Kang, Harry Qi

1 PROBLEM SETTING & OBJECTIVE

We aim to build a business recommender system for Yelp. In particular, we aim to recommend the top 10 businesses to a user by predicting his or her rating for a business yet to be rated. The prediction is based on the previous ratings and reviews by the user, in combination with those by other users and various characteristics of businesses, such as location and type.

2 METHODS & TECHNIQUES

2.1 TRAINING ALGORITHM

We propose focusing mainly on building a user-rating predictive model first via a collaborative filtering approach, and later by combining extra information such as reviews and locations with the rating model for the recommender to generate preferred candidates.

Two mainstream ideas of collaborative filtering are *neighbourhood methods* and *latent factor models.*[1] We implement a simple kNN based neighbourhood CF algorithm as a baseline. Then, we implement a more sophisticated hybrid approach integrating various techniques such as *singlular value decomposition*[1], *matrix factorization*, and *normalization of global effects*. We would roughly follow Yehuda Koren's strategy[2] building our hybrid latent factor model.

The neighbourhood method focuses on finding similarities between users, items (e.g. resturants), or both, Based on the reasoning that "neighbouring" users give similar ratings to a movie, neighbouring items would equivalently receive similar scores from a specific user. On the other hand, the latter approach attempts to model user-item interaction as inner product in a joint latent factor space, by mapping both users and items to some factor vectors. [3]

While the methods have already been studied before, we still face the challenge of dealing with the sparse nature of the dataset, also called the cold start problem. This requires a case-by-case solution based on the charactieristics of the data obtained.

2.2 EVALUATION METRIC

Evaluation is based on how accurate the rating predictions. We would take the average performance via K-fold cross validation, which adopts both the root mean square error (RMSE) and mean absolute error (MAE) metrics. Previous research has indicated that both metrics are highly accurate on rating prediction systems.[4]

However, we have not yet decided on the validation method for the quality of the final

[1] http://web.mit.edu/be.400/www/SVD/Singular_ Value_ Decomposition.htm

recommendation list. This is becuase such validation often requires sufficient data over a long period of time, and is also because lingering arguments on the notion of quality (e.g. people would sometimes prefer things not consistent with their previous taste, meaning a recommender producing overly similar results may not be actually preferred). We would leave this for further discussion.

3 Data Description

We plan to use the data provided by Yelp for the Yelp Dataset Challenge[2] in 2014. This particular dataset includes data from five cities – Phoenix, Las Vegas, Madison, Waterloo and Edinburgh – and consists of 42,153 businesses, 252,898 users, and 1,125,458 reviews. Because we are mainly interested in restaurant businesses, we will use Kaggle's data[3], which is just the restaurant portion of Yelp's data. While 11,537 businesses, 43,873 users, and 229,907 reviews are used for training, about 10 % of the data (1,205 businesses, 5,105 users, and 22,956 reviews) are used for testing. There are 15 attributes for the business data, 7 attributes for the review data, and 11 attributes for the user data. We would also generate some cross-joined featrues to represent information such as user bias and geographical bias.

Because they are nicely formatted and organized into multiple json files, processing the data is quite staightforward. To avoid any unwanted correlation, we would shuffle the data randomly before selecting the testing portion.

4 Milestone Goals

We expect to have finished processing all data into building our recommender system. We plan on having finished building our model by the milestone deadline and having tested and improved our model's performance through cross validation on all training data by randomly dividing them into ten "testing data" groups. After the milestone date, we would continue optimizing performance via trying other techniques to deal with "cold-start".

References

[1] Xiaoyuan Su and Taghi M. Khoshgoftaar, *A Survey of Collaborative Filtering Techniques*, Advances in Artificial Intelligence, 2009

[2] Yehuda Koren, Robert Bell and Chris Volinsky, *Factorization Techniques for Recommender System*, IEEE Computer Society, 2009

[3] Yehuda Koren, *Factor in the Neighbors: Scalable and Accurate Collaborative Filtering*, AT& T Research

[4] Guy Shani and Asela Gunawardana, *Evaluating Recommendation Systems*, Microsoft Research

[2] http://www.yelp.com/dataset_ challenge

[3] https://www.kaggle.com/c/yelp-recsys-2013/data