Recommending the Target Audience to Businesses

Dhiraj Ramnani
dramnani@usc.edu
University of Southern California

Parth Vaghani
vaghania@usc.edu
University of Southern California

Shaily Parikh
shailypa@usc.edu
University of Southern California

Virali Thakkar
vcthakka@usc.edu
University of Southern California

Krish Mehta
krishmeh@usc.edu
University of Southern California

1. Abstract:

Businesses that claim their target audience to be “pretty much everybody” or “anyone interested” exhibit diminishing economic growth. Identification of right customers helps a business in its crucial decision-making process and its success and survival. Finding a target audience is a challenge as it is influenced by various factors. These factors include personal preferences, geographical convenience, and business reputation. Taking into account these factors, we devised an ensemble approach which includes collaborative methods such as ALS(Alternating Least Squares), KNN, and Cross Recommendation to capture user’s personal preferences coupled with Location and Category based recommendation system which captures Geographical Convenience.

2. Keywords:

ALS(Alternating Least Squares), Sentiment Analysis, K-Nearest Neighbors, PageRank, Decision Tree, Cross Recommendation, Collaborative Filtering, Local Experts.

3. Introduction:

Recommendation systems have been utilized by almost every business and have resulted in a Win-Win situation for Customers and Businesses. These systems help in improving user experience and make a business profitable. Yelp aggregates review data from its users and recommend businesses to its users based on various factors which include budget, food categories, location, etc. During the last few years, reviews have become crucial to the success of a business, as every business owner knows that good reviews can boost popularity and profitability, whereas terrible reviews even have the potential of closing businesses down. That’s why it is crucial for businesses to understand the impact of review websites such as Yelp, or TripAdvisor and the role they play for the success or downfall of a business. Targeted advertising has been the prime focus of most of the online businesses. Once the business identifies its Target Audience, the next step is to offer potential discounts only to the Target Audience. From increased sales to improved reputation - discounts may be that one ingredient which can bring business success. The aim of this work is to come up with a system that can find new potential customers for businesses and help them channelize their advertising strategy. Few studies emphasize finding local experts in every area and consider their reviews critical to the business reputation [1]. On the other hand, one of the research focuses upon using check-in information to comprehend customer’s geographical convenience and the distance they are willing to travel for food [3]. According to a survey by BrightLocal [5], an individual is willing to travel at most 17 miles to visit restaurants. Few essential factors to find potential customers include the Friend Circle, location of the business, and business reputation. It is quite challenging to understand a customer’s decision-making process. In our work we have taken into consideration:

- Location of the customer.
- Sentiment detected from the reviews.
- Business Reputation computed through the reviews and similarity with other businesses.
- User’s social network.
- Local Experts of the area and how they influence the customers.

4. Approach:

4.1 About Yelp Dataset

The dataset that we worked upon has four main data tables namely reviews, tips, business, and user. In the dataset, only 5% of pre-processed data belongs to 2004 to 2010 timeframe. So, all experiments are performed on the sampled training data collected from 2011 and 2016. Data from 2017 onwards have been considered as testing data.

For the sake of simplicity, we considered the data of 1000 most reviewed businesses for our experiments.

4.2 Recommendation System

Every person has different preferences. Some people like Asian food while others like Mexican. Some care about taste only while others care about ambiance and services. This problem can be handled by identifying the user's preferences.

We have focused on predicting a score for every user-business combination. And suggesting the new potential users to a business based on the calculated score. We implemented various methods to find out those scores.

A) BaseLine

For our baseline, we have used the following method.

$$Score(u, b) = \boldsymbol{\mu} + stars(u) + stars(b)$$

Here, $\boldsymbol{\mu}$ is the average of average rating of all businesses, stars(u) is the difference between user u's average star rating and $\boldsymbol{\mu}$, stars(b) is the difference between business b's average rating and $\boldsymbol{\mu}$, and Score(u, b) is the predicted score for user u and business b combination.

B) K-Nearest Neighbors

Determined the K users who have similar preferences as user u. Predicted the user's rating for businesses by averaging ratings of K similar users, if possible. We employed Euclidean distance as a metric to compute the similarity between two users.

$$Sim(u1, u2) = \sum_{r \in R} (stars(u1, r) - stars(u2, r))^2 / |R|$$

Here, R is the list of businesses which both the users u1 and u2 have rated. *Sim(u1, u2)* is the similarity measure of both the users. Smaller the value, closer their preferences are.

C) K-Nearest Neighbors with clustering (Type of Food)

Due to the sparseness of data, it is unlikely that two users write a review or rate same business, resulting in an unreliable outcome. To resolve this problem, we formed clusters among businesses based on the categories (assuming it is the most influencing factor) and performed collaborative filtering (method-B) with more abundant information.

D) PageRank

Users can be given more importance based on the following factors:

- The number of reputed businesses he/she visits.
- Activeness of user on Yelp.

