Visualization and Analysis of Restaurant Ratings Based on Neighborhoods

Zongrun Li
zli867@gatech.edu

Xuejie Guo
xguo324@gatech.edu

Shuheng Gan
shuheng.gan@gatech.edu

Siying Cen
scen9@gatech.edu

Jinyu Huang
jhuang472@gatech.edu

1 INTRODUCTION

There are many restaurant review sites and customers would like to find references from these websites and rely on online reviews to make decisions. People assume that these reviews are reliable. But reviews are sometimes fake. Yelp also admits 25% of reviews are fake[2]. Besides, people sometimes just have general ideas when they find restaurants such as I would like to eat American food. Recommending neighborhoods which have higher density and ratings of restaurants will help people find fit restaurants.

Visualization and analysis of restaurant ratings based on neighborhoods is a project which dedicates to provide users with fair ratings for restaurants in Atlanta region by collecting and analyzing data from Yelp and using visualization methods based on neighborhoods. Users can find accurate and easy-understanding information such as rating distribution, density of restaurants from our interactive figures. They can also choose specific neighborhood, restaurant category and price range to find restaurants.

2 SURVEY

Ratings and visualization methods affect customers' behaviors. For ratings, Luca proved the causality effect of Yelp ratings on the revenue of restaurants with regression approach. Additionally, the impact of ratings is correlated with many other features such as number of reviews and certified reviewers[12]. To provide fair ratings, Jiang combined NLP with K-medoids clustering to cluster Yelp's review which may make us find different reviews' features[7]. Jindal and Liu detected duplicate reviews by using Jindal's detection model[9]. Xie et al., Lim et al. and Mahmudur et al. found that fake reviews used to be generated in a very short period of time[11, 15, 20]. Mahmudur established Marco system to catch fake review in Yelp which based on RSD, ARD and FRI modules[15]. And Mahmudur et al. also used another method based on CoReG, RF, IRR and JH algorithms to detect fraudulent behaviors in Google Play[19]. After recognizing the review spikes, Andrew et al. provided a mathematics method to detect abnormal ratings based on Bayesian reputation systems [16]. Li et al. and Oh et al. designed algorithms based on reviewers' attributes[6, 14]. Li et al. considered reviewers' IP and features [6]while Oh et al. adjusted the reputation based on the confidence of customer ratings which has been calculated based on customers' activity, objectivity and consensus[14]. Masound's compared different recommendation algorithms and provided us insights about how to choose fair recommendation algorithms[13]. Kim et al. dedicate to find a clear criteria for evaluating users in Yelp by using logistic regression, PageRank and SVM algorithms[10].

All the algorithms for detecting fake reviews have shortcomings because fake reviewers will change their methods to avoid their reviews be deleted based on fake detection algorithms. Some methods dedicate to find unreliable reviewers and some methods dedicate to find review spikes to avoid fraudulent reviewers. We decide to combine these two ideas to provide fair ratings.

For visualization methods, Yang et al. found that presenting information in a hierarchical way makes it easier to capture information[21]. And Jin et al. proved that labeling restaurants with keywords of reviews will help our users know the general comments and make decision efficiently[8]. Zhang et al. used DBSCAN to identify the center of their check-in area and then analyze the overall check-in probability over spatial distance. They found out that most users' check-in travel is no more than 2 km[1]. Sun et al. created Voronoi diagram to cut the area into regions of venues, then run A Multidirectional Optimum Ecotope-Based Algorithm to analyze the influence of geographical features

to venue's rating and revealed that the spatial features will influence the rating of certain kind venues[17].

These researches use mathematics methods to classify and divide areas. But actually, people are unfamiliar with these newly defined area. The natural way to divide cities are neighborhoods.Since that, we will consider about visualizing restaurants' distance and distribution information based on neighborhoods to help customers make choices easily and to avoid the shortcomings.

3 PROPOSED METHOD

3.1 Intuition

3.1.1 Innovations.

1. Restaurants are classified based on neighborhoods. It is a friendly way for costumers who only has general ideas to find a place to eat for a specific category of restaurants.

2. The restaurants' ratings are not simply based on all ratings' average. We use algorithm to detect days which have fraudulent behaviours and fake reviews in these days.

3. The project visualizes results from our analyses. Restaurants with different ratings and price have markers in different colors and sizes. Customers can also filter restaurants easily based on interactive website.

3.2 Approaches

3.2.1 Yelp Data Collection.

Yelp Fusion API and Google API are useful tools for getting restaurants' and users' data. However, Yelp API can only return up to 1000 results at this time and it only returns three selected reviews for each restaurant and google API can only return 20 results. Since that, we decide to use yelp data and write a program to scrape the users' and restaurants' data from Yelp web page.

