Yelp Visual Recommender

Alankrit Shah, Ajay Vijayan and Sree A Viswanathan, Varuni Shama Rao

Abstract— Yelp data consists of information about a variety of businesses and its reviews by various consumers. Its data consists of the detailed account of businesses, its text and star ratings by its reviewers, check-in information for a large set of cities from four different countries. Although this information is very useful, oftentimes users have to narrow down to a specific restaurant based on food and location preferences. Secondly, locating specific restaurants based on food choices and attributes becomes almost impossible to achieve. The project aims to build a visual recommender system with an interactive dashboard that would help both business owners and customers address the above problems by combining data from Yelp and Wikipedia.The dashboard would allow users to navigate to the relevant restaurant based on user's choice of country, state, and category. An interactive map would allow them to browse through the restaurant landscape and choose an appropriate restaurant based on location and average star reviews.In addition, the system recommends specific restaurants based on a person's social profile. For business owners, the visualizations display a comparison of restaurants opinion across different quarters. The owner can drill down to a specific year to understand how polarity (opinion) of their restaurant's quality of food item stands in comparison to its neighborhood. In addition, they can also understand which attributes of the restaurant such as service, ambiance and price are most important to the customers when selecting a particular restaurant and how their restaurant performs on average to the other restaurants in the neighborhood.

Index Terms—Natural language processing, sentiment analysis, visualization, yelp

1 INTRODUCTION

YELP is the collection of reviews crowd-sourced by various users and has become a key product when customers tend to make their restaurant choices. Each customer writes their reviews based on the various experience they face during their visit to the restaurant. Customer's choice of the restaurant can depend on a variety of factors like the name of the food item, quality of the food item, and the variety of other restaurant attributes like location, cost, and service. In addition, business owners would also like to prioritize on which items to improve and what attributes of the restaurant should be improved based on the reviews.

The major problem with the variety of these text reviews is that it takes quite some time to read those reviews and a customers and business owners won't be interested in reading a multitude of reviews and hence would make their choices based on reading a subset of reviews. However, these reviews may not be a representative sample of the reviews present on yelp since the user would not have time to read thousands of reviews. This would heavily bias the decisions they make which would also make the yelp platform less attractive.

Our analysis dashboard focuses on both customers as well as the business owners. Business owners would be interested to know if the quality of the food product had increased over time. If so, what specific food items should I prioritize with respect to competition in my neighborhood. Secondly, they would like to know how various attributes such as ambiance, service and wait time compare with other restaurants in the neighborhood. It would also highlight on what they could improve. Secondly, Customers would be given recommendations to restaurants based on the choice of specific food items and the social aspect. For the choice of food items, we recommend restaurants based on location as well as the food quality and for the social aspect, we engineer a recommender system that would highlight different restaurants based on the preferences of user's friends. The below paragraph describes the set of visualizations we used to address the above problems.

For business owners, we created an intensity map which displays variations of trends across different years and quarters and displays a bubble chart which shows the variation of polarity and average stars across different food items served in the neighborhood. For customers, we help them corner down to a particular restaurant based by using an interactive sunburst chart and the map. We then recommend them restaurants based on their friend's profile by creating a connected concept map.

Inorder to achieve our objectives, we required names of specific food items along with reviews of restaurants and quality of various attributes. So, we scraped names of food items from Wikipedia and then did extensive natural language processing on the review text to extract specific attributes. For instance, if a review contains "The service was average and ambiance was good", then the extracted attributes would be (service, average) and (ambiance, good). In addition, we also did full-text match on the reviews to extract specific food item reviews. Then a sentiment analyzer was used to extract sentiment of words to attach them with reviews. The document is organized as follows. Section 2 describes the review of the literature. Section 3 describes the design

- *Alankrit shah is with the Arizona State University, Tempe, AZ 85281. E-mail: ashah28@asu.edu*
- *Ajay Vijayan is with the Arizona State University, Tempe, AZ 85281. E-mail: ajay.vijayan@asu.edu*
- *Sree Aurovindh Viswanathan is with the Arizona State University, Tempe, AZ 85281. E-mail: sviswa10@asu.edu*
- *Vauruni Shama Rao is with the Arizona State University, Tempe, AZ 85281. E-mail: vshamara@asu.edu*

principles used in our project. Section 4 describes the system design along with various visualizations aimed to answer the above questions. In addition, it describes how the data analysis was carried out. Section 5 describes the results of the project via a case study of two different scenarios. Section 6 concludes the document.