We captured these factors using PageRank score and weighted the users accordingly.
(Users & Businesses are represented with 359007 nodes in the graph. Tips & Reviews are represented with 871648 edges)

E) Decision Tree Regressor

To find the pattern between activeness of user and popularity of a business, we decided to train a Decision Tree Regressor to predict the stars for unrated businesses by a user. We used the following features: Average stars of the user, Eliteness, Fans of the user, number of votes for a user, number of reviews for a business, and average stars of a business.

F) Cross-Recommendation

Businesses having a large set of similar users can be in the same geographical region and can have various common characteristics. We split the business pairs into 4 types and give weights based on the size of the set of similar users. For example, we have two businesses B1 and B2 who have 800 customers in common. Now all the users that are customers of B1 but not B2 are recommended to B2 with the weight of business pair. This happens for all pairs of businesses. A customer gets a cumulative score for a particular business based on other businesses that share customers with this business.

G) ALS with Sentiment Analysis

We performed Part of Speech Tagging on the user reviews and obtained adjectives from the review. Finally, we utilized SentiWordNet to compute score for these adjectives and performed aggregation over the scores to obtain average sentiment of the review. The rating is then modified by the sentiment score computed through SentiWordNet. This updated data is used as training data for the ALS model to fetch top users for each business, which improves the rating prediction accuracy.

H) Finding Potential Customers through Local Experts:

We have designed an algorithm that generates a list of Yelp users who are experts in their locality in a particular category. The problem can be formulated as follows: *Given a query q to find local experts in a category c(q) in a location l(q), find a set of users that are knowledgeable in category c(q) and are local to location l(q).* To ensure Topical Authority, we have developed a separate classifier for a few prominent categories which segregated users into two bins: experts and non-experts. For Local Authority, we obtained the home location of the user (Most familiar location). In order to obtain the home location, we clustered the locations of all the businesses, a user has visited, and selected the centroid of the densest cluster as his location. Finally, we generated ranked list of the user-business combination by recommending all businesses that a local expert has visited, to the entire network (friends) of each local expert. We ranked the recommendations based on the distance of the business from the user's location and business reputation (highly reviewed businesses). We performed a specific analysis on finding customers for Mexican Restaurants in Phoenix. We predicted about 300 user-business pairs using the training data and observed that about 20 out of 300 pairings actually visited those businesses in test data.

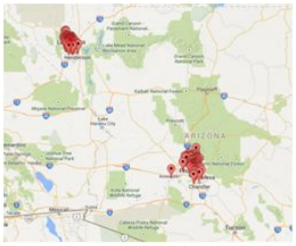


Figure 2.1
User Location Clusters

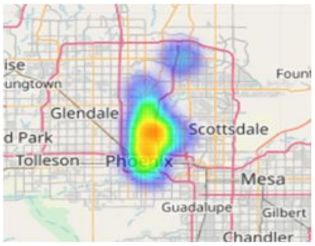
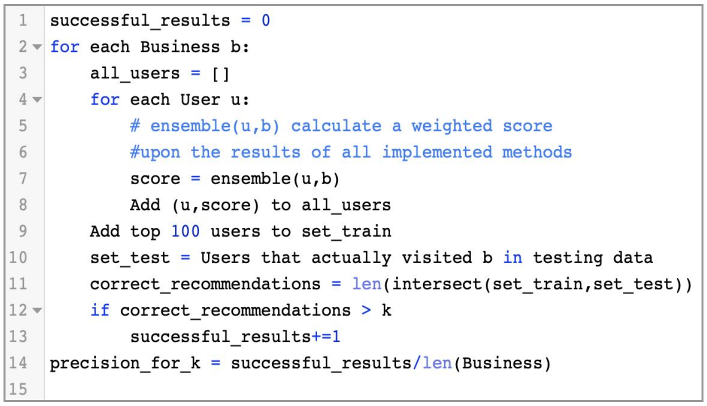


Figure 2.2
User-Restaurants Vicinity

Using the weighted scores from the above A to G methods, we assign an overall score to users for all businesses. Based on the overall score we identify top 100 new potential customers for every particular business.

