CS224W Project Proposal: Predicting Yelp Ratings From Social Network Data

Kelvin Do, Rotimi Opeke, James Webb
Department of Computer Science, Stanford University

October 16, 2015

1 Introduction

Yelp, is a consumer application popular that has consistently kept itself relevant among other web review applications by integrating business reviews within an extensive social network. While the ability to add friends and internally message other users isn't as strongly emphasized across the product, the social component of the application helps provide incentive for users contribute to the greater Yelp community. This graph-like community, of both users and businesses creates an invaluable network of information from which future user use behavior can be modeled.

In representing users, businesses and their relationships, our team wanted to take a second look at the relational Yelp data to make predictions about the actions that users would take on the network - specifically to make predictions on their future review scores. Through a study of the communities that exist within the graph and through several different metrics that we can use to study similarity we hope to serve unique predictions about a user's future review scores based on information made available through graphical analysis. Although machine learning algorithms can be run against this information in the absence of a graph, we hope to use the relational information provided through the graph to make even more accurate predictions.

2 Literature Review

2.1 Effects of User Similarity in Social Media (Anderson, Huttenlocher, Kleinberg, Leskovec, 2012)

This paper [1] explored the possibilities of predicting evaluations between users based on information about the users themselves, not their previous evaluation history. Specifically, for sites that allow user contribution to be rated, such as Wikipedia and StackOverflow, the research examined how a contribution would be received by a given audience. The researchers used a similarity score, based on overlapping interests between users, and studied how this corresponded with positive evaluation of user contributions.

Exploring social interactions on the web hits at the heart of social network analysis. Users rating each other's contributions could be modeled as nodes linked on a graph. Similarity measures provide more insight into the relation between nodes, and in this case were used to predict the weights of those ratings.

This research was strong in that it very succinctly demonstrated the efficacy of using similarity scores. Each result showed much promise for the relationship between users interests and their evaluation of each others input. What the paper lacked was any extension into addressing the content of user contributions. Although it established similar users tend to approve of each others opinions, it did not attempt to predict if similar users have similar opinions. Furthermore, the measure of similarity between users was limited to only an overlap of interests. This measure could be made more rich by considering social forces at play. For instance, for review sites that allow friendships or social connections, how do those topological factors impact the evaluation of user content? These were questions left unexplored by the research.

We would like to explore how concepts such as similarity between users can be used to predict the content of user contributions. For sites such as Yelp or Amazon, where user contributions come in the form of reviews, could similarity between users help to predict if an audience will tend to rate a product negatively or positively? This could be explored using similar features as extracted from user interests in this research and could be broadened to include social information such as friendship networks.

2.2 Evaluating Collaborative Filtering Recommender Systems (Herlocker, Konstan, Terveen & Riedl, 2004)

Recommender systems with collaborative filtering have risen to become the source of most suggested items on sites where users have many options to choose from like e-commerce and movie streaming. In [2], Herlocker et. al. discuss the main considerations when evaluating recommender systems since many implementations consider a wide range of metrics and data sets. While many of these implementations boast the most successful recommender system for their respective products, Herlocker et. al. point to three main considerations (i) recommendation goals, (ii) data set and (iii) accuracy metrics and explores different approaches to each.

Recommendation goals can vary widely based on its service to its users. While the goal for many recommendation services such as Netflix and Amazon is to **find good items** to recommend to users, other recommendation goals include **find all good items**, **recommend sequence**, or **just browsing**. It highlights how its important to consider that some recommender system success can vary widely if you consider what the goal of the recommendation is (ex. looking for a movie to watch versus looking to browse through a news feed). Additionally there should be consideration that a data set is not too sparse and is appropriate to fit the goals of the recommendation.

One of the most insightful parts was the discussion of accuracy metrics. Regardless of how accurate a recommender system claims to be, it may be useless if the accuracy metric incorrectly captures the goal of the recommendation. Some accuracy metrics to consider are **Mean Absolute Error (and related), Classification Accuracy Metrics**, and **Prediction-Rating Correlation**. While all metrics measure some form of accuracy, not all metrics consider the same recommendation goals. For example, if strongly penalizing bad recommendations is a goal, then a mean squared error can be a much stronger predictor of accuracy than mean absolute error. Accuracy metrics should be carefully considered.