We separate data extraction into two main steps: web crawl and web analysis. We first send the http request to get server response, then preprocess the scrapped html file with Beautiful soap. In order to get restaurants' and users' data, we use the framework described below. To prevent being block, we set timer and use random user agent.

3.2.2 Feature deduction.

The original data scratched from Yelp needs cleaning before being processed. For example, we observe duplicated notations of restaurant type, which is expected to be a significant predictor of ratings. In order to merge the similar notations, we will use hierarchical clustering method.

The data we have includes many features of each specific restaurant. In order to identify the key features in explaining the ratings, we will build up a linear regression model first and carry out LASSO feature selection. This will give us a hint as to what parts of information to show and the order they should be displayed on our interface. What's more, we will use k-fold cross validation to select the best model. Based on the selected model, we can obtain a predictive model with ratings as response.

3.2.3 PNPoly.

To efficiently offer users ideal restaurant choices and improve the entire dining experience, we decided to group restaurants into smaller and more specific zones. We decided to use neighborhood unit because the planning and economic development between neighborhoods are quite different[4]. After we collected the restaurant coordinates and neighborhood borders data, we transferred this to a classical point-in-polygon (PIP) problem, i.e. how to decide a given point is inside or outside of a polygon. Crossing number algorithm and winding number algorithm are two common solutions to PIP problem[3, 5, 18]. Since crossing number algorithm is very straightforward and the other one is preferred for a nonsimple closed polygon (e.g. one that overlaps with itself, which is not suitable for our 2D map situation), we chose the former method. The logic of crossing number algorithm is: if a point is inside of a polygon, any ray from this point will have odd intersection(s) with edges of the polygon; if outside, even (include 0) intersections. The Franklin's point inclusion in polygon (PNPOLY) algorithm[3] will test each edge of a polygon and count the intersection number when there is at least one. To avoiding redundancy, it always chooses a horizontal ray parallel to the positive x axis and pointing to the right of a point. Thus, when a point is on the left boundaries, it is considered to be inside; on the contrary, it is outside.

3.2.4 Fake Review Detection.

To detect fraudulent behaviours, we assume that fraudulent behaviours will continue in a relative short time(one day) and fake reviews happen on these days. Since that, we will first detect days with higher numbers of positive reviews than normal and then detect the fake reviews in these days. To detect abnormal days, Mahmudur Rahman1 et al. indicate that when the reviews number greater than 1/7 of all reviews number before that day, the ratings will obviously increase[15]. So, we will first select abnormal days using that condition. For fake reviewers detection, we assume there are enough fake reviewers on abnormal days and we also assume elite members are reliable. We use this assumption to label reviewers and use reviewers' attributes which have related to reviewers' credibility such as friends, photos, etc.[10]. We use Random Forest to train the model and use the model to judge whether a reviewer reliable or not. After judgement, we will delete unreliable reviewers' ratings on abnormal days and calculate the ratings as our result.

3.2.5 Recommendation.

We will use linear regression algorithm to evaluate correlations between restaurants features and ratings. To predict restaurant rating, features such as location, neighborhood, total review numbers, opening hours and category will be extracted from dataset; linear regression algorithm will be performed to create rating prediction model. Training and testing dataset are separated from the whole dataset. Once the model is constructed, we can figure out most relevant features to rating, so that we can optimize our filters in the interactive interface.

3.2.6 Visualization.

The website has two pages. Main page which includes a main-plot and a subplot and an About page which includes some information and user manual for the project. For the main page, the main plot is based on google map API and JavaScript. The Atlanta map are divided by different neighborhoods. Different neighborhoods have different colors based on neighborhoods' ratings. Users can select neighborhoods to get more information. The information displays on the right side of website. Restaurants with different ratings and price have different markers with different colors and sizes. Users can select these markers to have more information about restaurants. The subplot is based on d3.js. The subplot provide users with restaurants' categories and price selector. Users can filter restaurants by using subplot. For any users' operations, the recommendations for filtered data will display on the subplot.

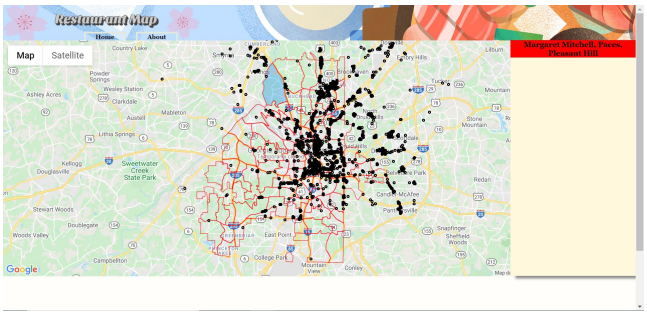

Figure 1: Website Skeleton

4 UPCOMING EXPERIMENTS/ EVALUATION

Our project includes three experiments now. First is fake reviews detection. Second is restaurants information classification. Last is restaurants recommendations.

