Improving Business: Mining and Summarizing Check-in and Customers Review

Cheng, Yu
School of Information Sciences, University of Pittsburgh
135 North Bellefield Avenue, Pittsburgh, PA 15260
yuc70@pitt.edu

Xi, Peng
School of Information Sciences, University of Pittsburgh
135 North Bellefield Avenue, Pittsburgh, PA 15260
pex7@pitt.edu

Chen, Shuli
School of Information Sciences, University of Pittsburgh
135 North Bellefield Avenue, Pittsburgh, PA 15260
shc107@pitt.edu

ABSTRACT
In this paper, we summarize some features of business and mine review text of customers to show that how the information extracted from our work can improve the success of business. We present four methods mainly to analyze what make highly-rated businesses be sought after by users: (1) Feature selection for filtering important business attributes. (2) Visualization of business stars and check-in. (3) Text mining of review. (4) Network mining of co-rated business-business. The dataset is sourced from Yelp and it includes information about local businesses, reviews, users and check-in. The large amount of data benefits us to retrieve and solve some small but pragmatic problems. And we extend our findings and pose several interesting insights which could be useful to business owners.

General Terms
Measurement, Design, Human Factors.

Keywords
Feature selection, visualization, text mining, word cloud, network.

1. INTRODUCTION
One of the critical features of Yelp is that different users can share information of various businesses with each other. Users can find the useful information and determine whether or not to eat, drink or entertain in a certain business based on what others said. On the other hand, Yelp also provides a profound effect on the success of businesses. We can find out what kind of features does a successful business should have and what aspects users care about mostly. Business owners should know the important information to improve their services. This paper focuses on businesses in Pittsburgh and the overall rating is in forms of stars.

In this dataset which has high-dimensional data, to model the important attributes of business, we must understand the dimensions of their features. In this paper, the target is to present what kind of business is welcomed by users. We achieve our goal in the following four aspects: (1) Select the most important attributes of business by analyzing their different features and determine what aspects of a business are most concerned by users. (2) Use check-in to analyze business services and plot the result to maps, which give an excellent visualization. (3) Use the review to do text mining and see the most common words in review and the importance of different terms. (4) Create a network to present the information of the relationship of business and users, and then identify the most important business in different methods. We combine these four aspects to give suggestions for business to maintain and increase their popularity of communities.

There are some benefits of combining these methods that predict what fields business can improve to cater to most of users: (1) It is comprehensive for business owners to know well the market situation. (2) The method enables a more accurate analysis of business environment and customers' preferences that are related to business quality. (3) It provides business owners with various choices to enhance their performance in competition because the analysis includes multi-aspects.

2. DATASET
Yelp Academic Dataset [1] provides the business profiles, user reviews over 10 cities across 4 countries. It covers several files and each one is composed of a single object type (business, users, etc.) and one json-object per-line. We convert them into csv files using python [2], and extract part of businesses in Pittsburgh, including 2,724 business, 206,321 review, 366,715 user, and 45,166 check-in.

2.1 Missing data
Although there are a lot of attributes presenting each business, most of them are missing in object business. Fig. 1 shows missing values of 100 businesses in Pittsburgh. Each line stands for a feature of business and missing value is represented in white block. The dataset has 50% missing values and basically businesses have similar filled data. Most of business attributes are categorical values which we count and proportionally fill into dataset as missing data. For example, if there are 1,000 missing value "NA", 400 "True", and 100 "False" of business attribute Outdoor Seating, and the probabilities of occurrence of "True" and "False" are 75% and 25% respectively regardless of the missing data. So we can replace the 1,000 "NA" with 750 "True" and 250 "False". The cases of other attributes are processed in the same way.

