Assignment 3

Interactive Visualization

Joris Reijrink
Student number: 0847198
j.reijrink@student.tue.nl

Chris Hoedemakers
Student number: 0661115
c.g.j.j.hoedemakers@student.tue.nl

January 22, 2015

1 Introduction

In this report we propose an interactive application for analyzing the restaurant & consumer data set provided by the UCI Machine Learning Repository [6]. However, before we present this application, we explain the contents of this data set (Section 2), argue for several analysis task that can be performed on this data set (Section 3), and discuss different visualization techniques that can be used in order to achieve these tasks (Section 4). In Section 5 we then present the proposed application and underpin the design decisions that were made. We conclude this report with several observations regarding the application design, in order to draw some final conclusions (Section 6).

2 Data Set

The dataset used for the visualizations described in this report is named *Restaurant & consumer data, recommender systems domain* and originates from the UCI machine learning repository [6]. The dataset consists of 9 files that can be separated into 3 categories:

Restaurant

1. chefmozaccepts.csv
2. chefmozcuisine.csv
3. chefmozhours4.csv
4. chefmozparking.csv
5. geoplaces2.csv

Consumer

6. usercuisine.csv
7. userpayment.csv
8. userprofile.csv

Rating

9. rating_final.csv

The restaurant data-files describe various properties of each restaurant, like location, cuisine or accepted payment method. The consumer files describe the characteristics of a consumer like age, weight and budget. But also preferred cuisines or smoker data-fields are present. The final data-file links the consumer to a restaurant. This files contains the restaurant rating of a consumer, a consumer can have ratings of various restaurants. A complete list of all data-items and their description is present in appendix A.

3 Tasks

In this section we formulate a set of tasks that should be performed on the data using the final application. However, before defining these tasks, we present a framework for formulating data analysis tasks. We then actually define possible tasks for the restaurant & consumer data set using this framework, and select a few of these tasks.

3.1 Framework for Defining Tasks

In order to formulate a nice and clear visualization task, we provide a framework for doing so as presented in [4]. Here we distinguish between five different dimensions for each task, namely the *goal* (why is the task pursued?), *means* (how is the task carried out?), *characteristics* (what does the task seek?), *target* (on which part of the data is the task carried out?) and *cardinality* (on how many instances is the task carried out?). We shortly explain these five aspects in some more detail.

For the goal of the task we distinguish between three different types of analyses. We can have an exploratory analysis (or undirected search) that aims at deriving an hypothesis from an unknown data set, a confirmatory analysis (directed search) that aims to verify a found or assumed hypothesis, and finally a presentation that simple exhibits analysis results.

For describing how the task is being carried out we distinguish between three main ways of doing so. These are navigation, (re-)organization and relation. Navigation subsumes all means that change the extent or the granularity of the shown data, but that do not reorganize the data itself. (Re-)organization includes all means that actually adjust the data to be shown by either reducing or enriching it, and relation encompasses all means that put data in context.

Characteristics of the data can be subdivided in low-level data characteristics and high-level ones. Examples of low-level characteristics are data values of a particular data object or the data objects corresponding to a particular data value. Examples of high-level characteristics include more complex data patterns such as trends, outliers, clusters, frequency, distribution, correlation, etc.

The target of the visualization task describes on which part of the data the task is being carried out. This often boils down to a certain (sub)set of the different data attributes.

The cardinality finally specifies which part of the data instances we consider when carrying out the task. For instance just a single one, a certain subset, or just all of them.

3.2 Possible Tasks

Given the restaurant & consumer data set as presented in Section 2, numerous analysis tasks can be formulated. We start by observing that we can distinguish between three main sets of tasks.

The first set of tasks involves the analysis of the restaurant data, including their ratings. Here we can identify tasks that for instance focus on finding correlations between restaurant features and their ratings. An example of a specific question that could be posed here is "Do restaurants receive better ratings when they are part of a franchise, provide internet, serve a certain cuisine, or are located in a certain geographical area?".

The second set of tasks involves the consumer data. In this set we find tasks such as identifying correlations between the consumers their preferred type of cuisine and their personal information. An example of a specific question within this set of tasks could be "Do consumers that are obese, are a heavy drinker, or have a small budget, have a common type of cuisine they prefer?".

The third and last set of tasks we can identify are those that link the restaurant data to the consumer data, using the ratings of the consumers for the restaurants. What is especially interesting here is finding contradictions (which are probably outliers). These could be found by posing questions such as "Are there non-smoking restaurants that are visited (and even rated good) by consumers that smoke?" and "Are there consumers that travel a long distance to visit a certain restaurant?". But also the discovery of certain correlations between restaurant features and consumer characteristics might bring forth interesting observations. Specific questions that could be posed include "Do consumers of a certain religion more strongly prefer restaurants with a formal dress code?" and "Do younger people prefer restaurants that are part of a franchise?".

