Interactive Data Analysis Tool

Big Data: Using R and Shiny

UNIVERSIDAD POLITECNICA DE MADRID

January 27, 2016

Authored by: Aditya Bhadoria & Adrián Ramírez del Río

Interactive Data Analysis Tool

Table of Contents

1. Problem Background

The task at hand is to develop an interactive data analysis tool which will provide a computer based visualization system to help people carrying out tasks more effectively using visual representations of datasets. The design of this tool, like any other interactive analysis tool, must solve the particular problems that an "end user" or "potential customer" has to deal with when facing some data that he wants to analyze in a work/normal session. Implementation process is carried out with the help of a web application framework of R known as Shiny.

In the existing literature, Visualization (Vis) is defined as the tool suitable when there is a need to augment human capabilities rather than replace people with computational decision-making methods. The design space of possible vis idioms is huge, and includes the considerations of both how to create and how to interact with visual representations. Vis design is full of trade-offs, and most possibilities in the design space are ineffective for a particular task, so validating the effectiveness of a design is both necessary and difficult. Vis designers must take into account three very different kinds of resource limitations: those of computers, of humans, and of displays. Vis usage can be analyzed in terms of why the user needs it, what data is shown, and how the idiom is designed.

In order to fulfill above said requirements, the visualization has to be effective in supporting user tasks. This means that the visual encoding and user interaction should display data in a correct, accurate and true way. In terms of visual channel used in a visual encoding, we have to ensure both expressiveness and effectiveness. The expressiveness principle states that the visual encoding should express all the information in the attributes of a dataset, and only the information from the dataset. Effectiveness principle on the other hand says that the importance of an attribute should be aligned with the salience of the corresponding channel. In a plain simple way, most important attributes in a visual encoding should be represented by the most effective channels. To measure the channel effectiveness, we have several parameters to evaluate such as – accuracy, discriminability, separability, pop-out and grouping.

2. Data Introduction

For fulfilling the visualization task, we decided to encode the classic Yelp dataset. Yelp users give ratings and write reviews about businesses and services shown on Yelp website. These reviews and ratings help other Yelp users to evaluate a business/establishment or a service and make a choice whether to select them or not for particular purpose. While ratings are useful to convey the overall experience, they themselves do not convey the context which led a reviewer to that experience. Yelp reviews and ratings are crucial source of information to make informed decisions about a business/establishment.

Important features about the original Yelp data:

- 2.2M reviews and 591K tips by 552K users for 77K businesses
- 566K business attributes, e.g., hours, parking availability, ambience.
- Social network of 552K users for a total of 3.5M social edges.
- Aggregated check-ins over time for each of the 77K businesses

Geographies Covered:

- U.K.: Edinburgh
- Germany: Karlsruhe
- Canada: Montreal and Waterloo
- U.S.: Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison

Details about the data which we used in our project:

Business

```
{   'type': 'business',
    'business_id': (encrypted business id),
    'name': (business name),
    'neighborhoods': [(hood names)],
    'full_address': (localized address),
    'city': (city),
    'state': (state),
    'latitude': latitude,
    'longitude': longitude,
    'stars': (star rating, rounded to half-
stars),
    'review_count': review count,
    'categories':    [(localized    category
names)]
    'open':  True  /  False  (corresponds  to
closed, not business hours),
    'hours': {
        (day_of_week): {
            'open': (HH:MM),
            'close': (HH:MM)
        },
    },
    'attributes': {
        (attribute_name):
(attribute_value),
       },
}
```

Review

```
{
    'type': 'review',
    'business_id':   (encrypted   business
id),
    'user_id': (encrypted user id),
    'stars':  (star  rating,  rounded  to
half-stars),
    'text': (review text),
    'date':  (date,  formatted like '2012-
03-14'),
    'votes': {(vote type): (count)},
}
```

One crucial point to consider is that while developing the visualization tool, we had to keep in mind the necessity of making this tool interactive, since interactivity is quite effective for developing tools that handle large and complex dataset. As evident from the important features of yelp data, this data is really varied, complex and huge. On extraction of the dataset, the size is found to be 1.2 GB approximately.

