CS 761 Project Proposal

Submitted by: Sneha Rudra and Clarence Cheung

Resnick and Varian [1] define recommendation systems as systems that use the opinions of a community of users to help individuals in that community more effectively identify content of interest from a potentially overwhelming set of choices. These can be thought of as information filtering systems that predict how users would 'rate' a product that they have not used or experienced before. Today, recommender systems have become common and most commercial and social websites suggest options such as - movies, music, news, books, research papers, search queries, social tags, restaurants etc to users [2].

For this project we propose to apply the ideas of recommendation systems to Yelp dataset, available for download from Kaggle https://www.kaggle.com/c/yelp-recsys-2013. Yelp hosts and markets a web and mobile application by the same name [3]. The application publishes crowd-sourced reviews and ratings on variety of local businesses ranging from restaurants to health and medical services. It enables users to locate and search for local businesses of interest, to provide ratings and reviews for businesses that they may have tried and to connect with other users.

The data set corresponds to Yelp's dump data for Pheonix, Arizona area. The training data contains information about 11,537 businesses, 8,282 checkin sets/notification information, 43,873 users and 229,907 reviews. Business data is characterized by several features- Business ID, Type, Name, City, State, Latitude, Longitude, Neighbourhoods, Stars, Review Counts, Categories and a boolean variable indicating whether it is open/closed. User data is characterized by the following features- Type, UserID, Name, Review Count, Average Stars (which is a floating point variable) and Votes (useful, funny, cool). Reviews and Checkins are similarly characterized. The test data is similar in format as the training data but most data entries have several missing fields. The goal is to use feature extraction/model building to predict ratings for the test data.

Approaches to generating recommendations/ predicting ratings can be broadly classified into two main categories : (i) Content Based Methods (ii) Collaborative Filtering. In content-based recommendation methods [4], the system tries to generate user profiles to learn patterns in the user's past preferences and then items which are similar to those preferences are suggested. It is often common to request the user to fill out questionnares inorder to gather more data to create user profile. Content based methods are used in practice to recommend items that are labelled with tags/keywords. Hence if every item is characterized by a content profile -a set of keywords (which may have been obtained through feature extraction), then the user profile (vector indicating the relevance of a given keyword to the user) can be obtained from the content profiles of the items liked by the user in the past through means such as Rocchio algorithm [5]. Other content based approaches include building Bayesian classifier/decision tree to estimate probability of an item being liked. Usually classifiers take a long time to construct and are mostly used for smaller problems [6].

Collaborative filtering methods take a different approach and try to suggest items to user based on the items previously rated by other like-minded users (those who would rate items similarly as the given user) [4]. Breese et. al [7] organize collaborative filtering algorithms into memory-based and model based. Memory- based methods predict the rating that the user would assign to an item as an aggregate of the ratings given to the same item by other similar users. Common aggregation strategies include simple average, user-similarity weighted sum etc. Several approaches such as Pearson correlation, cosine based and graph theoretic approaches [8] have been used to compute the user-similarity. It is expected that several user-specified ratings would be needed for this method to work well and in the case of missing ratings, experiments show that imputing values can help [7]. Model-based algorithms on the other hand first learn a model from a set of known ratings and use the model to make predictions. Several model-based collaborative methods that have been explored in the literature include statistical model for collaborative filtering [9], bayesian model [10], linear regression [11] etc.

For this project, we propose to apply model based collaborative filtering algorithms to the Yelp dataset and in particular we propose to apply regression models learnt in class such as - Linear Regression, Ridge Regression and Lasso Regression. To deal with missing data in the test set, we propose to explore Column Subset Selection [12] and Matrix Completion Methods [13]. Finally we propose to evaluate the different models based on metrics such as Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Normalized RMSE and MAE, Average RMSE and MAE [14].

REFERENCES

[1] Resnick, Paul, and Hal R. Varian. "Recommender systems." Communications of the ACM 40,no.3(1997): 56-58.
[2] Koren, Yehuda, Robert Bell, and Chris Volinsky. "Matrix factorization techniques for recommender systems." Computer 8 (2009): 30-37.

[3] http://www.yelp.com

[4] Adomavicius, Gediminas, and Alexander Tuzhilin. "Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions." Knowledge and Data Engineering, IEEE Transactions on 17, no. 6 (2005): 734-749.

[5] J.J. Rocchio, "Relevance Feedback in Information Retrieval," SMART Retrieval System—Experiments in Automatic Document Processing, G. Salton, ed., chapter 14, Prentice Hall, 1971.

[6] M. Pazzani and D. Billsus, "Learning and Revising User Profiles: The Identification of Interesting Web Sites," Machine Learning, vol. 27, pp. 313-331, 1997
[7] J.S. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," Proc. 14th Conf. Uncertainty in Artificial Intelligence, July 1998.
[8] C.C. Aggarwal, J.L. Wolf, K-L. Wu, and P.S. Yu, "Horting Hatches an Egg: A New Graph-Theoretic Approach to Collaborative Filtering," Proc. Fifth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining, Aug. 1999.
[9] L.H. Ungar and D.P. Foster, "Clustering Methods for Collaborative Filtering," Proc. Recommender Systems, Papers from 1998 Workshop, Technical Report WS-98-08 1998
[10] Y.-H. Chien and E.I. George, "A Bayesian Model for Collaborative Filtering," Proc. Seventh Int'l Workshop Artificial Intelligence and Statistics, 1999.
[11] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-Based Collaborative Filtering Recommendation Algorithms," Proc. 10th Int'l WWW Conf., 2001.
[12] Balzano, Laura, Robert Nowak, and Waheed Bajwa. "Column subset selection with missing data." In NIPS Workshop on Low-Rank Methods for Large-Scale Machine Learning, vol. 1. 2010.
[13] Ganti, Ravi Sastry, Laura Balzano, and Rebecca Willett. "Matrix Completion Under Monotonic Single Index Models." In Advances in Neural Information Processing Systems, pp. 1864-1872. 2015.
[14] Shani, Guy, and Asela Gunawardana. "Evaluating recommendation systems." In Recommender systems handbook, pp. 257-297. Springer US, 2011.