Browser-based Categorization of Data Towards Automated Visualization

Steven Diviney

April 15, 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 rapidly. 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. Additionally 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

10/04/2013
changes in
onenote

This section introduces the dissertation topic and explains the motivation behind the work. A research question is purposed and a number of objectives are outlined in an attempt to satisfy it. This is followed by evaluation criteria and finally by a summary of the document structure.

1.1 On Visualization

There are three fields subfields of Data Visualization. The boundaries between them are not particularly distinct and they are referred to somewhat interchangeably in academic literature. There are many other types but these three are of particular interest as their primary concern is the visualization of large volumes of data almost always with the aid of a computer.

1.1.1 Information Visualization

Information visualization is perhaps the most broadly used and can be thought to encompass all of the fields of visualization. The term today is generally applied to the visual representation of large-scale collections of non-numerical information, such as text in a book or files on a hard-disk. The distinction between this definition and that of Data Visualization seems to be quite poor. The type of the data in question is used to distinguish the two. The terms "information" and "data" are not so easily distinguished. Information can be thought of as a level of abstraction above data. The mere fact that a dataset is non-numerical does not identify it as information. A set of ordinal labels or categories is just as meaningless as a set of numbers. Information is created by organizing such data and presenting it with context.

Information visualization is concerned with representing more abstract topics. A good example is process visualization. Each element in a process visualization represents a complex topic.

1.1.2 Data Visualization

Data Visualization is the science of visual representation of data, defined as "facts and statistics collected together for reference or analysis" (oed31, 2013). As stated the distinction between data and Information Visualization

is not very concrete. This distinction is generally not regarded as important but this paper is concerned with Data Visualization. The synthesis of new information through the creation of visual artifacts.

1.1.3 Scientific Visualization

Scientific Visualization is concerned primarily with the visualization of objects in three dimensional space with an emphasis on realistic rendering. This emphasis on realism is primarily what distinguishes it from other forms of visualization. This is not to say that the other forms may distort the data, rather that abstract data does not necessarily have a spatial dimension. How would one realistically visualize the lines of a book? Novel ways must be invented to accomplish this.

1.2 Motivation and Description

Data Visualization is defined as an internal construction in the mind. The field of Data Visualization is connected with creating visual artifacts in order to facilitate individuals in building an internal representation of a dataset (Spence, 2001). Data Visualization is not a new field but it has only become an established area of scientific research in recent years . The volume of data a typical computer can generate and process far exceeds what a human can comprehend. Data Visualization is becoming an increasingly popular technique to aid this comprehension.

Cite journal dates

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 . The articles presented in these journals typically focus on the creation of specific visualizations, new and interesting ways to graphically represent specific types of data. There is also a suitable body of literature concerning 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.

Cite IEEE Visualization Journal, Sage and others

Mazza, Shneiderman, Approche Graphique, Haskell etc

The majority of new techniques exists in some degree of isolation. Individual implementations typically exist in near complete isolation, often with the dataset hard coded into the software. There are a number of products that attempt to create a complete end to end process for visualiz-

How do you back up a claim like this? Perhaps get rid of it

Ref state of the art or just leave for SOA?

ing 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 .

This paper 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.

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

Excel

Either cover this in SOA or get rid of it

Show a really bad example of excel output

This is another rather bold claim

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

1.3 Research Question

The objective of this project is to determine to what extent can suitable visualizations be dynamically and automatically generated using browser based technologies. The Automatic Classifier and Data Visualizer, or AC↯DV is presented. The goals for the project are as follows:

- Design and develop software capable of accepting arbitrary data in a specific format and display it using suitable visualizations. This should be done without an intervention from the user.
- Access the level of benefit that the visualizations can bring to potential users from various fields.
- Investigate to what extent visualizations can be generated given only an input dataset.
- Elaborate on the potential of a more sophisticated version of the software using additional techniques to determine features of input datasets.

Browser based technologies have been selected due to their wide-scale support. The goal of the system is to simplify the process of visualization creation. Browser based technologies require no setup or configuration and thus simplify the process.

But it's not an end user app so is this sufficient justification?

1.4 Evaluation

The system will be evaluated by inputting a number of datasets and assessing the output against the previously stated goals. A number of use cases have been drawn up to determine how beneficial such a tool is in aiding various users in visualizing data. . ACϟDV is not intended as an application suitable for end users so the user experience is not relevant to the evaluation. The goal of the system is to quickly produce effective visualizations that represent the data accurately.