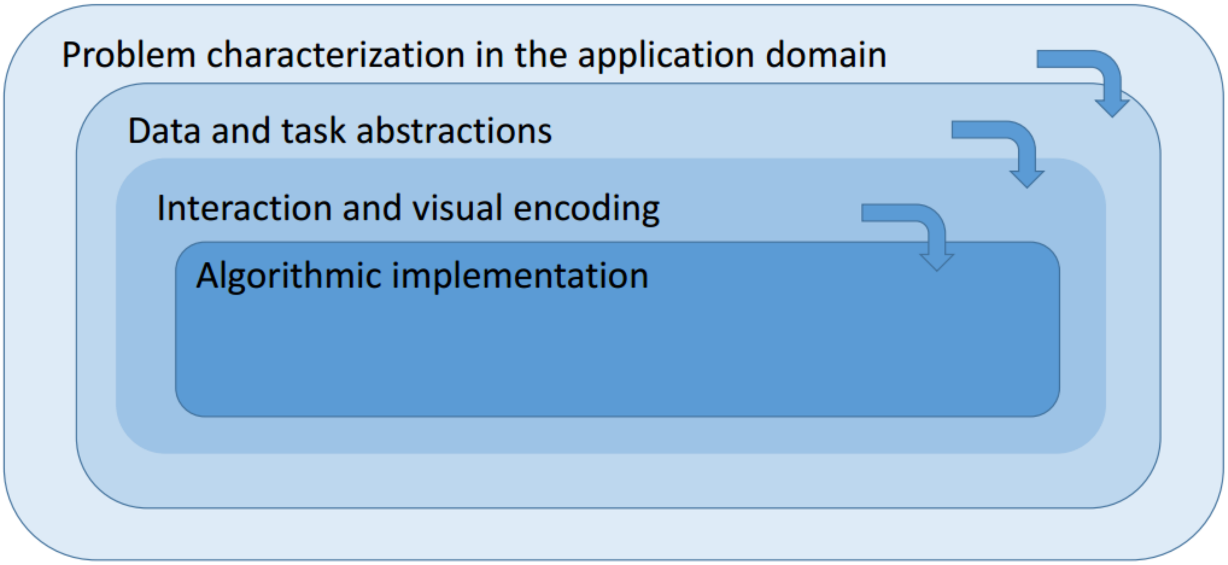
For datasets as large as Yelp, the limitations of both people and displays prevent even only the view of everything at once or on a single scale; interaction comes in this situation very handy, where user actions can change the view to change or navigate in several directions. Additionally, a uni-static view can only represent one aspect of the dataset. For some tasks and datasets, this is more than sufficient, but for more complex datasets, interactivity makes it possible to get an overview of these huge data by providing the choice to user to change and view different aspects of data.

For our dataset, an interactive visualization tool will support the exploration at multiple levels of detail, varying from a very high-level overview of dataset to straight multiple levels of summarization along with fully detailed view of a small part of it.

Another important aspect is the platform on which our tool will be used, since our tool is web based computer graphics, there is possibility of interactivity. With the computing power of present systems and rich graphics, it is entirely possible to represent this huge Yelp dataset in different ways of presentation and summarization in a way that helps in understanding the connection between different features and attributes of dataset.

3. Validation and Design Choices

Validation is necessary to have effective and expressive designs because the design space is huge and most of the designs are ineffective for a particular design problem. There are many possible ways for validation and this whole complex problem has been divided into four cascading levels which are as following:


Problem characterization in the application domain
Data and task abstractions
Interaction and visual encoding
Algorithmic implementation

3.1 Problem Characterization in application domain

This block represents the top level for design abstraction levels, this describe the specific domain and encompasses the group of target users who will be using our tool to accomplish some objective. Each domain has its own terms, vocabulary, type of data and problems. In this level, we try to identify the situation blocks which are an understanding that the designer reaches about the needs of the user. The methods to do this include interviews, observations, or careful research about target users within a specific domain.