For first experiment, we will answer which days have fraudulent behaviours and who are reliable reviewers on these days. Based on each days reviews number for each restaurants, we will use spike detection method to find abnormal days. And we assume fraudulent behaviours happened only on these days. To make these days ratings reliable, we will use Random Forest to find who are reliable reviewers. For Random Forest, we will label elite users as reliable users and the others are unreliable. The dataset is from abnormal days and attributes are from data we scratched such as reviewers' friends number, reviews number, photos, etc.. We will put 70% of data to train the model and 30% to evaluate the model.

For second experiment, we will answer the restaurant is belong to which neighborhood and what is the restaurant's category. To classify restaurants based on neighborhoods, we will use PNPoly algorithm. We will provide the program with three-dimension array. Each elements in the array is vertexes of neighborhoods. Each vertexes latitude and longitude are provided. We will also use two-dimension array to store restaurants' latitude and longitude positions. Using PNPoly and

these data, we will give each restaurants a neighborhood attribute. To classify restaurants based on categories, we have analyzed data we have and find there are 213 kinds of restaurants. Some categories are subsets of others. We will base on the relationships between categories to narrow down the category number.

For the restaurant recommendations, we will answer which neighborhood or restaurant is the best for having food and which neighborhood or restaurant is the best for having food for a specific category and specific price range. We will use neighborhoods' attributes and restaurants' attribute and restaurants' ratings to do regression analyses. After analyzing, we will find which attributes are related to ratings and our recommendations are based on high correlation attributes. Recommendations for specific restaurant category and price range will be based on filtered data and interactive website functions.

5 DISTRIBUTION OF TEAM MEMBER EFFORT

5.1 Effort

All team members will contribute similar amount of effort

- `Zongrun`: Building HTML and CSS skeleton; completing main-plot and implementing interactive functions; implementing reliable ratings algorithm.
- `Xuejie`: Using Yelp API to retrieve data; building up scratcher to scratch data from website; design data structure and analyzing data.
- `Shuheng`: Implementing PNPoly algorithm; using PNPoly algorithm to classify restaurants; completing subplot based on d3.js.
- `Siying`: Implementing Random Forest algorithm; implementing recommendation algorithm; completing subplot based on d3.js.
- `Jinyu`: Building up crawler to scratch data from website with Xuejie; cleaning data; design data structure and analyzing data;

5.2 Plan

5.2.1 *Completed Plans.*

- `Feb.24 - Mar.8`: Decided the website layout and data structure. Discussed for algorithms. All members participated in this part.
- `Mar.8 - Mar.19`: Xuejie collected data from Yelp and Siying collected data from google map. Zongrun built up HTML and CSS skeleton.
- `Mar.19 - Mar.30`: Shuheng decided to use the PNPoly algorithm to classify restaurants. Xuejie tried to build up scratcher. Siying and Jinyu decided the algorithm for recommendation. Zongrun decided the algorithm for fake review detection.
- `Mar.31 - Apr.3`: Did midterm check and wrote progress report. All members participated in this part.

5.2.2 *Future Plans.*

- `Apr.3 - Apr.10`: Xuejie will build up crawler and scratch data from Yelp's websites. Shuheng will implement PNPoly algorithm. Siying will implement Random Forest algorithm. Zongrun will implement algorithm for detecting spike days.
- `Apr.11 - Apr.18`: Jinyu will clean the data scrathed. Jinyu and Xuejie will analyze data and provide the data in final format. Siying and Shuheng will complete subplot. Zongrun will complete interactions between subplot and main-plot.
- `Apr.19 - Apr.21`: We will write final report and presentation slides. We will also finish the About page in our website. All members participated in this part.

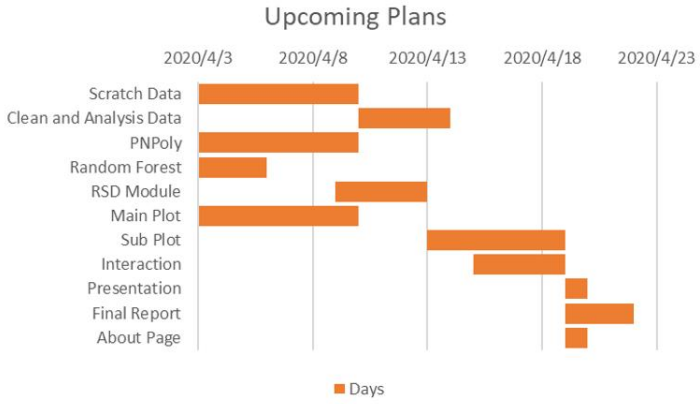

Figure 2: Upcoming Plans

REFERENCES

[1] 2016. *WWW '16: Proceedings of the 25th International Conference on World Wide Web.* International World Wide Web

Conferences Steering Committee, Republic and Canton of Geneva, CHE.