Effective and meaningful evaluation of recommender systems is challenging and little work has been done prior to compare and contrast factors that lead to successful recommender systems. The article excels in its comprehensive coverage of different considerations when evaluating recommender systems. The article has reviewed much of the literature related to implementations of recommender systems and notes the potential sources of bias that could make a result seem successful. The questions that remain for us are what are other ways we can augment the success of recommender systems beyond the traditional collaborative filtering methods. Since this article didn't push the frontier of recommender system implementations, we hope to add to the literature and explore new ways other similarity measures can make accurate predictions.

2.3 Random-Walk Computation of Similarities between Nodes of a Graph (Fouss, Pirotte, Renders, Saerens, 2007)

This paper [3] explores the possibility of relating nodes in a movie/viewer database through similarity measures. With this calculated similarity between viewers, between movies, and between viewers and movies as the engine behind a collaborator recommendation system. In particular, the research explored the use of random walks and Markov chains as a measure of similarity between two given nodes. This problem space relates to the topics of random walks algorithms such as PageRank that have been discussed in class.

To measure similarity between two nodes, the researchers calculated the Euclidean Commute Time Distance (ECTD), which is defined as the square root of the average number of hops a random walker would take to get from Node A to Node B and return to Node A. The researchers then proceeded to prove the ability to compute this value via principal component analysis on the graphs Laplacian matrix.

The greatest strength of this research was the thoroughness with which it explored many different scoring algorithms for its recommendation system. The novel ECTD method was compared against 11 other algorithms such as average commute time, k nearest neighbors, maximum-frequency, cosine coefficient, etc. In doing so it introduces many useful methods for gauging the similarity between two nodes. One area the paper admittedly did not tread was the use of weighted edges (i.e. the review rating of movie watchers) and instead opted for use of watched versus not watched edges. Including these weights could have been useful in generating a better approximation of how much a viewer will like a given movie.

We hypothesize our project, a rating predictor, could be modeled similarly to a recommendation system. Our research could take these similarity measures, especially those relating to ECTD, to improve the Yelp restaurant recommendation process. In particular, it would be valuable to rely solely on whether nodes in the graph have rated a restaurant, but exactly what they rated it. By generalizing the research in this paper to consider weighted edges could serve as an even better predictor of how a given user would rate a restaurant.

3 Discussion

Our review of the current research in this problem space has left us with several important notions. In the tasks of rating prediction and recommendation systems, calculating similarity scores between customers (movie-watchers, restaurant-goers, etc.) and between products (movies, online purchases, restaurants, etc.) can greatly improve accuracy. We also learned that there are many nuanced ways of obtaining similarity measures. Fouss et al explored network-based similarity scores such as random walk-based ECTD and k nearest neighbors, whereas Andersen et al focused on qualities of the nodes themselves (overlap in expressed interests).

While [1] explores social graph data to make predictions about ratings, it doesn't use a a recommender system to make there predictions. [2] dives into the literature of recommender systems, but doesn't put forth any new applications of recommender systems. [3] begins the discussion of node similarity, but only explores one method of random walk applications. Notably absent from these studies was the use of social, friendship-based network information between nodes to find similarity to make recommender system based recommendations. We hypothesize that using these implicit measures of similarity in conjunction with explicit social connections would ultimately increase recommendation/ranking prediction accuracy.

4 Project Proposal

4.1 Introduction

Yelp is one of the more unique social networks in the sense that it thrives off of highly qualitative user-inputted reviews. Each one of these reviews, tailored to a multitude of businesses serve as an unbiased platform in which the demographic has an equal voice in ranking and comparing local businesses. Luckily enough, Yelp has disclosed this data to academic institutions in the hope of learning more about the data that they are collecting. While taking a original stance on historical data posed a challenging task, our team saw this as an opportunity to explore options to predict future reviews.

