Yelp Chanllenge

Tianshu Fan
Xinhang Shao
University of Washington

June 7, 2013

1 Introduction

In this project, we took the Yelp challenge and generated some interesting results about restaurants. Yelp provides data about businesses, reviews, users, and check-in sets in the Greater Phoenix, AZ metropolitan area. The original data were in JSON format. They were parsed and imported into Postgres using JDBC. Several questions were raised by the Yelp online contest, mainly focusing on finding useful information from the data. One question was to predict the rating of a restaurant from the review text only, which was the one we tried. Basic TFIDF method was used, achieving a mean absolute error (MAE) of 1.13 and a root mean square error (RMSE) of 1.49. Some interesting results were generated by queries.

2 Data Importing

The original data are in JSON format, with one JSON object in each line. Some fields are lists, and some are nested JSON objects. To convert to relational schema, a java toolkit called JSON.simple was used to parse JSON format to a flat table. It is a SAX-like parser that processes data in a streaming fashion without using up the main memory. A JSON object is like a map entry, whose value could be retrieved by its name. It can also handle nested JSON objects and lists. Lists are JSONArray, which is a java.util.List essentially. At first, parsed data were written in text files and then copied into database, as in Homework 1. However, the text of reviews caused some troubles. One was the encoding problem. The other was some special characters, such as slash and newline. With JDBC, these problems were never encountered. JDBC also allowed us to update TFIDF score of reviews conveniently. Figure 1 is the ER diagram of yelp data. We mainly focused on restaurants, which have two thirds of total reviews (159,429 out of 229,906). Check-in sets were not used. Some fields were discarded due to lack of information (null for most records) or uselessness in the queries. All businesses are in the greater Phoenix, AZ metropolitan area, so city and state were omitted. One difficulty was that most restaurants belonged to more than one category. At first concatenated strings or arrays were considered for the category field, but due

to searching efficiency, is a relation was used. 10 is a tables were used to represent cuisines from different countries or areas. Some close categories with few restaurants were merged together. For example, Table Japanese has 174 records, merged from 125 Japanese and 94 Sushi Bars from the original data, which means that 45 restaurants have both tags. Some categories were neglected, such as Barbeque and Steakhouses. One restaurant may belong to several such tables, but duplication was minimized after we reassigned restaurants to the merged categories.

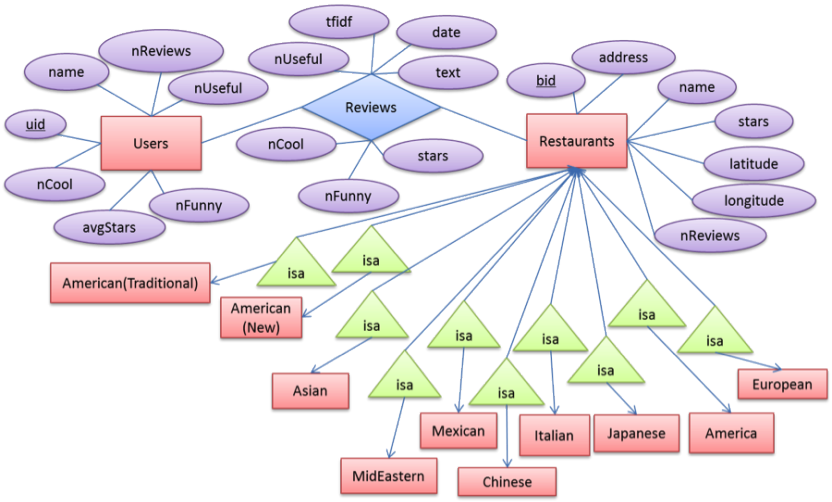

Figure 1: The E/R diagram of the relations

3 Review rating prediction

First, we preprocessing the review text to replace punctuation. The reason for this it to increase the accuracy of the prediction and decrease the variance of the word list at the same time. However, we still keep the facial expression like :), =) and so on. In the J. Martineau and T. Finin[1], it only predicts whether the movie review is more positive or more negative. Our case is more complicated. Second, in order to calculate the tf-idf score, we used the following steps:

1. Import all 5 star reviews to Java, count the frequency of each word using HashMap, sort according to the word count by TreeMap. Manually pick a list of words with highest frequencies, discard stop words, non-relevant words and low frequency words. This is the positive list which includes 183 words.
2. Repeat for 1 star reviews, and get a negative list with 103 words.
3. Use all reviews to find idf of each word in the lists, which is the total number of reviews divided by the number of reviews that word appear. The equation of idf is shown below:

$$idf_w = log\frac{|D|}{|D_w|}$$

4. For each review, count the frequency of each word in the positive list and negative list, as tf with equation below:

$$tf_{w,d} = 0.5 + \frac{0.5 \times f(w,d))}{maxf(w,d), k \in d)}$$

$$tfidf = tf \times idf$$

5. Compute the average of tfidf score of each word in the positive list as P, and the average of tfidf of each word in the negative list as N. Third, use three methods to predict the star. The first method is mentioned in G. Ganu, N. Elhadad, and A. Marian[2]. They talked about sentiment-based text rating using formula:

$$TextRating = [\frac{P}{P+N} \times 4] + 1$$

which gives a score on scale from 1 to 5. However, this method assumes a linear distribution of positive tfidf. The second method is to predict star value based which range of the original star accumulated percentage of each star the positive tfidf percentage it falls in. The method three try to deal more with the case when people use negation of positive value to express negative feeling.