2 Review Of Literature

There are several methods for extracting topics from the text, from a group of documents to visually display insights. The Latent Dirichlet Algorithm is a Bayesian generative model for text and it is the most common way to perform this. In "Improving Restaurants by extracting Subtopics from Yelp Reviews," Huang et al used LDA to uncover latent subtopics in Yelp reviews and visualize the extracted results. "Therefore, LDA is used to discover repressed topics in a text document. It assumes that a core of the text documents covers a collection of K topics. Topics are defined as the multinomial distributions over a word dictionary |V| words. The Algorithm was applied within a certain sub-topic resulting in the predictions and then displayed visually." [10] Although our focus is not by applying an algorithm, we similarly hope to provide insights about the yelp data which also provide recommendations for the selected user.

As mentioned in the paper "Visualization of gene expression information within the context of the mouse anatomy" by Andy Taylor, Kenneth McLeod, Chris Armit, Richard Baldock and Albert Burger, a sunburst takes information organized within a tree structure and displays the tree structure in a radial layout. "Assuming the information is organized as a tree no organizational data is lost. The center of a sunburst diagram is the root node of the tree, with children of the root node being the first layer of blocks in the sunburst. Children sit directly around in the next layer of the sunburst, and so on until the leaf nodes are reached at the edge of the diagram. The size and position of the blocks within the sunburst are used to indicate the structure, and organization, of the data. Data attribute values are presented by coloring the nodes." [9]

For the paper, "Visualizing Yelp Ratings: Interactive Analysis and Comparison of Businesses", by Akhila Raju, Cecile Basnage, and Jimmy Yin, they had interviewed quite a few business owners who actively compared their business to businesses in other cities. "One business owner had explained that he reached out to business owners who are not close in geographic proximity as it removed them from competition, and he could gain useful insights about what works for a business and what does not". [5] They implemented sentiment analysis of reviews that provided more insights into a review than the current rating system. It allowed them for more in-depth analysis and comparison between businesses with the same rating. They also wanted to implement a popular keywords feature, which allowed the users to see what the popular words for businesses across categories were.

3 Design Principles

In this section, we briefly describe the questions that we want our system to answer and the set of design principles that we followed to map the data to one of the visual variables. The following were the set of questions that we wanted to answer based on the data from Yelp and Wiki.

1. How do the restaurant perceptions change over time?
2. How do people perceive my restaurant's food quality when compared to my neighborhood?"
3. What factors influence the opinion of restaurants in my neighborhood?
4. How to choose the restaurant based on my geo-location and category preferences?
5. What restaurants will my friends be interested in?"

For question 1, we chose to use a stacked bar chart, as we had to display comparisons between various categories levels of categorical data. In addition, the goal was to show aggregated polarity values across all the years. Hence we used equal interval classification scheme [16] to divide the range into four equal bins. Since the polarity value ranged from -1 to +1 we chose to use a divergent scale to color the bar values [12].

For question 2, we chose to use a bubble chart as we wanted to show variations of polarity across different food items. We chose the position of the bubble to denote variations in polarity as that would provide us with much better visual perception. Secondly, a qualitative color scheme [12] was used to color different food items on the chart. This was used since there was no ordering between different food items.

For question 3, we used a parallel coordinate plot. This plot was chosen as it is well suited for expert driven data exploration for the large set of variables and variations in polarity values [17]. We are aware that the major problem with the parallel coordinate plot is the visual clutter. Inorder to avoid that we had a selection on the coordinate axis so that we can only visualize the lines we are interested in. Since the ordering [18] of variables are done per the popularity the business owner should prioritize, this plot was suitable for answering the question. This ordering would heavily influence the appearance of the parallel coordinate plot and hence would help make the correct decision for the user.