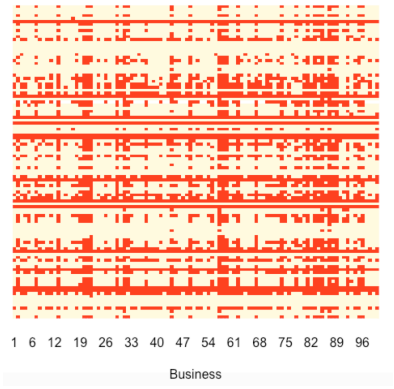
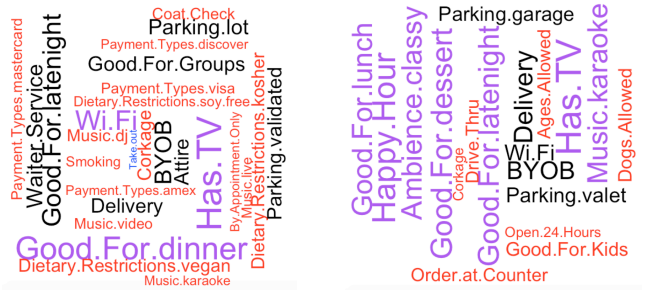


Fig. 1: Missing Values (marked in white) of 100 businesses

3. THE PROPOSED TECHNIQUES

3.1 Important Features Identification

There are 78 attributes which describes some features of each business in Pittsburgh, e.g., hours, parking availability, and ambience. We use feature selection technique to obtain important attributes in order to predict business stars by extracting the key words from different attributes of businesses which have higher rating star. The weighted importance of attribute is represented in word cloud (Fig. 2), which display the most important information in a beautiful, informative image that communicates much in a single glance in an intuitive way. The words in larger fonts stand for more important attributes for business stars. Although restaurants and bars are food businesses and they have some common important attributes such as Has.TV and Good.For.latenight, there are many differences between them. For example, Good.For.dinner is one of important factors of restaurants means that consumers usually care about if restaurants are good for dinner. While consumers go to bars weigh classy ambience, good for lunch or dessert, happy hour and music more than others so that they rate business stars.



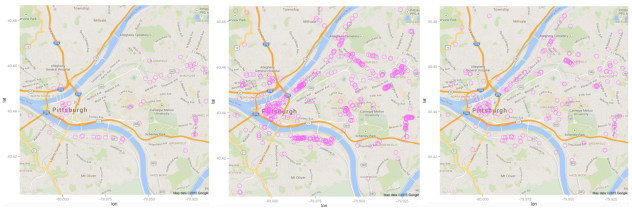
Restaurants **Bars and Cafe**

Fig. 2: Important Attributes for businesses

3.2 Business Visualization

3.2.1 Business Stars

Rating star is an important indicator to business's success and it is the most important influence on users' judgement. In fact, thereis an economic research has shown that star ratings are so critical to the Yelp businesses that even an extra half-star allows restaurants to sell out 19% more frequently [10]. Stars are correlated with many factors such as services, ambience, and availability of parking lot. And Yelpers usually rank stars and filter restaurants, bars, pet stores or some other businesses. Here we distribute businesses in Pittsburgh into geographical map according to the rating stars (Fig. 3). We can see from the following graphs that although there are many more businesses of star 4 compared to star 2 and star 5, the distributions have few differences across the area, so location is an insignificant factor to rate business in the area.

