Browser-based Categorization of Data towards Automated Visualization

March 6, 2013

Abstract

Data is being generated in ever greater amounts. Visualization is one the most effective ways to quickly analyse and explore a large amount of data. There is a good deal of literature surrounding various individual visualizations. However the complete end to end process has not received much treatment, that is taking an arbitrary dataset and visualizing it effectively. Publications that do deal with this topic generally rely on an expert to manually complete the process.

This paper investigates the feasibility of an automated end to end system that can visualize an arbitrary data set. Emphasis is placed on areas where previous publications have relied on human actors to complete the process.

A browser based system is presented. It consists of two main components. The first is a simplified chart API. Each chart automatically visualizes an input dataset. The second is a classification engine that determines which chart to use for an input dataset.

A number of visualizations have been implemented. These visualizations form an API that is useful without any additional components. Each visualization has an associated set of properties. These properties are then matched against an input dataset to determine if the chart is suitable. The properties of a dataset are determined at runtime.

Summary

Automatic visualization of an arbitrary dataset is possible to a point. Creating a system that can visualize data is relatively simple. Determining if the output visualization is suitable or even logical is a much harder problem. It is difficult, for example, to determine the relations of variables in a dataset without any additional information, or even to determine the type of an individual variable. These tasks have typically been left to human actors. For this reason several assumptions have been made.

Visualizations are selected based on properties of a given dataset. They are rated on how effectively they can represent these properties. There has been a good deal of research on the limits of human cognition regarding graphics. The findings presented in these studies are used to determine how effectively a set of properties can be represented by a given visualization.