5. Results and Discussion:

We have derived inspiration from click-through rates in advertising for the evaluation metric. Out of 100 potential recommendation that we make for a particular business, if k of those users actually visited(reviewed) in the testing period then we are considering our recommendation for that business successful. We experimented with k ranging from 1 to 5.



```
successful_results = 0
for each Business b:
    all_users = []
    for each User u:
        # ensemble(u,b) calculate a weighted score
        #upon the results of all implemented methods
        score = ensemble(u,b)
        Add (u,score) to all_users
    Add top 100 users to set_train
    set_test = Users that actually visited b in testing data
    correct_recommendations = len(intersect(set_train,set_test))
    if correct_recommendations > k
        successful_results+=1
precision_for_k = successful_results/len(Business)
```

Evaluation Algorithm

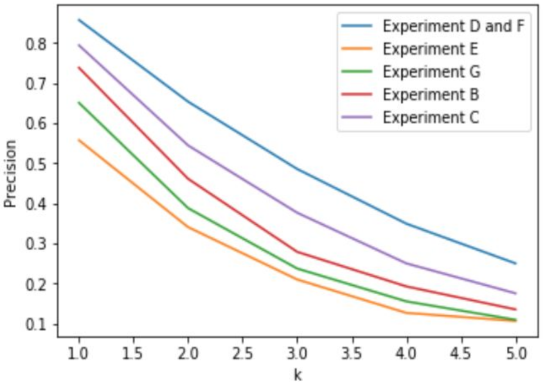


Figure 5.1

Figure 5.1 shows the results for 5 individual experiments. Each line depicts a different experiment for $\forall k$. Cross Recommendation with Pagerank(Experiment D and F) provides better results than others. Interestingly, the improvement from experiment B to C represents fine-grained suggestions based on a category of a business, is an important factor for the recommendation. In experiment C we fine-grained the recommendations for one category but this can be extended for all categories to achieve better results in the future.

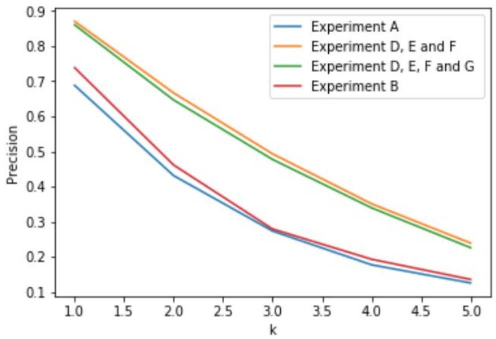


Figure 5.2

Figure 5.2 depicts ensemble approaches which combine different individual approaches. Combining the results of ALS, Cross Recommendation, PageRank, and Decision Tree improves results over individual experiments for $\forall k$ but the best results are observed when combining Cross Recommendation, PageRank and Decision Tree. We get 87% precision for k=1 for this experiment which suggest that out of 100 recommendations at least 1 user visits the business in the test period for 870 out of 1000 businesses.

Experiment	k=5	k=4	k=3	k=2	k=1
A	0.12	0.17	0.27	0.43	0.68
B	0.135	0.19	0.278	0.461	0.738
D, F	0.25	0.34	0.48	0.65	0.85
D, E, F, G	0.22	0.33	0.477	0.64	0.86
D, E, F	0.23	0.35	0.49	0.66	0.87

Table 5.1

We were able to enhance precision by 27.94% from the baseline result for k=1 and by 91.67% for k=5.

Next, we separately performed experiment H for businesses in 3 categories in Phoenix city.

Category	Success Rate	Classifier Accuracy
Mexican	20/300	0.822
American	19/300	0.832
Pizza	18/300	0.812

Table 5.2

According to experiment H, we generated potential customer-business pairs which were not a part of the training set. Out of all the combinations, we filtered pairs by the distance (within 10 miles) and composed a list of 300 pairs ranked by business reputation. For example, in the Mexican category, 20 user-business pairs were successful recommendations.

6. FUTURE WORK:

As we have observed from the above results that category-based recommendations provide better results, we can extend our methods to fine-grain results in every business category. Also, we want to focus primarily on summarising reviews for every business and use important topics from summary as features for better recommendations.

7. REFERENCES:

[1] T. Jindal, Finding local experts from Yelp dataset, https://www.ideals.illinois.edu/handle/2142/78499.

[2] K. Balog, L. Azzopardi, M. Rijke, et al. Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 43–50. ACM, 2006.

[3] R. Li, C. Ju, J. Jiang, W. Wang, et al. CORALS: Who are My Potential New Customers? Tapping into the Wisdom of Customers' Decisions. Yelp Academic Challenge Dataset, 2017.

[4] Y. Koren, Collaborative Filtering with Temporal Dynamics. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 447-456. KDD, 2009.

[5] Bright Local. 2017. Local Consumer Review Survey. https://www.brightlocal.com/learn/local-consumer-review-survey/. 2017.

8. Appendix:

Repository:
https://github.com/24qwerty/Data-Mining-Project
Dataset: https://www.yelp.com/dataset
Data Exploration: Krish Mehta, Parth Vaghani
Experiment A, C: Virali Thakkar
Experiment B: Virali Thakkar, Krish Mehta
Experiment D, F: Shaily Parikh
Experiment E: Shaily Parikh, Parth Vaghani
Experiment G: Krish Mehta, Parth Vaghani
Experiment H: Dhiraj Ramnani
Ensemble Approach: Krish Mehta, Parth Vaghani, Shaily Parikh, Virali Thakkar