Browser-based Categorization of Data Towards Automated Visualization

Steven Diviney

March 31, 2013

Abstract

This paper presents a system to automatically generate suitable visualizations given arbitrary data. Data Visualization is an increasingly common technique used to reinforce human cognition. In many areas of human activity the volume of data being generated is increasing. New methods must be employed to assist in the comprehension of this data. There has been a good amount of research performed to assess what factors contribute to the creation of an effective visualization. Many new and novel visualizations have been created. However, automatic generation of visualizations has received little attention.

Such a tool would help to combine these two areas of research. An understanding of what factors contribute to an effective visualization are encoded into the system presented. Given an arbitrary dataset the system attempts to select the most appropriate visualization. This paper also discusses to what extent this process is viable.

Using the limited amount of information contained in a raw dataset it is possible to select a comprehensible visualization. As the dataset becomes increasingly complex the effectiveness of such a system diminishes. A number of datasets are input to the system and an evaluation is undertaken. The results presented show that such a system can be used to assist users in the creation of suitable visualizations while avoiding the creation of inappropriate or even misleading visualizations.

Contents

1 Introduction

1.1 Motivation and Description

DRAFT Information visualization is a growing area of research. Presently there is a good amount of discussion surrounding new and novel visualization techniques and how to create effective visualizations [**?**] but very little on the process of creating a visualization. There have been a number of books published on the subject . These works outline the steps needed to visualize different types of data but do not attempt to automate the process.

The majority of new techniques exists in some degree of isolation. The majority of individual implementations exists in near complete isolation, often with the dataset hard coded in There are a number of products that attempt to create a complete end to end process for visualizing data sets. These products are either highly specialized or lack complex visualizations and require user input throughout the process .

Highly specialized applications such as gretl can afford to make numerous assumptions about the input dataset. It is assumed they will be used as part of a specific suite of tools and as such are able to directly process the proprietary output of such tools. Such output is often rich with meta-data which is used to assist the visualization process .

General purpose tools such as Microsoft Excel contain a number of simple charts and graphs. They require a basic level of user training to create and offer no assistance in selecting the most appropriate chart for a given dataset. This often leads to unsatisfactory, confusing or even misleading results .

With a few exceptions, such as gretl, these tools are proprietary and lack documentation on the techniques they use .

They system presented here aims to address these deficiencies by providing a fully automated end to end visualization tool for arbitrary datasets. There have been some notable projects that accomplish such a goal. Emphasis has been placed on areas where previous works have relied on human actors to complete the process.

-User Need -Gap in research -Expose deficiencies of state of the art -Outline what was taken from SOA, outline what was added.

1.2 Research Question

I need help with phrasing this. -Objectives. Be clear and consistent.

Mazza, Shneiderman, Approche Graphique, Haskell etc

How do you back up a claim like this?

ref state of the art

again, state of the art, gretl would be a good example, need more

Excel

Example of stock trading software, although this might stray into NDA territory. Excel also does this for stock charts

Show a really bad example of excel output

This is a rather bold claim

NAME it

Polaris, Mackinlay. Need to look back over these and pick out what was expanded upon.

1.3 Evaluation

-Probably going to be paper based

1.4 Overview

-Go into contents of each chapter a bit -Expose the argument thread. THIS is the focus. Stay with argument.

2 State Of The Art

2.1 Automated Visualization

2.1.1 Polaris

2.1.2 A Presentation Tool

Published in 1986, perhaps too old?

2.2 Visualization Process

2.2.1 The Eyes have it

2.2.2 Introduction to Information Visualization

3 Design

3.1 Technologies used

3.2 Overview of Data Visualization Process

This will probably be broken down

3.3 Architecture

Section may be somewhat redundant due to section 2.2

This will probably be broken down