For question 4, we used a combination of Sunburst chart and an interactive map. We chose a qualitative color scheme for the Sunburst chart as there were no implicit ordering [12] of elements. The size of the sundial chart was used to denote the number of restaurants in that region. We chose to use a sequential color scheme for the map markers as it indicates the density of the restaurants in the neighborhood.

For question 5, we used a concept map. Small bubbles were chosen to depict the names of the users. The rectangles denote the name of each restaurant which has

been recommended to the user. We ordered the restaurant rectangles as per the users' preferences. A sequential color scheme denotes the common interest among the number of friends who might visit the restaurant.

As discussed above, appropriate visual elements were chosen as per requirements of each question and the choice of the color was chosen such that order [12], separation [13] and aesthetics [14] of the visualizations were maintained. Also, all the colors were selected based on Brewer's colors [19] schemes.

4 System Design and Implementation

In this section, we briefly describe the design of our system and how various visualizations are built to answer each specific set of questions. We used Shneiderman visualization mantra of visual analytics "Overview First, Zoom and Filter, Details on demand" as a framework in our visualizations. A live demo of our product can be viewed at https://youtu.be/cnacb2HXgD8 . The system is divided into two different set of visualizations that would help answer questions from business owners and customers

4.1 Business Owner - Centric Visualizations

Business-centric visualizations would help the business users to make prioritize on the factors to increase the number of customers visiting the business. There are three visualizations that would cater to the businessownwers and is described below.

4.1.1 Restaurant (food quality) opinion over Time – Intensity Map

This visualization addresses the question about "How does the restaurant perceptions change over time?" The intensity map shows how average polarity value of reviews for the specific restaurant had changed over time. The x-axis denotes time in months while the y-axis shows the distribution of polarity by a sequential color scheme. When we hover over a column, the percentage of items with a bad polarity/opinion is shown. The user can explore more about the specific part of that chart by clicking on it. This would take us to our next visualization which explains the specific set of food items.

4.1.2 Food Item quality comparisons - Bubble Chart

The visualization addresses the question about "How do people perceive my restaurant's food quality when compared to my neighborhood?" The perception polarity for the restaurant is denoted by the radius of the circle. And circle color denotes whether the information is related to owner's restaurant or the neighboring restaurants. The colors are chosen to keep in mind visual clarity in times of overlap. The x-axis of the graph corresponds to the star rating of the restaurant and y-axis corresponds to the review count or popularity of the restaurant. The user can explore the region by hovering on a specific bubble which represents the restaurant.