$$if(P_{tfidf} > N_{tfidf}), 3 + 2 \times \frac{P_{tfidf} - P_{min}}{P_{max} - P_{min}}$$

$$if(N_{tfidf} > P_{tfidf}), 3 - 2 \times \frac{N_{tfidf} - N_{min}}{N_{max} - N_{min}}$$

4 Results

4.1 Data inconsistencies

When importing data, we found some inconsistencies among tables. About 1700 users in Table Reviews could not be found in Table Users, so foreign key constraint could not be added. For some users, the total number of reviews he/she wrote calculated from Table Reviews did not match the number of reviews shown in Table Users. Review count was also inconsistent between Table Restaurants and Table Reviews. One reason could be that only Phoenix users were in Table Users, while people from other places may also comment on restaurants in Phoenix. However, its more likely that the database was not updated concurrently, so the information is less valuable. If possible, Yelp should improve the database maintenance.

4.2 Review rating prediction result

The accuracy of the prediction is measured by MAE(mean average error) and RMSE(root mean square error) metioned in F. Li, N. Liu, H. Jin, K. Zhao, Q. Yang and X. Zhu[3].

Table 1: Error of Each Methods

METHOD	MAE	RMSE
Method1	1.13	1.49
Method2	1.64	1.95
Method3	1.25	1.67

The two values for each method are listed in Table 1. As we can see, the first method gives the best result, and method 3 also gives a good prediction. However the method 2 are poor. The results are highly depend on the word list we chose. also plot three figures for each method. The x-axis of the figure is review number, and the y-axis is star value. The blue lines are true star value and the red dot lines are the prediction star value. As we can see, the method 2 has a poor prediction on star 5, and the method 3 has a poor prediction on score 3 where method 1 have good predict on each star. However, there are many cases that the prediction of star bias with true star by 1-3.

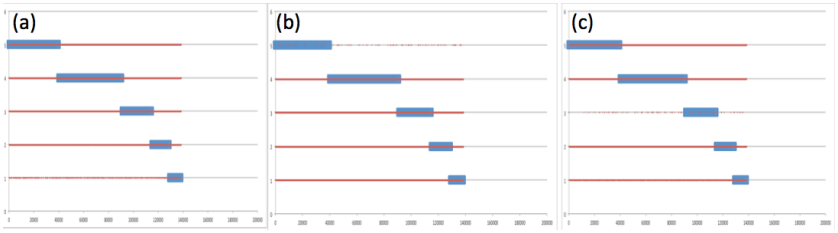

Figure 2: The E/R diagram of the relations

4.3 Results from queries

4.3.1 Spatial distribution of restaurants

There are 4503 restaurants in total. From the spatial distribution in Figure 2, we can see that restaurants are concentrated in small areas and most space has no restaurant at all. The number of restaurants was counted within each area of 0.02 degree latitude by 0.02 degree longitude.

4.3.2 Relationship between rating stars and other facts

Figure 3 (a) (c) (d) show the relationship between the ratings and other facts, like the number of restaurants, the average number of reviews per restaurants, and the number of funny/useful/cool votes. They all have the same distribution. Restaurants with 3.5 stars 4.5 stars are most popular, and have more reviews and review votes.