3.3 Chosen Tasks

From the three main sets of data analysis task that can be defined for the restaurant & consumer data set, we focus on the the first and third one in this project.

Within the first set of tasks (regarding the restaurant data and their ratings), we concentrate on exploring correlations between restaurants their features and their received ratings. In order to specify valuable visualization techniques, we formulate the following set of specific analysis questions:

Do restaurants receive a better rating when they

- are part of a franchise?
- provide internet?
- serve a certain cuisine?
- are located in a certain area?

formal task descriptions

In order to formally describe this first task, we use the framework presented in Section 3.1. For the goal of this task we observe that we want to derive hypotheses from the unknown data, indicating an exploratory goal. We furthermore recognize that we want to do this by means of relation-seeking, in order to characterize correlations between the set of restaurant features/attributes (denoted by S_{rest}), and their rating (denoted by S_{rate}). When performing this task, we consider all instances of the data set, giving rise to the following formal definition for the task:

(exploratory, relation-seeking, correlations, $S_{rest} \cup S_{rate}$, all)

For the third set of tasks (that links the restaurant data to the consumer data), we concentrate on exploring correlations between restaurant features and consumer characteristics (i.e., what kind of people rate what kind of restaurants well or bad?).

- "Do younger people prefer restaurants that are part of a franchise?"
- "Do consumers of a certain religion prefer restaurants with a formal dress code?"

Furthermore we are especially interested in possible outliers. Specific questions that we could formulate regarding this are:

- "Are there non-smoking restaurants that are visited (and even well-rated) by consumers that smoke?"
- "Are there non-alcoholic restaurants that are visited (and even well-rated) by heavy drinking consumers?"
- "Are there consumers with a high budget that visit restaurants with low prices? Do they like them? And what about the other way around?
- "Are there consumers that travel a long distance to visit a certain restaurant?"

In order to formally describe these tasks, we make a similar observation as with the first task. The main difference is that we are also actively looking for outliers here, and that we include the set of consumer attributes (denoted by S_{cons}) in order to perform the described tasks. This therefore yields the following formal description of the second task:

(exploratory, relation-seeking, correlations|outliers, $S_{rest} \cup S_{cons} \cup S_{rate}$, all)

4 Techniques

In order to achieve the tasks that were defined in the previous section, we will argue how suitable different visualization techniques are for doing so. We start with considering the first of these two tasks that is formulated as "Are there any correlations between restaurant ratings and one or more of their attribute values?". For this task, fairly simple visualization techniques could suffice, because we simply want to explore the relation between two attribute values (i.e., the restaurant its rating and one of its attribute values). Examples of suitable visualization techniques for this task are a Scatterplot [5] or (Normalized) Bar Chart. More involved techniques such as a Parallel Coordinate Plot could also be considered, but are not necessarily better. Often these techniques have a tradeoff between the amount of data they are able to show, and the level of detail of the presented data. Because the data for the first is rather small, a more basic technique might even be preferable.

When looking at the second analysis task however, the amount of data we consider increases quite a lot. We especially want to be able to view more data simultaneously, because we want to link restaurant attributes to consumer attributes, using the rating that the consumer gave the restaurant. This means that we want at least three data objects to be shown: the restaurant ID, the consumer ID, and the rating. However, we could still do this using a basic Scatterplot by placing the ID's on the axes and introducing color coding for the rating. If we would want to explore more attribute relations at once, we could even consider to use a Scatterplot Matrix [**?**]. Though this increases the amount of space that is needed to show all of the plots and makes it harder to interpret them, this issue could be anticipated on by introducing an additional view that shows a single plot in detail.

Another technique that could be used to explore attribute relations is a Parallel Coordinate Plot [3]. With this technique we could use three axes, where the left axis represents a restaurant attribute, the right axis a consumer attribute, and the middle one the rating that the consumer gave to that restaurant. In other words, each line would represent a rating by a consumer with a certain attribute value on a restaurant with a certain attribute value. However, because most of the attribute values are discrete, Parallel Sets [2] might be considered as a better option. This technique on the other hand disables the analyst to easily distinguish single ratings.