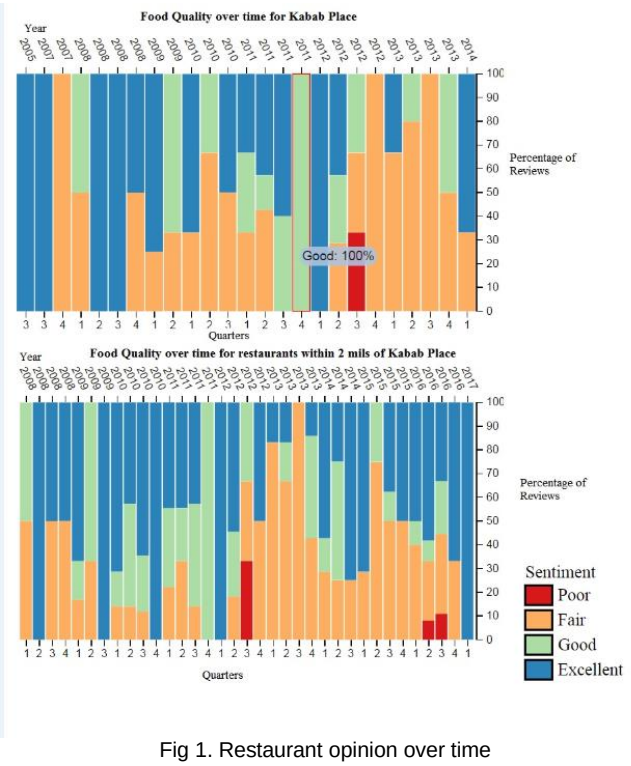

Fig 1. Restaurant opinion over time

This would highlight two bubbles on the screen one for the target restaurant and the other denoting the aggregation polarity score of neighborhood restaurants. All the other food items of the neighboring restaurants are faded away.

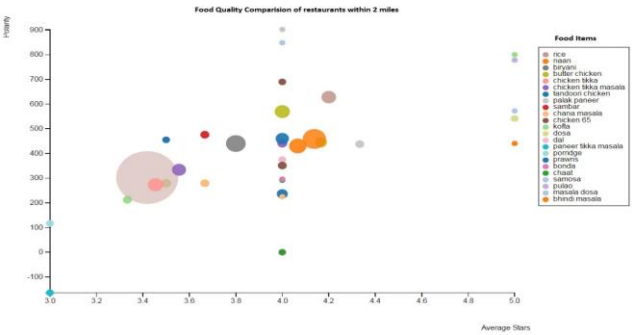

Fig 2. Food Quality comparision over time

4.1.3 Attribute Opinion comparisons - Parallel Coordinate Plot

This visualization addresses the question about "What factors influence the opinion of restaurants in my neighborhood?" Specifically, the parallel coordinate plot is utilized to compare the different attributes of a restaurant across user reviews. The attributes are represented by vertical lines within the plot. The intersection of these attributes with horizontal path marks the polarity of that attribute. Higher intersection cuts denote better polarity for the selected attribute. The path thus formed can be used to measure the overall quality of the business. The average of these attributes is also depicted by a bolded path in the plot.

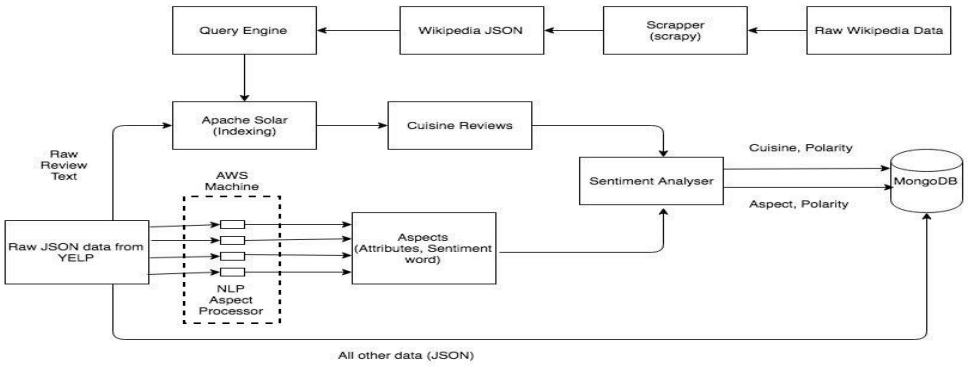

Fig. 1. Data Processing Infrastructure – Interaction of Apache Solr, MongoDB, and Raw Data.