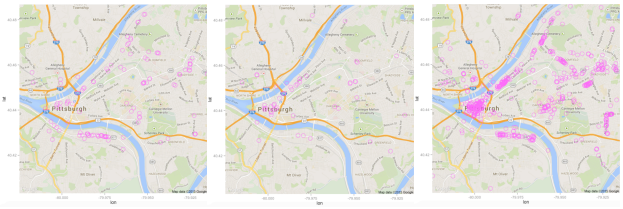


star = 2 **star = 4** **star = 5**

Fig. 3: Geographical distribution of businesses by star

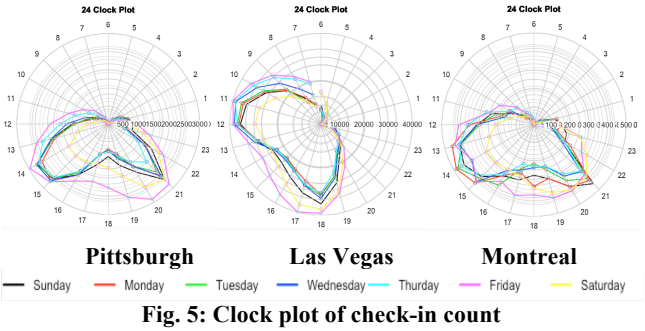
3.2.2 Business Check-in

The information of business check-in contains check-in count, days and hours for each business. It can be used to analyze business services and users' behaviors. Fig. 4 plots the distributions of businesses into geographical map by check-in hour. Comparing with graphs which are generated according to the check-in count, we can come to a conclusion that there are many more users consuming in the evenings than the hours in the mornings and afternoons. In addition, it addition, the count of check-in people in different hours can be plot in the clock diagram as in Fig. 5. We can compare the numbers in different cities. A point on the radius of one concentric circle is represented of the number of check-in at a specific clock. It can be concluded that there are almost the same check-ins from Sunday to Thursday in Pittsburgh, and the same to Las Vegas and Montreal. And on Fridays, users consume more often but check in less on Saturdays. Besides, the information of check-in in Pittsburgh and Las Vegas are similar and it implies that users' behaviors are closer in comparison with those in Montreal which is in Canada.



9:00 - 12:00 **12:00 - 18:00** **18:00 - 23:00**

Fig. 4: Geographical distribution of businesses by check-in hour



Pittsburgh **Las Vegas** **Montreal**

Fig. 5: Clock plot of check-in count

3.3 User Classification

For improving business, it is better to know something about the information of clients, namely users under this circumstance. To determine whether a certain user is an information sender or an

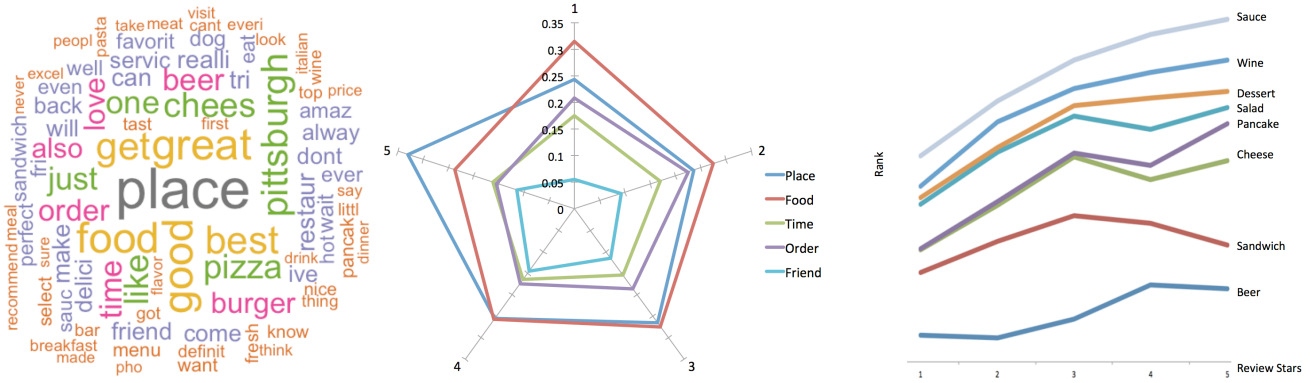
information receiver and whether a certain user is a social people, business can provide information senders or social people with some discounts and coupons to encourage them share more comments and experience with their friends.

