Activity Distribution Analyzer

Shubham Goel, School of Information, UC Berkeley

Abstract—This tool is designed to analyze the user action logs generated from tools like Tableau. It enables us to inspect the frequency of different tasks performed, especially when the tasks are categorized using a multi level hierarchy. This tool highlights the tasks that are commonly performed together giving a unique perspective of the data. The tool can be adapted to be used with any similar dataset.

Index Terms—exploratory data analysis, data visualization, tableau, multi level hierarchy.

1. Introduction

Data analytics is commonly used to examine user behavior. Usually the goal behind such practice is to make the tool or service better. The analytics enables the tool creators to understand the part of the system that are working well and the ones that are buggy. In our case we had a slightly different goal. As a part of the 'Exploratory Data Analysis(EDA) research seminar' we wanted to identify some of the common trends and best practices in the world of EDA. To that effect all the students in the class performed exercises exploring different data sets using Tableau Desktop analysis tool[1]. Some of these exercises were performed with specific questions in mind. Others were general data analysis to understand the dataset and extract insights from the data. The students recorded each step they performed in a tabular format. They recorded this data along with timestamps and screenshots of their actions. For each action they also reported the goal, reasoning and the end result of that action. The software generated logs were also submitted.

In this paper we are going to look at a tool build to analyze all this custom data reported by the students. The goal is to visualize this complex data so that it can be more clearly understood. The data may have many hidden patterns in it, like do different students perform their analysis in a similar way for the same dataset and the same objectives, are there some tasks that are repeated over and over, are some actions that are always followed by another specific action and does this behavior depend on the user.

2. Data Cleaning And Transformation

There were a total of three exercises. In the first exercise the students picked one out of three datasets and came up with a hypothesis to explore. In the remaining two exercises the class analyzed the same data set. In one case they had to formulate their own hypothesis and in the other they were given a question to answer. The reported data consisted of:

- Automatically generated Tableau Logs
- Actions reported along with goals and results
- Screenshots

The tableau logs contains a ledger of different actions that the software performs like loading libraries, system calls, authentication calls and user performed actions. For the purpose of this tool I used the user actions extracted by Sara Alspaugh, our Graduate Student Instructor for the course. These actions had been extracted by looking for the key "command-post" in the log files and storing the corresponding values as a list of json objects. Another key resource I used was the multi-level taxonomy/hierarchy developed by Hassan Jannah[2]. See table 1 for a small subsection of the taxonomy. At the root level all actions are grouped into five basic categories which are further sub divided into action level 2 and action level 3 items. Action level 3 items represent a small collection of specific Tableau actions.

[1] http://www.tableausoftware.com/products/desktop

[2] https://docs.google.com/spreadsheets/d/12z-_9Eg43U8jWZh1swYJVSovNDeW6fdI6z8fMj705_s/edit#gid=0

Action Level 1	Action Level 2	Action Level 3	Tableau Actions
App Actions	Datasource Actions	Analyze	tabdoc:show-me
App Actions	Document Actions	Close	tabdoc:move-dashboard-edge
App Actions	Datasource Actions	Connect	tabui:save-workbook
App Actions	Datasource Actions	Connect	tabdoc:set-sheet-formatting
App Actions	Datasource Actions	Connect	tabdoc:move-free-form-zone
App Actions	Document Actions	Exit	tabdoc:move-zone

Table 1. Registre Tableau Taxonomy

2.1 Analysis

I imported these action logs for all users and the corresponding taxonomy into an iPython notebook. I wanted to analyze these logs and look at the distribution of different actions and see if any of the actions were commonly performed together. To this effect I transformed the data and build new logs that incorporated the taxonomy terms in them. Essentially these new log files where a collection of level 1, 2 and 3 actions. I wanted to look at actions that are commonly performed together. So, I grouped the actions together as bigrams and trigrams for each level of hierarchy. To look at the distribution I plotted the histograms of some of the data. Figure 2 shows the distribution of bigrams for one of the assignments plotted at action level 3 from the taxonomy. Even thought the top two bars in the figure are not very informative, the next few bars have some interesting results. We can see that actions like query and remove are commonly performed together.

Next I wanted to get the data in a format where I could import it into a visualization using the D3 javascript library. The entire data set was divided into a collection of files. Each student exercise had three corresponding files, one for each action level in the taxonomy. I chose to convert this data into a json format. I picked a format which could store a tree data structure. Its a basic graph with no cycles, one predefined root node and all nodes are connected to the network. Each node in the json file had a name, key and size associated with it. Action level 3 nodes had a color associated with them too. The size represented the frequency of that action or action n-gram and the color coding indicated the associated level 1 action. For a total of 23 valid assignments 69 files were generated using a python script.