[2] BBC. [n. d.]. Yelp admits a quarter of submitted reviews could be fake. www.bbc.co.uk/news/technology-24299742.

[3] W Randolph Franklin. 2006. Pnpoly-point inclusion in polygon test. http://www.ecse.rpi.edu/Homepages/wrf/Research/Short_Notes/pnpoly.html. Accessed: 2020-03-31.

[4] Atlanta Government. [n. d.]. Neighborhood Gentrification Pressure Areas. https://www.atlantaga.gov/home/showdocument?id=33833. Accessed: 2020-03-28.

[5] Eric Haines. 1994. Point in polygon strategies. *Graphics gems IV* 994 (1994), 24–26.

[6] Bing Liu Xiaokai Wei Huayi Li, Zhiyuan Chen and Jidong Shao. 2016. Spotting fake reviews via collective positive-unlabeled learning.. In *2014 IEEE international conference on data mining*.

[7] Yimin Liu Jiang, Renfeng and Ke Xu. 2015. A general framework for text semantic analysis and clustering on Yelp reviews.

[8] Wei Jin, Hung Hay Ho, and Rohini K Srihari. 2009. OpinionMiner: a novel machine learning system for web opinion mining and extraction. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1195–1204.

[9] Nitin Jindal and Bing Liu. 2008. Opinion spam and analysis. In *Proceedings of the 2008 international conference on web search and data mining*. 219–230.

[10] Caroline Kim, Gordon Lin, and Honam Bang. 2015. Discovering Yelp Elites: Reifying Yelp Elite Selection Criterion. *University of California-San Diego* (2015).

[11] Ee-Peng Lim, Viet-An Nguyen, Nitin Jindal, Bing Liu, and Hady Wirawan Lauw. 2010. Detecting product review spammers using rating behaviors. In *Proceedings of the 19th ACM international conference on Information and knowledge management*. 939–948.

[12] Michael Luca. 2016. Reviews, reputation, and revenue: The case of Yelp. com.. In *Harvard Business School NOM Unit Working Paper*. 12–16.

[13] Bamshad Mobasher Robin Burke Mansoury, Masoud and Mykola Pechenizkiy. 2019. Bias disparity in collaborative recommendation: Algorithmic evaluation and comparison.. In *arXiv:1908.00831*.

[14] H. Oh, S. Kim, S. Park, and M. Zhou. 2015. Can You Trust Online Ratings? A Mutual Reinforcement Model for Trustworthy Online Rating Systems. *IEEE Transactions on Systems, Man, and Cybernetics: Systems* 45, 12 (Dec 2015), 1564–1576. https://doi.org/10.1109/TSMC.2015.2416126

[15] Mahmudur Rahman, Bogdan Carbunar, Jaime Ballesteros, and Duen Horng (Polo) Chau. 2015. To catch a fake: Curbing deceptive Yelp ratings and venues. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 8, 3 (2015), 147–161. https://doi.org/10.1002/sam.11264 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/sam.11264

[16] Mahmudur Rahman, Mizanur Rahman, Bogdan Carbunar, and Duen Horng Chau. [n. d.]. *FairPlay: Fraud and Malware Detection in Google Play*. 99–107. https://doi.org/10.1137/1.9781611974348.12 arXiv:https://epubs.siam.org/doi/pdf/10.1137/1.9781611974348.12

[17] Jorge David Gonzalez Sun Yeran Paule. 2017. Spatial analysis of users-generated ratings of yelp venues. *Open Geospatial Data, Software and Standards* 2 (2017). https://doi.org/10.1186/s40965-017-0020-9

[18] Dan Sunday. 2012. Inclusion of a Point in a Polygon. http://geomalgorithms.com/a03-_inclusion.html. Accessed: 2020-04-01.

[19] Andrew Whitby, Audun Jøsang, and Jadwiga Indulska. 2005. Filtering Out Unfair Ratings in Bayesian Reputation Systems.

[20] Sihong Xie, Guan Wang, Shuyang Lin, and Philip S Yu. 2012. Review spam detection via temporal pattern discovery. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 823–831.

[21] S. J. Woo Y. J. Ah, K. H. Jeong and L. S. Ho. 2011. Visualization of restaurant information on web maps.. In *The 5th International Conference on New Trends in Information Science and Service Science, Macao*.