So we cluster users using k-means with k = 3 based on user features such as review count, fans, compliment count, vote count and friends count, and compute mean (centroid) in Euclidean space:

$$\mathbf{m}_j = \frac{1}{|C_j|} \sum_{\mathbf{x}_i \in C_j} \mathbf{x}_i$$

$|C_j|$: the number of data points in cluster C_j

The distance from a data point x_i to a cluster centroid m_j is computed with



(a) Word cloud from review of star = 5 **(b) Important factors - review stars** **(c) Food-review stars**

Fig. 7: Significant terms from review text

$$dist(\mathbf{x}_i, \mathbf{m}_j) = \|\mathbf{x}_i - \mathbf{m}_j\| = \sqrt{(\mathbf{x}_{i1} - \mathbf{m}_{j1})^2 + (\mathbf{x}_{i2} - \mathbf{m}_{j2})^2 + \cdots + (\mathbf{x}_{ir} - \mathbf{m}_{jr})^2}$$

Fig. 6 represents the users clustered based on review count and vote count. The users reviewing frequently maybe vote at a wide range of number. While overall, users have more friends and fans varying they review more businesses. It implies that those users are probably social people.

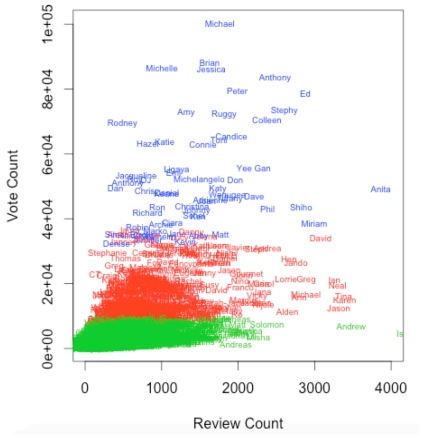


Fig. 6: User Clustering vote count vs. review count

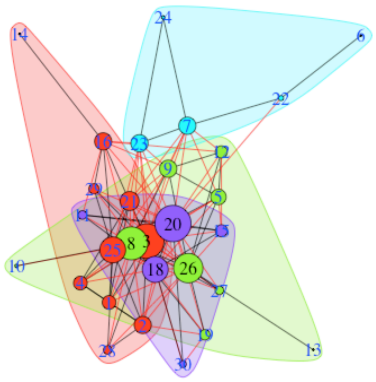
3.4 Review Text Mining

Users wrote reviews have become an important element for business to get the feedback from customers, like what users like or dislike so that business owners can improve their service or products according to different opinions. In this part, we extract 10,247 out of 206,321 reviews to analyze what users care about mostly in terms of different restaurants in Pittsburgh. And the reviews are divided into 5 parts based on stars from 1 to 5. Facing large amount of review text, business owners cannot get the point quickly while a text visualization would speed up the process of getting feedback. So here we use word cloud technique which is an explicable visualization tool to plot important words. Fig. 7 (a) shows the frequent terms (occurrences is greater than 80) used in review text of star 5. The terms are usually single words and the importance of each word is shown with different font sizes and colors. The more frequent a word is, the bigger it gets in size. Place and food are always mentioned in review text for all businesses, which means these two attributes of business are significant for all the users. It can also be concluded from Fig. 7 (b) which shows the occurrences of five critical factors place, food, time, order and friend. Basically, place and friend occur more often in review text of large stars 4 and 5, while food and order frequently appear in review of stars 1 and 2.

From the result of review mining, food is key element for users to rate restaurants as we expect, so we filter some food such as cheese, dessert, salad, sandwich, pancake, beer, wine and sauce and represent the accumulated frequencies by stars in Fig. 7 (c). The stacked lines are independent of each other. Users refer to dessert, pancake, cheese, wine and beer more varying rating higher restaurants, but the contrary is the case to salad and sandwich.

3.5 Business-Business Network

The business-to-business network represents the relationship among businesses which are correlated with users. Businesses are associated with each other when there are common users who post review for them. We build a structure to consolidate users and business from review dataset, and generate a network for co-rated businesses in Fig. 8. This network shows top 30 businesses in Pittsburgh, which are most frequently reviewed and the edge weights > 10, which means that two businesses have a link if they are reviewed by at least 10 common users.