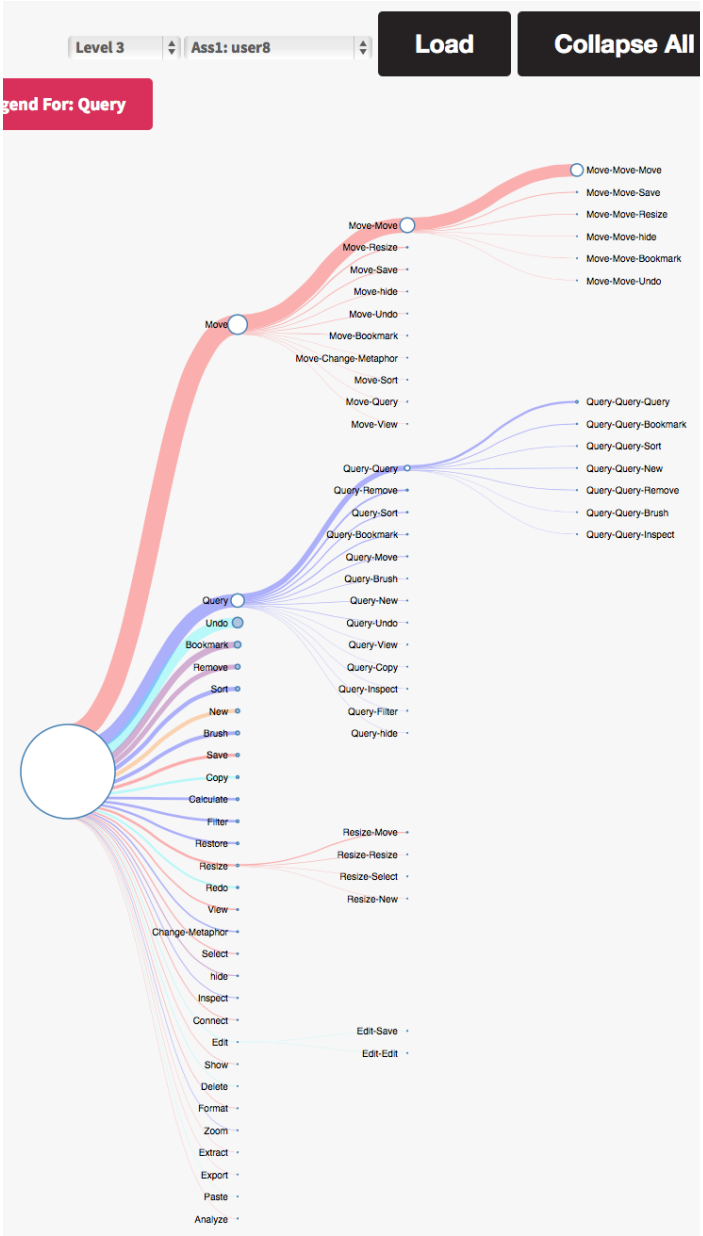

Fig 1. Activity Distribution Analyzer tool

3. Tool

http://shubhamgoel.me/viz/ADA/

The histogram is a very effective way of looking a the distribution of a data but it does not support any further interactions. I wanted to build a tool that would allow the user to drill down on the actions they were interested in to see what commonly followed and to be able to switch between different action levels and exercises quickly. I chose a