In order to fully take advantage of this network to make future predictions about user reviews for businesses, we have to feed graph data into recommender systems to provide a good gauge of future outcomes. One of the major decisions we had to make in this project was figuring out how to devise these recommender systems, but leveraging techniques in Collaborative Filtering (CF), Markov Chains and some heuristics of our own, we hope to better take advantage of the graph data to make review predictions.

4.2 Data and the Problem

Yelp's Public Academic Data provides us with the reviews of the 250 closest businesses to 30 universities around the United States. In whole, the data set provides us with json-objects for the relevant users (and their reviews) for these 250 businesses. While this large and interwoven data set provides us with a massive repository of reviews to make inferences about the quality of businesses, the current organization of information does necessarily suit our future goals to make predictions about user's future business reviews.

With this information also structured as a json-object, the information wasn't inherently relational but gave us highly-parsable text to store information about users, businesses and reviews. Beyond the "stars" measure that would be used in both the review and business context, the user_id would serve as a join key for both information about the user and review.

4.3 Representing Our Graph

One of the crucial decisions that we made early on was going to be about structural network of our graph. Because we were interested in seeing how the use of the graph would improve our overall recommender accuracy, we had to choose carefully between the many structures that would have served as sufficient options.

We ended representing both of our users and businesses as nodes, and using the reviews that went from users to the specific businesses as the premise for our edges (carrying the weight of the review's star rating). Because of the limited scope of our data (we knew that all users had made reviews for the 250 businesses that surrounded each of the 30 academic institutions), we knew that there was generally going to be a high level of connectedness in the graph. Although the user-to-user connectivity may pale in comparison between the user-to-business connectivity, establishing these connections may later help us generate a "friendship heuristic" for the recommender.

4.4 Predicting User Ratings with Recommender Systems

All user review predictions will be based on recommender rystems. Traditional recommender system approaches such as movie recommendations or similar product suggestions, use the method of Collaborative Filtering. In this approach, users would be measured for how similar they are to other users based off of their actions these users can take, such as rate movies or products. Examples of metrics that have traditionally been used to measure the similarity between two users are the **Cosine Similarity Measure** [6]:

$$sim(x, y) = arccos(r_x, r_y) = \frac{r_x \cdot r_y}{||r_x|| \cdot ||r_y||}$$

and the **Pearson Correlation Coefficient** [6]:

$$sim(x, y) = \frac{\sum_{s \in S_{xy}} (r_{sx} - \bar{r}_x)(r_{sy} - \bar{r}_y)}{\sqrt{\sum_{s \in S_{xy}} (r_{sx} - \bar{r}_x)^2} \sqrt{\sum_{s \in S_{xy}} (r_{sy} - \bar{r}_y)^2}}$$

To go from a similarity metric to a prediction, the following method and equation can then be applied: Let r_x be the vector of user x ratings Let N be the k most similar users to x who have rated item i Prediction of item s for user x [6]:

$$r_{xi} = \frac{\sum_{y \in N} sim(x, y) \cdot r_{yi}}{\sum_{y \in N} sim(x, y)}$$

The way we will extend this research is by calculating user similarity based on properties of the user social graph instead of the above similarity measures.

4.5 Random Walk Computation of Similarity

One possible measure of similarity based on the user social graph is a random walk computation of similarity. The assumption of using this metric is that nodes that can be randomly walked between with higher probability are likely to be more similar. This can be conceptually explained by the reasoning that people who are direct friends or have many friends of friends likely have more similarity that those who are only related by a friend of a friend. Similarly, random walk algorithms with take less time on average to traverse between two nodes.

One such random walk measure of similarity is **Euclidean Commute Time Distance**. $[n(k|i)]^{1/2}$ is defined as the square root of the average number of steps that a random walker, starting in state $i \neq k$, will take to enter state k and then return to state i. Thus, the Euclidean Commute Time Distance is the expectation of this quantity:

$$\begin{aligned} ECTD &= [n(k|i)]^{1/2} \\ &= [m(k|i) + m(i|k)]^{1/2} \\ &= [E[T_{ik}|s(0) = i] + E[T_{ki}|s(0) = k]]^{1/2}. \end{aligned} \tag{1}$$

4.6 Community Similarity

Another measure of similarity is whether or not two nodes are in the same community. Using community detection as well as other data that we have about users in the Yelp data set, we can make recommendations considering nodes in the same cluser as the most similar.