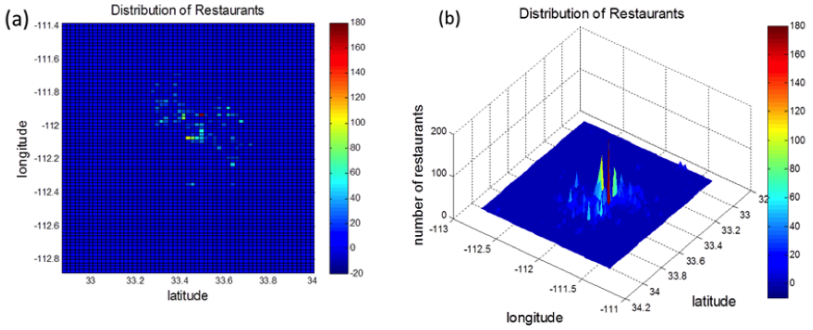

Figure 3: The E/R diagram of the relations

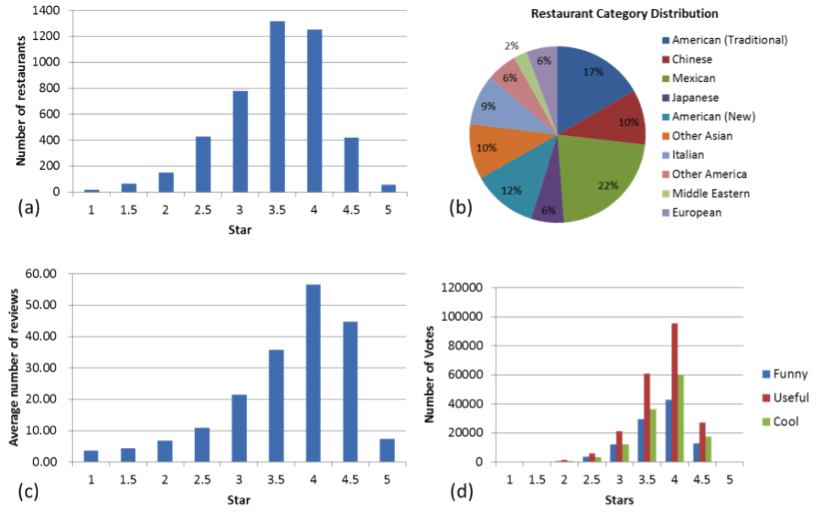

Figure 4: (a) Restaurant distribution in 2D. (b) 3D view of the spatial distributions of restaurants. The height represents the number of restaurants

4.3.3 Facts about restaurant categories

Figure 3 (b) and Figure 4 (a) (c) show the relationship between restaurant categories and other facts. European food and Middle Eastern food have least number of restaurants, yet have the highest average rating. Mexican food is most popular, but the rating is among the lowest. People go to American (New) restaurants also like writing reviews (the average number of reviews per restaurant is the highest), while Chinese restaurants have least number of reviews per restaurant. American (Traditional) restaurants get the lowest rating.

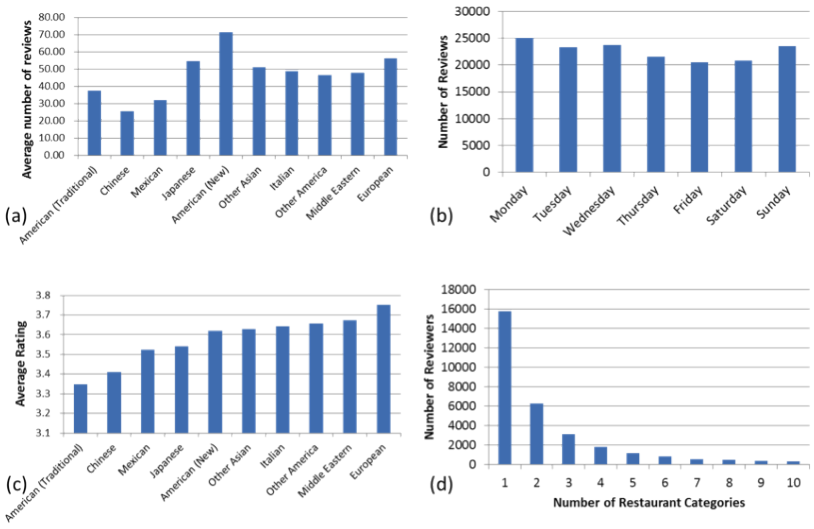

Figure 5: (a) Number of restaurants for each rating. (b) Restaurant category distribution. (c) The average number of reviews per restaurant for each rating. (d) The number of votes for funny, useful and cool reviews for each rating

4.3.4 Review categories

Figure 3(d) is a histogram that shows the distributions of the number of restaurant categories people write review for. The x axis is the number of categories of restaurants that people have reviewed, and the y axis is the number of people who wrote reviews for that number of categories. There are 36,473 distinct users (reviewers) from Table Reviews. 47This query was one of the most complicated. A temporary table was created containing user id and categories. For each record in Table Reviews, if the business id can be found in a category table, insert the distinct user id and the category name combination to the temporary table. Repeat for each category table. Then do group by and count twice on user id and the count for the number of categories, respectively.

4.3.5 Review count for days of the week and months

Figure 3(b) shows the total number of reviews on a certain day of the week. From the check-in information from other groups, people go to restaurants on Thursdays and Fridays most frequently. However, the number of reviews does not vary too much on each day of the week. Friday has the least number of reviews, and Monday has the most. It can be inferred that people usually write reviews in the next one or two days. The same statistics was done for months. The number of reviews for each month is also pretty close, with a maximum of 14,707 in August and a minimum of 11,957 in February.

5 Summary

We imported the Yelp data about restaurants into Postgres, and found some inconsistencies between the original tables. A simple TFIDF method was used to predict the rating stars from pure review text, achieving a MAE of 1.13 and a RMSE of 1.49. Some interesting results from queries were also shown using the information from the data.

References

[1] J. Martineau & T. Finin. Delta tfidf: An improved feature space for sentiment analysis in Proceedings of the 3rd AAAI International Conference on Weblogs and Social Media, 2009, pp. 258-261.
[2] G. Ganu& N. Elhadad& and A. Marian. Beyond the stars: Improving rating predictions using review text content in 12th International Workshop on the Web and Databases, 2009.
[3] F. Li, N. Liu& H. Jin, K. Zhao& Q. Yang &X. Zhu. Incorporating reviewer and product information for review rating prediction. in Proceedings of the Twenty-Second international joint conference on Artificial Intelligence-Volume Volume Three, 2011, pp. 1820-1825.