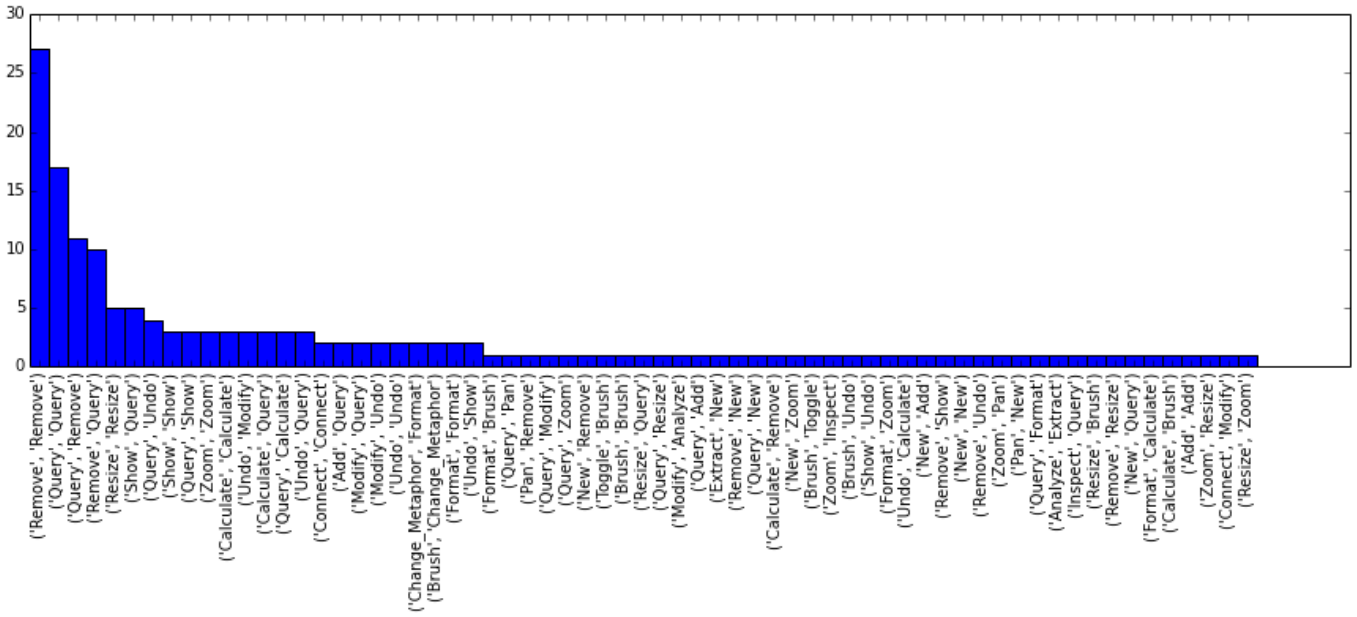

Fig 2. Histogram of action level 3 bigrams

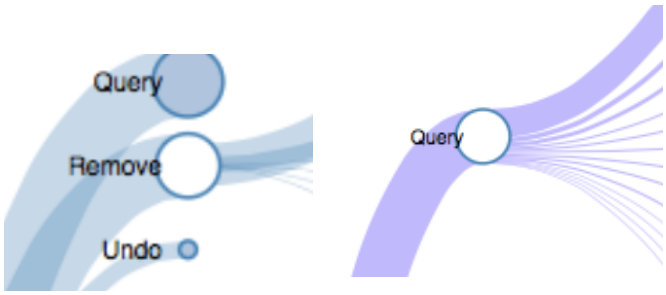

Fig 3. Distribution of the outgoing links
Left: before Right: after

collapsing tree structure to represent the data. This would allow the user to drill down on the data and also only look at the information they were interested in. The nodes at the first depth level indicate the action unigrams. The user can click on a node to look at the tasks that are performed after the action unigram and further drill down on the tree to look at the trigrams. The weight of the links connecting the nodes indicate the frequency of the action sequence. The input data and action levels can be changed using drop down menus. The user can look at the entire exploded view of the tree using an “expand all” button or close the tree using a “collapse all” button. One of the small but import feature of the visualization is the way incoming and outgoing links are presented. Usually in a network graph the links originate and end at the center of the node. In our case that lead to overlapping of originating links from a node. Although this is functionally correct, it created a visually displeasing effect. To fix this I made modifications such that the thickness of the incoming link is the same as the thickness of the outgoing link as can be seen in figure 3.

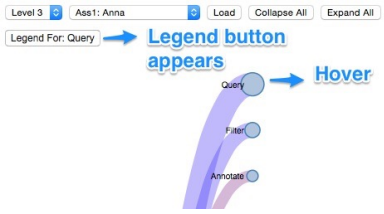

Fig 4. Legend Button for Query Action

3.1 Dynamic Legends

One of the unique challenges with this tool was to show case the associated legends and help the user understand the taxonomy without referencing the documentation. To solve this problem I use a new kind of dynamic legend. I created a new tree which represented the taxonomy action levels. In this tree the root level nodes are action level 1 items followed by level 2 and level 3 actions respectively. The leaf nodes are the actual tableau actions. In the main visualization a button appears when the user hovers over an action level 3 node. This can be clicked to reach the taxonomy tree. The taxonomy tree dynamically opens the relevant branches to the action level 3 node selected, making it easy for the user to see the underlying