Switching between tenses a bit here

Key to the evaluation is the notion of suitable visualizations. In order for a visualization to be considered useful it must meet several criteria. This will be outlined in detail in later sections . Visualizations are composed of many individual elements with different attributes such as colour, size and spatial location. These elements have been the subject of previous research and a number of guidelines exists outlining their usage . However these guidelines are not hard rules and their exists no formal way to evaluate a complete visualization. Methods from the field of Human Computer Interaction are typically used, specifically user evaluations and trails. ACϟDV generates visualizations that are well understood, thus eliminating the need for lengthy user evaluations concerning the effectiveness of the individual output visualizations. The guidelines stated above are used by the system to pick appropriate visualizations. These guidelines will be used to benchmark the effectiveness of visualizations produced by the system.

Ref appropriate section

Cite Mazza, Jock etc. CALL them retinal variables

1.5 Overview

This chapter is followed by a survey of the State of the Art in the areas of automated visualization and the visualization process. Section 3 examines the design of the visualization tool and gives an overview of the process of Information Visualization. Section 4 gives a complete description of the projects implementation and any problems encountered. Section 5 is an evaluation of the project and a discussion of its merits and failures. The

extent to which such a system is viable is also discussed in this section. The paper concludes with a look at potential future work that could extend AC↯DV.

2 State Of The Art

This section presents work currently being undertaken in two main areas of research discussed in this paper; automated visualization and the visualization process. An analysis of these projects conclude this section.

2.1 Visualization Process

2.1.1 The Eyes have it

2.1.2 Introduction to Information Visualization

2.2 Automated Visualization

The focus of this paper is automated visualization, taking a dataset and generating a visualization from it without any user input. There are a few projects that incorporate some kind of automated data visualizer. Generally such a visualizer is part of a larger project and requires user interaction to function. This section presents two such projects with varying degrees of automation.

2.2.1 Polaris

Polaris extends the well known pivot table interface to display information visually. Multiple visualizations are displayed on a pivot table which the user can interact with by selecting or "brushing" data-points to filter the displayed data. The visualizations act as interactive query builder that allow user to explore large datasets quite rapidly. First published in 2000 it has evolved into Tableau, a commercial visualizations product.

citation

Polaris consists of two main components. A graphic generator and a database query generator. The database query generator will be omitted in this discussion as it is not relevant to AC↯DV. The graphic generator makes several assumptions about the nature of the input data. These assumptions are also made by AC↯DV. Data is characterized as either quantitative or ordinal. Nominal data is assigned an ordering by the system and treated

Should this just go in design?

as ordinal data. Polaris assumes all nominal fields are dimensions, or independent variables, and all quantitative fields are measures, or dependent variables. Three different types of graphics can be generated based on the input data. User input is then used to refine this to a single generated graphic.

The user interacts with a relatively complex UI. The UI contains a number of "shelves" onto which the user places data-sources and data records. The user also selects what type of mark to associate with each record. The selected mark and nature of the data is used by the system to determine what type of visualization to render. For example, if the data-source contains an ordinal and quantitative field and the user selects a bar as the mark to use a bar chart is generated. If the user selects a dot as the mark a dot plot is generated. An understanding of Bertins retinal variables (Bertin & Barbut, 1973) is encoded into the system to ensure the visualizations are comprehensible. It is not clear how Polaris handles input that cannot be effectively displayed using these rules, e.g. when the input data set is too large to be displayed.

Polaris allows multiple series of data to be overlaid on the same visualization but it does not appear to allow more than two dimensions of data to be displayed at the same time. As it is a data exploration tool it requires user input too function. While it does contain an understanding of retinal variables these are used to automate the generation of mappings between data and graphical marks. It does not contain any rules governing how the various retinal variables are combined so it may be possible for a user to create ineffective or incomprehensible visualizations .

Go into a bit more and cite

2.2.2 A Presentation Tool

Published in 1986, perhaps too old?

3 Design

3.1 Technologies used

3.2 Overview of Data Visualization Process

This will probably be broken down

Section may be somewhat redundant due to section 2.2

3.3 Architecture

This will probably be broken down

4 Implementation

5 Evaluation and Discussion

6 Future Work and Conclusions

References

Bertin, J., & Barbut, M. (1973). Sémiologie graphique. Mouton; Paris: Gauthier-Villars.

Oxford English Dictionary Online, 2nd edition. (2013, July). http://www.oed.com/.

Spence, R. (2001). Information visualization. Addison-Wesley.

A Appendix A