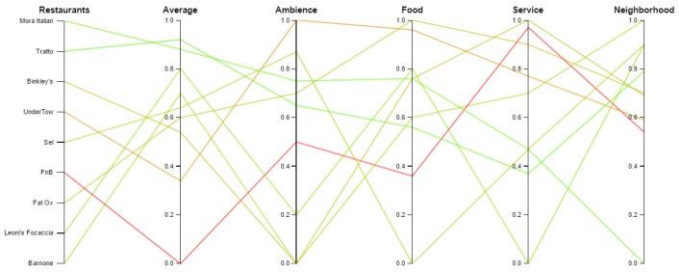

Fig 3. Attribute Opinion comparision

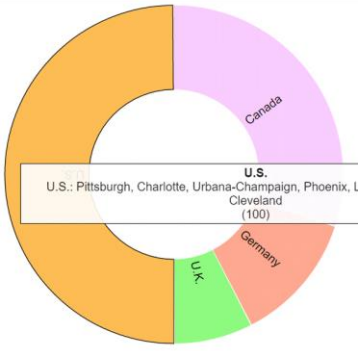

Fig 4. Navigation of hierarchy of Country, State & Cusine

The plot can also select only a subset of paths that correspond to a defined range of an attribute. This selective analysis feature will allow the business owner to conduct fine-grained analysis on the business state.

4.2 Customer Centric Visualization

Customer centric visualizations would help the customers to make choose a restaurant for themselves based on food quality or to choose restaurants based on friends' preferences.

4.2.1 Hierarchy of Country, State, City, Cuisine - Multi-level Pie Chart

This visualization addresses the selection problem. Specifically, this multi-level Pie Chart is a zoomable sunburst visualization that is to be used to navigate through the various levels of the hierarchy of country, state, city, zip code, cuisine, etc. When a part from second layer hierarchy is selected, the previous layer collapses itself, shrinking to the center of the screen and efficient management of the screen space.

4.2.1 Location preference- Interactive map

This visualization addresses the selection problem and the question "How to choose the restaurant based on my geolocation and category preferences?" Specifically, this interactive map would change depending on the selections made on the navigation wheel.

The possible levels could be country, state, city and lastly the food cuisine (category). When a country and state are selected, the map moves to the appropriate location and highlights the total number of restaurants as a bubble. The number at the center indicates the number of restaurants at that bubble. Selection of a particular bubble would expand and display all the restaurants covered by the specific bubble. Users can scroll and zoom the map to select the desired restaurant.

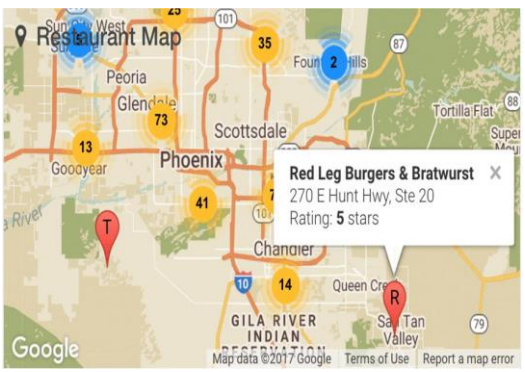

Fig 5. Interactive map for navigation

4.2.2 Restaurant Recommendation based on friends - Concept Map

This visualization addresses the question of "What restaurants will my friends be interested in?" Specifically, this concept map provides the restaurant recommendations for a selected user based on his/her

friends' interests. Each friend is linked to many restaurants depending on their food preferences and restaurant attributes. The higher the number of linking to a restaurant, the higher is the chances of the recommendation of that restaurant. Moreover, the position of the restaurants would be in the order of most to least recommended restaurants. The graph also includes color coding for each restaurant which signifies the overall rating of the restaurant. Selecting a user's friend from the graph would also highlight their linking to the restaurant showing the relation between the person and the restaurant

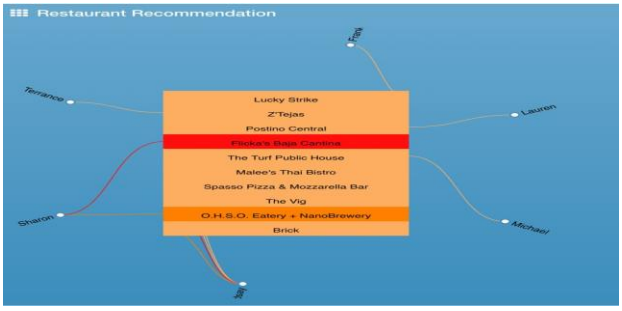

Fig 6. Restaurant Recommendation

4.3 Interaction among various visualizations

The interactive visualization recommender allows the user to explore the data by choosing various attributes using the Sundial.It reduces the number of restaurants on the interactive map. Once the restaurant is selected, the comparison of food quality ratings over time and the quality of specific food items for that restaurant is shown. Customers can be interactively selected by the interactive concept map.