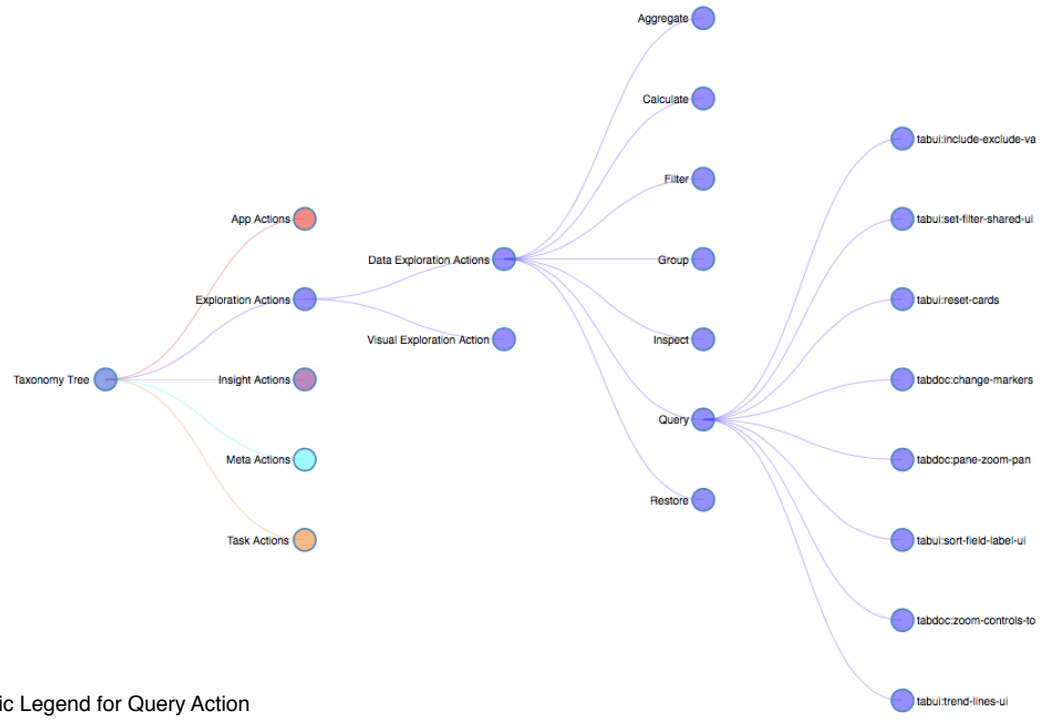

Fig 4. Dynamic Legend for Query Action

tableau actions. In figure 4 we can see the legend button for the query node. Figure 5 shows the corresponding dynamic legend tree for the query action.

3.2 Technology

I used the D3 javascript library to build this tool as its one of the most advanced visualization frameworks out there for building networked graphs that work natively on the web. For the initial version I started out by modifying one of the existing sample codes on the mbstock website[3]. I was able to find one example of D3 based tree diagram with weighted edges and color coding, but I was unable to take any inspiration from it because the code had been obscured.

Some other notes about the tool:

- all the private information of class students has been removed
- tool has been open sourced under the MIT License

3.3 Utility

This tool makes it easy to look for patterns in the data if any. One example would be to look at data from user 13 at level 3 taxonomy. Query and undo actions are used more heavily as compared to other users and some of the query are followed by an undo action. When compared with user 9 we see that this user uses the show function and the move function more. Similar comparisons make it easy to see that there is a difference between the workflow of these users, even though they both use the query function the most. This gives the query function a context which would have been hard to see without looking at the bigram distribution.

4. Other Usecases

This visualization can be extended to other logs and datasets. One example could be visualizing activity data from health tracking devises like the Basis watch[4]. This wearable is capable of tracking data like the heart rate, number of steps walked, body temperature, perspiration etc. The data can be extracted from the basis website as a csv that has a row for every minute of data. This data has a lot of granularity. Like the activity data can be

[3] http://mbostock.github.io/d3/talk/20111018/tree.html

4 http://www.mybasis.com

classified as walking, running, biking or sleeping, which in a way is a multi level hierarchy. Our tool would enable the user to see patterns in this data set like what is the typical heart rate range associated with a workout or the body temperature associated with sleep.

Another data set that can be analyzed in a similar manner is the console log generated by mac applications. The console log has details about when the application was opened, usage duration, sometimes the actions performed, network calls, etc. Visualizing this data in a multi level hierarchy could be potentially useful for identifying the sequence of actions that lead to a crash or a bug.

5. Conclusion

Building this tool was a great learning exercise for me. It forced me to think about the different facets of a network dataset. The log data can be simply thought of as a list of actions in chronological order, but when those actions are aggregated it creates a complex directed graph with cycles. Categorizing this data with a multi-level hierarchy adds another dimension to the data.

By exploring the data using this tool I found out that the trigrams were not very useful. They did not show me a side of the data that was interesting. Bigrams on the other hand allowed me to dive deeper into actions and understand the context they are being used in.

6. Links

Live: http://shubhamgoel.me/viz/ADA/index.html

Github: https://github.com/shubhgo/EDA-drill-down-tree

Video: http://youtu.be/ju6HRIIAJMc

7. Acknowledgements

I would like to thank Prof. Marti Hearst and Sara Alspaugh for their guidance and feedback.