For our visualization tool, the target user can be anyone who wants to go out to eat food, for drinks or takeout. This is very important because for our tool, we ourselves, our friends and the regular people we meet are target users. This has helped us in evaluating the *needs* of a typical user and particular problems that the target audience would *solve* from this tool. Our typical user wants to go to a restaurant, pub or any similar establishment (in this document, we have used word “establishment” to denote restaurants and businesses), and he wants the best possible experience available in his/her budget and conforming to other desired features. One way in which we have done this is by a field study, where we *asked* and *observed* several potential users to verify the problem characterization in real world settings, asking them their *approaches* to this problem and different *mechanisms* they follow in performing this task.

Just to confirm, the users are performing all these tasks on a computing platform using a graphical user interface (GUI), so we have to think on the similar lines of making our tool as a web based tool favorable for laptops, desktops and mobile browsers. The crucial thing we found during these *interactions* with our target users is the importance of location, neighborhood, ratings, number of reviews, the content of reviews, proximity of other establishments and categories while selecting the place to eat and drink.

This has also helped in *determining* the structure of our visualization tool and the order of different widgets on the dashboard for a smooth user design and experience.

3.2 Data and task Abstraction

This design abstraction level includes abstracting the specific domain questions and the data pertaining to the specific domain which we have characterized in the previous abstraction level. Different domain questions can help in realizing the exact *domain-independent tasks* which user is trying to accomplish using the visualization tool. The *abstract data blocks* on the other hand have to be curated as per the design requirements. Its more than just identification of data, since, we have to find the exact way in which data will be useful to the user and many times, we have to *transform* or *derive* some new data to make it more relevant and appealing to the end user.

In Yelp data, we have first performed the data abstraction i.e. extracting the data to be represented and summarized visually in the tool. The data types in our dataset are *Items, Attributes and positions*. The items are the individual establishments (like- restaurants) which have different attributes such as ratings, opening time, neighborhood, review count and categories. The *position* is a derived attribute using two quantitative cyclic value attributes which are latitude and longitude. Each review is further an item for every establishment (nested approach) and each review has following attributes – stars (Ordered key attribute) and dates (sequential value attribute). Dataset type for the derived position is *Geometry (Spatial)* and later on the interactive map, *derived clusters* to show aggregation of locations has also been applied over the original spatial data.

Next in the analysis comes the task abstraction, the tasks user wants to perform using the visualization tool, these are further divided into *actions*, that define user goals. The other is *targets*, that thing on which the above said actions are performed. In *actions*, users can *analyze* using the tool. For example- a user can *discover* the geographic location where the most of the dinner restaurants are present, can *present* this to some other individual or group of people, can *annotate* one particular restaurant on the map, *derive* deep insights like which location has restaurants with good ratings. For *search* part of *actions*, user can do all the four tasks which are *lookup, browse, locate and explore* using the interactive map widget on the visualization tool.

Further, in the *query* part of *actions*, a user can *identify* the establishments in the interactive map using the pop-outs which shows the exact address of establishment along with other attributes. User can also *compare* two different restaurants' ratings or two different areas having different distribution of ratings (this is very crucial because we found out by *asking* people that people prefer to go to areas which have lot of good places opposed to single best place, as they want to have choices just in case or other reason is that the overall ambience in these areas with lot of good ratings is pleasant and amicable as an whole.)

The *targets* in task abstraction are *features* and *spatial data*, the definition of features suggests these as any particular structures of interest. Users want to *analyze* and *query* features like which are the best places for breakfast, do they give takeouts. For the *spatial data* (i.e. interactive map), users can analyze and search different business or single one of interest.

3.3 Interaction and Visual Encoding

In this abstraction level, we perform the creation and manipulation of visual representations for the abstract data block and abstract tasks chosen at previous levels. Each of this distinct approach is known as *Idioms*. Major factors while designing idioms are visual encoding (controls what users see) and interaction (controls how users change what they see). In our case, we will be using *multiple idioms* for visual encoding and user interaction purpose.

In our dashboard, we have a *barplot* as an idiom on left side, where data is star ratings (ordered key attribute) along with a derived quantitative value attribute (counts of establishment for each rating) and the line marks are encoded expressing the value with aligned vertical position, separating key attribute with horizontal position. For this encoding, we chose bar plot over pie chart, because of effectiveness of rectilinear layouts over radial layouts, studies have shown in literature that in rectilinear, perception speed is faster, and accuracy tended to be better. For showing differences, rectilinear layout outperforms radial layout.