5 System Setup

This section describes our system infrastructure along with data processing infrastructure that was used to clean and process data.

5.1 Dataset

The overall data comes from two disparate sources namely, Wikipedia and Yelp. Wikipedia data is scraped to obtain pairs of (cuisine, food item). The Yelp raw data is in JSON format and has raw review text along with tip data. This raw text data is then used to extract by aspects and quality of food items. The dataset and instructions to set up can be found at https://www.dropbox.com/sh/0k2qg1ln1rw38nw/AAAXqK4QA-RJ31jstWH22LmOa

5.2 System Components

Our system consists of the front-end which renders the set of visualizations with d3 and javascript technology. Since the data consisted of multiple values for different columns of the data we chose to use a NoSQL database. The database chosen for our product is MongoDB. Extensive scripts were stored in the database to perform aggregation and enable faster retrieval. In addition, we built indexes on geo-location and other columns that require faster retrieval. The middleware was setup using flask framework in Python.

5.3 Data Preprocessing

Extensive natural language processing and sentimental analaysis were used on review text to make meaningful charts on the data. The overall data preprocessing steps is provided in the block diagram of Fig 1. Key blocks are described in detail below.

5.3.1 Wikipedia Scraper

An initial set of cuisines were extracted from Yelp dataset. For each cuisine, a set of Wikipedia pages were extracted by Google search. For each page, all the cuisine names were scraped by using Scrapy (a python library). These data are then stored in separate JSON files.

5.3.2 NLP Aspect Processor and Extraction:

A very simple way to describe this process is to identify nouns in the sentence and then look for nearest adjectives around it. There are obvious shortcomings of the above approach and are overcome by extracting syntactic dependencies between words and output word forms by the Stanford CoreNLP tool. We used open source libraries that use Stanford's tool to extract various noun and the adjective pairs and these pairs are stored in JSON. These pairs of data are called Attribute Pairs. For example, some of the attribute pairs include (corn, sweet), (ambiance, horrible). The overall natural language processing time for the extraction of 2.5 GB of data was high. Hence we used an optimized amazon EC2 Linux cluster machines to parallelize this process and the resultant output was collected back to a single machine.

5.3.3 Query Engine and Apache Solr (Indexing)

The entire review text was put into a Lucene which has tokenization and full-text search capabilities inbuilt to it. The Query engine takes the Wikipedia pairs (cuisine, food item) and queries each food item and makes a full-text search on the Solr index. This returns the set of reviews for each food item. Now the sentence containing the food item is extracted out and is stored as a JSON file.

5.4.4 Sentiment Analyser

For each sentence obtained from the Solr output, the sentiment (polarity) of the sentence is extracted. Similarly, for each sentiment word in the aspect pair, the polarity of the word is determined. These data are then passed and stored into MongoDB for further querying.

6 Results

We designed the Yelp recommender system to help both business owners and users take informed decisions based on the dataset. The following description illustrates about the exploratory analysis capabilities of our system and how it can be used to analytically explore variations in trends and quality of the food items and provide

recommendations to users based on their friend's preferences.

6.1 Customer Centric Use Case

Our first set of customer-centric visualizations would help customers choose a particular restaurant based on their location and category preferences. Let's consider a scenario in which user named John would like to explore the places in a particular city say Tempe. The user would navigate through the hierarchy by the Sundial chart. This chart displays how the number of restaurants in is distributed at each city and at each cuisine level. The below charts (Fig 7) displays the wheel at each level of hierarchy.