Let us now consider the specific questions "Do restaurants receive a better rating when they are located in a certain area?" and "Are there consumers that travel a long distance to visit a certain restaurant? And how do they travel and confirm they rate the restaurant well". These questions clearly refer to a geospatial analysis. Therefore some kind of geographic (symbol) map [1] might be suitable in these cases in order to quickly get a clear overview of the data. But also in other cases a map might be interesting. For the restaurant & consumer data set we could for instance use Mike Bostock his Symbol Map to get a quick overview of where all consumer live that rated a selected restaurant. Additionally we could show the rating of that consumer using a tooltip when the analyst hovers over the line that connects the restaurant to the consumer.

5 implementation

In the end scatterplot might show data in easiest way to analyze. Adding additional interaction can overcome limitations

6 Observations

6.1 Observations Task 1

6.2 Observations Task 2

References

[1] Mike Bostock. Symbol map airports, 2008.

[2] Jason Davies. Parallel sets.

[3] Jason Davies. Parallel coordinates, 2011.

[4] Hans-Jorg Schulz, Thomas Nocke, Magnus Heitzler, and Heidrun Schumann. A design space of visualization tasks. *Visualization and Computer Graphics, IEEE Transactions on*, 19(12):2366–2375, 2013.

[5] Nikhil Sonnad. Scatterplot: Social trust vs ease of doing business, 2013.

[6] Irvine University of California. UCI Machine Learning Repository restaurant & consumer data set, 2012.

A Appendix: Dataset

The data attributes are described using the following format:
attribute name [type]: description

1. chefmozaccepts.csv

placeID [number]: unique identifier of the restaurant
Rpayment [text]: describes the payment methods the restaurants accepts

2. chefmozcuisine.csv

placeID [number]: unique identifier of the restaurant
Rcuisine [text]: describes the present cuisines of the restaurants

3. chefmozhours4.csv

placeID [number]: unique identifier of the restaurant
hours [text]: opening hours of the restaurant in 00:00-23:59 format
days [text]: days in the week the restaurant is open, days are separated with semicolon

4. chefmozparking.csv

placeID [number]: unique identifier of the restaurant
parking [text]: describes the parking possibilities of the restaurant

5. geoplaces2.csv

placeID [number]: unique identifier of the restaurant
latitude [number]: latitude of the restaurant location
longitude [number]: longitude of the restaurant location
the_geom_meter [number]: geospatial name of the restaurant
name [text]: restaurant name
address [text]: restaurant address
city [text]: city where the restaurant is located
state [text]: state where the restaurant is located
country [text]: country where the restaurant is located
fax [text]: fax number of the restaurant
zip [text]: zip code where the restaurant is located
alcohol [text]: describes if the restaurant serves alcohol
smoking_area [text]: described if the restaurant permits smoking inside
dress_code [text]: restaurants type of dress code
accessibility [text]: accessibility for disabled
price [text]: overall pricing of the restaurant, this can be low, medium or high
url [text]: restaurant website url
Rambience [text]: ambience of the restaurant, this can be familiar or quiet
franchise [boolean]: is the restaurant part of a franchise
area [text]: area of the restaurant
other_services [text]: other services provided by the restaurant

6. usercuisine.csv

userID [number]: unique identifier of the consumer
Rcuisine [text]: the preferred cuisine of the consumer, a user can have multiple

7. userpayment.csv

userID [number]: unique identifier of the consumer
Upayment [text]: payment methods the consumer has used in restaurants

8. userprofile.csv

userID [number]: unique identifier of the consumer
latitude [number]: latitude of the consumer's home
longitude [number]: longitude of the consumer's home
the_geom_meter [number]: geospatial name of the consumer's home location
smoker [boolean]: is the consumer a smoker

drink_level [text]: drinking level of the consumer
dress_preference [text]: dress code preference of the consumer
ambience [text]: ambience preference of the consumer
transport [text]: transportation preference of the consumer
marital_status [text]: marital status of the consumer
hijos [text]: children status of the consumer
birth_year [text]: birth year of the consumer
interest [text]: interests of the consumer
personality [text]: personality of the consumer
religion [text]: religion of the consumer
activity [text]: work status, this can be student, professional, unemployed or working-class
color [text]: favorite color of the consumer
weight [number]: weight of the consumer
budget [text]: food budget of the consumer, this can be low, medium or high
height [number]: height of the consumer

9. rating_final.csv

userID [number]: unique identifier of the consumer
placeID [number]: unique identifier of the restaurant
rating [number]: the overall rating of the consumer at a specific restaurant
food_rating [number]: the food rating of the consumer at a specific restaurant
service_rating [number]: the service rating of the consumer at a specific restaurant

B Appendix: Work breakdown

Joris	Chris
Report	
data	introduction
appendix A	tasks
	techniques
Implementation	
geographic map pane	parallel coordinate plot pane
average rating pane	
restaurant / consumer information pane	