Next in design is the *interactive map* as idiom, where data is location of restaurants which is the position derived from latitude and longitude and encoding used is *given use* with *point marks* specifying each and every location of establishments on the map. Maps come handy as the most common choices for cartographic generalization and geo-visualization. The reason for choosing map for representing is the most crucial and informed decision of the project, first reason is that technically, this is the most effective and expressive choice for completing most of the tasks of a potential user.

Other substantial reason is the familiarity of the users with this kind of interface, since our target user is a regular person with not so specific technical background; we have to develop the tool while keeping in mind the mass user acceptance. This kind of interactive map interface is already very common to users thanks to Google maps, Bing maps and Here maps. Regarding the choice of *channel* as spatial region is self-explanatory as this is the top ranked channel for *categorical variables* (different establishments) and points marks have been used to pin-point the exact location of establishment.

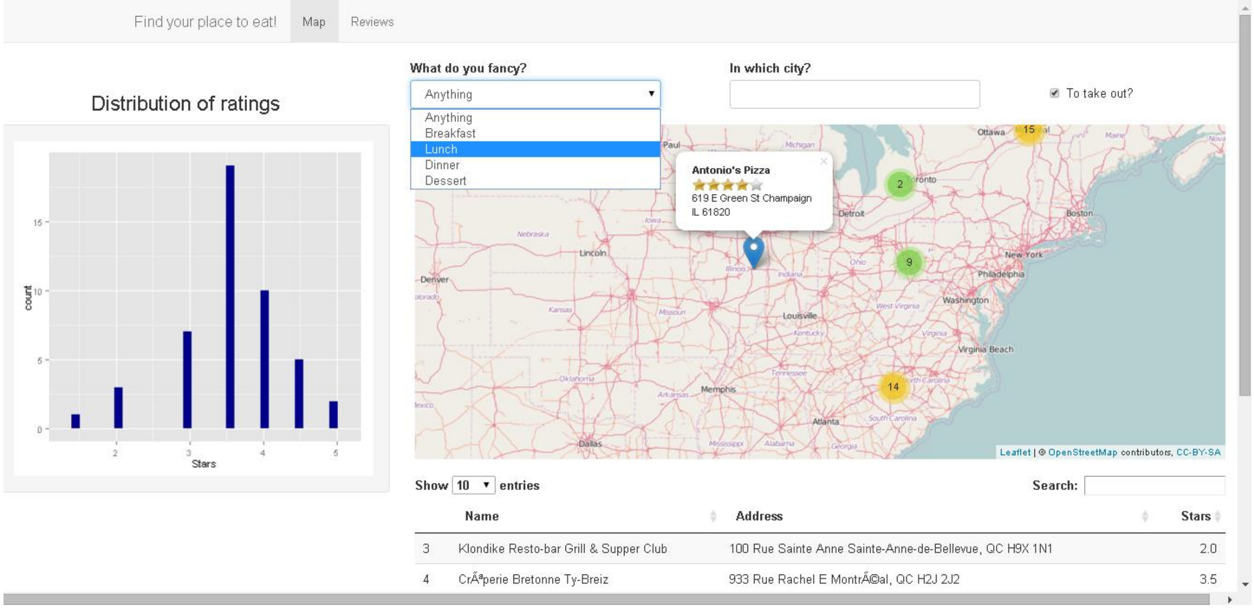
In *manipulation*, we have used the *Select* and *Navigate* both to present the complex and huge data is a very simple and approachable way. *Selection outcomes* represent an additional action where a pop-out comes out from the selected point mark and gives information about address and rating of establishment. Point to note is that in this pop-out, we have used pictures of ratings corresponding to *shapes* as channel to represent an ordinal ordered attribute.

Navigation has been used to change the point of view from which things are drawn, it can be understood in a way that camera moves closer to or farther from the plane. Zooming the camera will show less number of items, though they will appear larger and zooming out will show more items, and now they will appear larger. Clearly, the outcome of this navigation is a combination of *filtering* and *aggregation* (overview). *Pan/Translate* has been employed to move at same zoom level in interactive map, its equivalent to panning the camera moving it parallelly to the plane of the image, either up and down or from side to side. Additionally, a *derived cluster* encoding to show aggregation of locations while zooming out has also been used, it uses *color channel of hue* to represent a derived quantitative value attribute (number of establishments present in the cluster as zooming out).