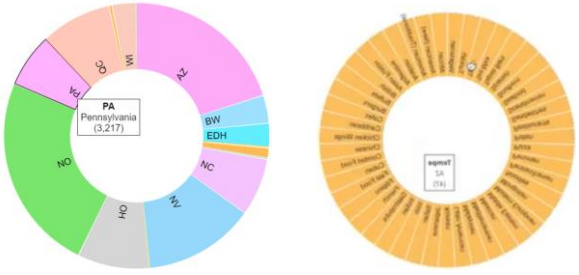

Fig 7. Sun dial Exploration

A navigation map shown in Fig 5 would help the user explore various places on the map. The user can scroll the map which would allow them to choose a particular restaurant. We had clustered the restaurant based on the closeness for the user to easily navigate through the map. A click on the particular restaurant cluster would expand to show all the restaurants in Tempe. This is used along with the Sundial chart to help reduce visual clutter and increase the ease of use.

Once the user had selected the particular restaurant, the system would then recommend the top reviewers for that restaurant. A click on the username would provide a recommendation of the list of restaurants the user and their friends would be interested in. A Sequential color scheme was adopted to show the restaurants. The list is ordered based on the user's preference and the color denotes the common interest restaurant among the friends' circle.

6.2 Business Centric Use Case

The next scenario, we explored is the business-centric one, where a business owner attempts to look for trends in the restaurant opinion based on the sentiment of the review text. With the help of the sunburst chart and the user can navigate the map and select the restaurant. For our use case let us consider that the business owner had chosen Kabab place. Then a comparison chart shown in business-centric view as shown in Fig 1. We could observe from the first graph that, the restaurant had not performed well over time. The initial distribution of the restaurant shows excellent ratings and as it progressed, the quality of food items had reduced over time.

The business owner can now select a particular year and quarter on the map where there is a drop in quality and observe how their food items are rated in comparison with the competition. This would lead the user to the chart shown in Fig 2. This chart illustrates all the food items of the neighborhood restaurants within the specified radius. Specific food item details can be obtained by hovering on it. The below chart displays how a specific item is highlighted and all the other bubbles are grayed out.

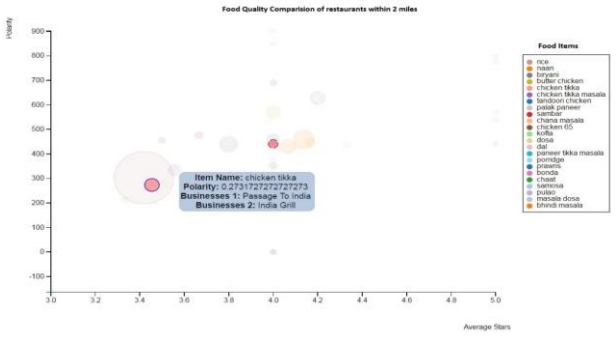

Fig 8. Food item comparison

The above chart displays the tooltip which displays the average polarity of the item in the nearby businesses. Finally, the business owner might attempt to understand which attributes of the restaurant the users in the neighborhood are interested in. We sorted the attributes based on the popularity in the reviews and display the top 5 attributes in the neighborhood. The business owner can select a particular restaurant and observe the trend line across different attributes. As shown in Fig 9, we could observe that food, place, price, service were the top attributes in that neighborhood and we could observe the trend line in blue for the restaurant Kohinoor Cuisine.

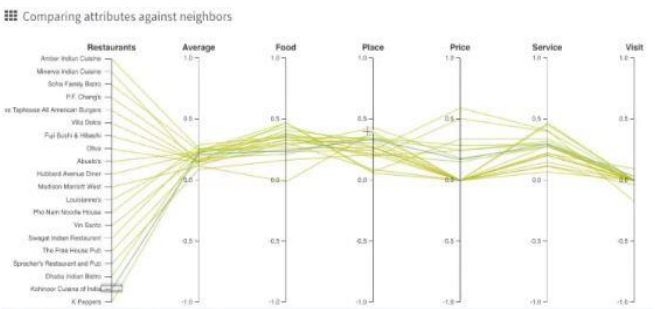

Fig 9. Selection of Restaurant and its Parallel Plot

7 Conclusion

The visual recommender system provides insights about the yelp data to both the business owners and the customers for taking informed decisions. The system provides recommendations for the selected user with the help of interactive dashboard which uses a combined dataset from Yelp and Wikipedia. Here, we have tried to provide answers to the four questions after analyzing the data.