We know that business ratings don't occur in a vacuum. Users in the same community of a user could lead to influencing a user's rating accordingly. For example, members of a community may have visited a restaurant together and during the meal, complained terribly about the experience. As a result, if one member gives the restaurant a low rating, it's likely other users will give the restaurant a low rating as well.

We propose running the same recommender system with collaborative filtering, but strictly considering nodes in the same cluser as most similar and as a result are the main predictors of a users rating for a business. To separate users into respective communities, we can run community detection algorithms such as the **Minimum Cut Method** on the social graph of friends on Yelp.

4.7 Friendship Overlap Similarity

Our final measure of similarity is based on the proportion overlap of friends for two given users. The basis for this is that users with the most similar friend sets may also have a similar set of business reviews on Yelp. This is just another way to use the social network data to derive information about similarities between two users that aren't possible with the user-review information alone.

Given two users x and y, let F_x be the set of all nodes that have an edge to x and $F(y)$ be the set of all nodes that have an edge to y. The friendship overlap similarity can be given by:

$$sim(x, y) = \frac{|F_x \cap F_y|}{|F_x \cup F_y|}$$

4.8 Evaluation

From the recommender system, we will be able to predict ratings that users give to businesses on Yelp. We also have the ground truth data from the Yelp data set of what ratings users actually gave to businesses. From these two set of numbers, we can then evaluate our prediction using known accuracy metrics.

As discussed in (Citation Number for Herlocker Article), Herlocker et. al. point out that an appropriate accuracy metric that aligns with the goal of the recommendation is crucial to the success of the recommender system. One such accuracy metric that makes sense for the goals of our recommender system is the **Root Mean Squared Error (RMSE)** since we don't care as much about small differences in predictions, but we care much more about large differences between predicted and actual ratings. The advantage of using RMSE is that it's a realtively understandable metric since it's clear how error is being measured. Additionally, RMSE has well studied statistical properties that provide for testing the significance of a difference between mean errors of two systems. Let R_p be the prediction rating and R_a be the actual rating. RMSE is given by:

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(R_p - R_a)^2}$$

Another possible adaptation of using RMSE is to only calculate the RMSE on the top k reviews that the user has given. The reason for is that as a service that gives recommendations for Yelp businesses, we care most about recommending what we would consider the best recommendations. It's less important what the middle and bottom recommendations are, as long as our best recommendations are accurate.

4.9 Project Goals and Deliverables

By the end of the quarter we hope to:

1. Implement all algorithms to calculate the similarity measures proposed (Randomized Walk Similarity, Community Similarity, and Friendship Overlap Similarity)

2. Predict user ratings based on similarity calculations

3. Evaluate success of predictions by comparing to ground truth ratings and applying RMSE

References

[1] Anderson, A., Huttenlocher, D., Kleinberg, J., & Leskovec, J. (2012). Effects of user similarity in social media. *In Proceedings of the fifth ACM international conference on Web search and data mining*, (703-712).

[2] Herlocker, J. L., Konstan, J. A., Terveen, L. G., & Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)* **22(1)**, 5-53.

[3] Fouss, F., Pirotte, A., Renders, J. M., & Saerens, M. (2007). Random-walk computation of similarities between nodes of a graph with application to collaborative recommendation. Knowledge and data engineering *IEEE transactions on*, **19(3)**, 355-369.

[4] Chevalier, J, and Mayzlin, D (2006). The effect of word of mouth on sales: Online book reviews. *Journal of marketing research* **43.3**, 345-354. APA

[5] Leskovec, J, Huttenlocher, D, and Kleinberg, J (2010). Predicting positive and negative links in online social networks. *Proceedings of the 19th international conference on World wide web.* ACM.

[6] Leskovec, Jure. "Recommender Systems and Collaborative Filtering." (n.d.): n. pag. Stanford CS 246. Web. 15 Oct. 2015.