Facet information has been utilized to connect the ratings barplot and interactive map. These two idioms have been *juxtaposed* side by side and connected using the *share navigation*. That is, when manipulating the interactive map by zooming in and zooming out, the ratings in barplot will correspond only to the establishments which have locations present in the geographical area represented in the interactive map.

To further manage the complexity of dataset, we have used *reduction* strategy with this dynamically changing data in interactive map. We have used *filters on items* such as – type of food to eat (breakfast, lunch, dinner), city of choice and option of takeout. Using proper features from R shiny, *Aggregation of items* has done for the city of choices to find food to eat.

A Sample view of the tool:



What do you fancy? Anything

In which city?

☑ To take out?

Antonio's Pizza
★★★★☆
619 E Green St Champaign
IL 61820

Show 10 entries

Search:

	Name	Address	Stars
3	Klondike Resto-bar Grill & Supper Club	100 Rue Sainte Anne Sainte-Anne-de-Bellevue, QC H9X 1N1	2.0
4	CrÃªperie Bretonne Ty-Breiz	933 Rue Rachel E MontrÃ©al, QC H2J 2J2	3.5

Lastly, to make this visualization tool complete as a *proper product* for customers, we have used two more things, one is a list below the interactive map listing a number of establishments with several attributes and another is a kind of *partition side by side views* to *lookup* the details of the selected establishment by using second tab available on dashboard. Once, a user will select a particular establishment on map or either on the list, all the available reviews of that one specific restaurant will be shown on the next tab and will be presented sorted with time and other details available from Yelp, so that user will be able to confirm his final decision or choice. The list below the interactive map shows only the establishments whose location at a moment is present in the interactive map geographically, similar to how ratings *barplot* and interactive map are *faceted* using *share navigation*.

3.4 Algorithmic Implementation

This is the innermost most level which involves creation of all the design choices which we discussed above. The objective is to efficiently handle the visual encoding, user interaction and the data processing regarding memory and performance of the computing system.

We have implemented the whole visualization tool using R and shiny. Due to large structure of Yelp dataset, in the starting, we performed the analysis using only a subset of Yelp data. But, our tool works fine even with a heavy subset of yelp data set too and show all the possible restaurants in 4 countries. Major library which came handy in the whole implementation exercise is the leaflet and DT library, along with this we have also used ggplot2 library. Leaflet is a standard and widely used library in the context of thematic cartography and geo-visualization. An important implementation issue was to employ the reactive expression for interactive map using shiny, a crucial and time-taking decision for making the whole thing computationally faster. Another issue was implementing the share navigation reduction strategy between the barplot and interactive map, this took us a lot of time and significant programming experimentation to achieve.

4. Conclusion and future work

This whole project of designing an interactive data analysis tool gave us lot of useful insights and tips for creating interesting and useful visualization tools. The Yelp dataset itself is huge dataset perfect for machine learning, clustering, visualization, NLP and social graph mining. The complex, varying and huge nature of this dataset presents novel challenges in front of data scientists which are very beneficial for learning process. From the day one of assignment, we had a feeling that we are working on developing a visualization product, not just another assignment. Unlike other assignments, we didn't jump on programming and developing codes from the start, but for this tool, we actually followed the real life approach of thinking on the behalf of users.

In the starting classes, we had several discussions with our professor who himself acted as a potential user and gave us some very useful insights such as – the zooming feature, the concept of reactiveness, and the overview of yelp dataset. After making our first prototype on R and shiny, we interacted with other classmates and friends and everyone like the concept of the tool, and gave us valuable suggestions for improvement and possible directions. There are still lots of possibilities left to explore with a tool of this type and with the kind of dataset Yelp is. This whole exercise has guided us an immensely in contemplating from the user side too when designing a visualization too cum product.

But the visualization tool is still in a very early phase, a lot of things can be improved. For example the reviews tab layout is probably not the best one for presenting the available data. Also improvements in terms of performance could help the tool to be more interactive and to handle more data.