1	Enrico's Tazza D'oro Cafe & Espresso Bar	16	Lidia's Pittsburgh
2	Whole Foods Market	18	Tram's Kitchen
3	Casbah	20	Tessaro's
4	Quiet Storm Vegetarian & Vegan Cafe	21	Thai Cuisine
6	Beto's Pizza & Restaurant	25	Soba
8	D's Six Pax & Dogz	26	Harris Grill
9	Olive Or Twist		

Fig. 8: Co-rated business-business network and modularity-based community

It can be easily to identify from the following graph that these five business: Casbah, D's Six Pax & Dogz, Tessaro's, Soba, Harris Grill and Tram's Kitchen which have the highest centralities are reviewed by the most common users. And these businesses are modularized into four different communities and in each community the businesses are correlated with each other closer than other businesses. The three communities that Casbah, D's Six Pax & Dogz, and Tessaro's belong to respectively are overlapped and much relevant among them. Based on above observations, we can infer that those co-rated businesses can be cooperative partners or potential competitors in the communities and the network.

4. FUTURE WORK

Business attributes are very useful information to cluster businesses and predict what successful factors are. We have identified the important attributes and features which can be utilized by businesses to improve their services or products. However, there are several ways to improve our work in the future. For example, attributes and features can be extended to predict business income and correlate them with review text to refine the importance of the features in future work. And another insight is that the business-business network will be a great way to build a recommender system to different users. The business frequently co-rated by consumers could be recommended to users who have similar consuming behaviors. We can summarize the user profile to cluster them in some useful way to get their tastes of different businesses and append more information on the clustered layout. And do some sentiment analysis of review information and apply NPL technique to process the review so that we can better understand the users' preferences. Furthermore, for those who are not friends or fans but share reviews of the same businesses in Yelp, it is a good idea to connect them in social circles.

5. CONCLUSION

To explore interesting and useful insights for business improvement, we firstly analyze the important attributes of various businesses in Pittsburgh based on Yelp Academic Dataset, and extract some important features and then rank them in a word cloud. Secondarily, business distributions in geographical map visualize the implications from stars and check-in information, according that we can summarize some critical information towards the time that users check-in most frequently. But the location seems not to be such an important factor for business success as people expect. Finally, we build a business-business network based on co-rated users and give some hints to businesses how to interact with other partners or competitors. Through what we find in the above four aspects, we can provide business owners with some useful suggestions to maintain their popularity and improve the quality to cater to more users. Overall, business is an interesting search object, although large amount of data that Yelp provides is widely used in data analysis, our work focuses on some practical problems for business improvement.

6. ACKNOWLEDGMENTS

In this project, our thanks to Dr. Lin, Yuru for giving us a lot of suggestion on how to refine the proposed objectives and apply data mining technology.

7. REFERENCES

[1] http://www.yelp.com/dataset_challenge

[2] https://github.com/Yelp/dataset-examples

[3] Huang, J., Rogers, S., Joo, E. 2013. Improving Restaurants by Extracting Subtopics from Yelp Reviews. University of California, Berkeley.

[4] Hood, B., Hwang, V., King, J., Inferring Future Business Attention. Carnegie Mellon University.

[5] Wang, J., Zhao, J., Guo, S., North, C. 2014. Clustered Layout Word Cloud for User Generated Review. Virginia Tech and University of Toronto.

[6] Hu, M., Liu, B. 2004. Mining and Summarizing Customer Reviews. Department of Computer Science. University of Illinois at Chicago.

[7] Soups, R. 2015. Yelp Dataset Challenge is Doubling Up! DOI = http://engineeringblog.yelp.com/2015/02/yelp-dataset-challenge-is-doubling-up.html

[8] Martin, T., Kita, R. Good Food Bad Service Restaurants. DOI = http://www.goodfoodbadservice.com/

[9] Hung, K., Qiu, H. 2014. Yelp Dataset Challenge 2014 Submission. University of California, San Diego. DOI = http://kevin11h.github.io/YelpDatasetChallengeDataScienceAndMachineLearningUCSD/#show-naive-bayes-math

[10] Linshi, J. Personalizing Yelp Star Ratings: a Semantic Topic Modeling Approach. Yale University.

[11] Anderson, M., Magruder, J. 2011. Learning from the Crowd. *The Economic Journal*.