The sentiment analysis implemented on the reviews have provided the system with more insight into the reviews than the normal unprocessed rating. Through this, the system could consolidate and inter-relate all the graphs with respect to both the business user and the customers. The system predominantly uses four graphs to visualize the questions. The initial graph provides a selection guide to various facets based on the distribution of the items inside each category. The food quality over time intensity maps provide us with an overall view of the customer's change in food quality perception over time. This is further extended to show the food item quality comparison and attribute opinion comparison. Finally, the system recommends the customer with a list of restaurants based on their friends' preferences which are well connected with a concept map.

REFERENCES

[1] Chi, EH-H., et al. "A spreadsheet approach to information visualization." Information Visualization, 1997. Proceedings, IEEE Symposium on Information visualization. IEEE, 1997.

[2] Gross, Markus H., Thomas C. Sprenger, and J. Finger. "Visualizing information on a sphere." Information Visualization, 1997. Proceedings of IEEE Symposium on Information visualization. IEEE, 1997.

[3] Sperka, Martin, and Peter Kapec. "Interactive visualization of abstract data." Science & Military Journal 5.1 (2010): 84.

[4] Keim, Daniel A., Christian Panse, and Mike Sips. "Information visualization: Scope, techniques, and opportunities for geovisualization." (2004): 1-17.

[5] Raju, Akhila, Cecile Basnage, and Jimmy Yin. "Visualizing Yelp Ratings: Interactive Analysis and Comparison of Businesses."

[6] Michael Luca. "Reviews, Reputation, and Revenue: The Case of yelp.com"

[7] James Huang, Stephanie Rogers, Eunkwang Joo. "Improving Restaurants by Extracting Subtopics from Yelp Reviews"

[8] Nima Hejazi. Capstone Project: Clustering and Topic Modeling with Yelp Reviews.

[9] Andy Taylor, Kenneth McLeod, Chris Armit, Richard Baldock and Albert Burger. "Visualization of gene expression information within the context of the mouse anatomy".

[10] James Huang, Stephanie Rogers, Eunkwang Joo. "Improving Restaurants by extracting Subtopics from Yelp Reviews".

[11] Shneiderman, Ben. "The eyes have it: A task by data type taxonomy for information visualizations." Visual Languages, 1996. Proceedings. IEEE Symposium on. IEEE, 1996.

[12] B. E. Trumbo, "Theory for Coloring Bivariate Statistical Maps," The American Statistician, vol. 35, no. 4, pp. 220-226, 1981

[13] -H. Levkowitz and G. T. Herman, "Color scales for image data," IEEE Computer Graphics and Applications, vol. 12, pp. 72-80, 1992

[14] K. Moreland, "Diverging Color Maps for Scientific Visualization," Proceedings of the 5th International Symposium on Visual Computing, December 2009.

[15] Heer, Jeffrey, and George Robertson. "Animated transitions in statistical data graphics." IEEE transactions on visualization and computer graphics 13.6 (2007): 1240-1247.

[16] E. K. Cromley and R. G. Cromley. An analysis of alternative classifications schemes for medical atlas mapping. European Journal of Cancer. Series B (Methodological), 26(2):211-252, 1964.

[17] Edsall, Robert M. "The parallel coordinate plot in action: design and use for geographic visualization." Computational Statistics & Data Analysis 43.4 (2003): 605-619.

[18] Yang, J., Peng, W., Ward, M.O., Rundensteiner, E.A., Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. In Proc. of IEEE Symposium on Information Visualization (2003), pp. 105–112.

[19] Harrower, Mark, and Cynthia A. Brewer. "ColorBrewer. org: an online tool for selecting colour schemes for maps." The Cartographic Journal 40.1 (2003): 27-37.