Browser-based Categorization of Data Towards Automated Visualization

Steven Diviney

April 18, 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

Chapter 1

Introduction

10/04/2013
changes in
onenote

This chapter 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.

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 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 .

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.

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:

Cite IEEE Visualization Journal, Sage and others

Mazza, Shneiderman, Approche Graphique, Haskell etc

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

Ref state of the art or just leave for SOA?

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.

- 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. . AC↯DV 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. AC↯DV generates visualizations that are well understood, thus eliminating the need

Ref appropriate chapter

Cite Mazza, Jock etc. CALL them retinal variables

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.

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. chapter 3 examines the design of the visualization tool and gives an overview of the process of Information Visualization. chapter 4 gives a complete description of the projects implementation and any problems encountered. chapter 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 chapter. The paper concludes with a look at potential future work that could extend AC↯DV.

Chapter 2

State Of The Art

This chapter 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 chapter.

2.1 Visualization Process

2.1.1 Introduction to Information Visualization

Introduction to Information Visualization (Mazza, 2009) is one of the most recent texts providing a state of the art in the field of information visualization. As well as listing many examples of newer more novel visualization techniques in the later chapters it also contains a substantial amount of information about the process of information visualization. This set of first principles has influenced the work presented in this paper quite heavily. The author acknowledges that given an arbitrary set of data there is no way to determine, automatically or otherwise, what visualization is a suitable representation of that data. There is in fact no agreed criteria to evaluate a finished visualization.

However there are certain characteristics of data that can be used to help select an appropriate type of visualization. These steps are outlined as follows:

- Define the problem: Is the goal of the visualization to communicate some meaning in the data, allow the user to explore the data or confirm a hypothesis.

- Examine the nature of the data to represent: What type of data is being dealt with; ordinal, quantitative or categorical.
- What is the dimensionality of the data: The number of attributes of the dataset. In a univariate dataset one attribute varies with respect to another. Bivariate and trivariate extend this with multivariate extending to four or more dimensions.
- What is the structure of the data: This can be more easily understood by defining what types of data structures can be used to store the data; linear data structures such as vectors and tables, temporal data, spatial data hierarchies and networks.
- What type of interaction will the user have with the data: can the user transform the data by modifying it, can the modify parameters used to display the data or is the visualization static.

This process is summarized in Table 2.1 (Mazza, 2009). It will be revisited in later sections.

Problem	Data type	Dimension	Data Structures	Type of Interaction
Communicative	Quantitative	Univariate	Linear	Static
Explore	Ordinal	Bivariate	Temporal	Transformable
Confirm	Categorical	Trivariate	Spatial	Manipulable
		Multivariate	Hierarchical	
			Network	

Table 2.1: Variables to consider when designing visual representations.

This process eliminates visualizations that are not suitable for a given data set. Deciding on an individual visualization to use is a decidedly harder task. The effectiveness of a visualization is dependent on the perceptual capabilities of the viewer. As there is no empirically verified model of human perceptual abilities one is purposed. This model is based off the work of Jacques Bertin (Bertin & Barbut, 1973). Bertin was the first person to define a set of "retinal variables"; how different properties of graphics can be used to convey various types of information. His work has been extended since it's original publication and the model presented here is perhaps one of the most up to date. A list of retinal variables is presented in order of effectiveness conveying different types information. This idea will be

revisited throughout this paper and a complete model will be detailed in a later chapter.

Mazza's work is a very complete introduction to information visualization and draws attention to all of the factors that need to be considered when creating a visualization. It is perhaps the most complete state of the art currently available and builds on a large amount of influential publications in the field. It has been very influential in the design of the system presented in this paper.

2.1.2 Structure of the Information Visualization Design Space

Structure of the Information Visualization Design Space provides an overview of the field of data visualization as it was in 1997. An organization of information visualization literature is purposed and several examples are provided. The organization technique is then used as a framework for new designs. The paper builds on Bertins work (Bertin & Barbut, 1973) which has been expanded on by Mackinlay (Mackinlay, 1986). Visualization techniques are grouped based on similarities of their data to visualization mappings.

A series of steps that need to be preformed to create a visualization are given.

Data Data is defined as a set of values taken on my a set of variables. It is an obvious prerequisite of any visualization. Different types of data have their own characteristic operations and subcategories have yet more unique operations. For example patent text or a financial report have their own characteristics and unique operations. In order to do any useful processing of data a more general model is used. Data is divided into three categories.

- Nominal, denoted by N
- Ordered, denoted by O
- Quantitative, denoted by Q

Data is further divided into the original dataset D and data D' that has been selected from this set and possibly transformed by some filter F.

Visualizations Visualizations are defined as being composed marks, their graphical properties and elements requiring human controlled processing. Human visual processing works on two levels.

- Automatic Processing: Works on visual properties such as position and in colour. These are Bertins retinal variables (Bertin & Barbut, 1973). This processing occurs at a subconscious level and as a result is very fast but limited in power.

- Controlled processing: This can be easily thought of as conscious thought. It is very powerful but limited in capacity. These limitations are fairly well understood (Miller, 1956).

Visual presentations consist of a set of marks (points, lines etc), a position in space (X and Y) and a set of retinal properties (colour and size). This distinction is a somewhat different grouping to the ones purposed by Bertin and Mackinlay. One can only assume the reason for this grouping is to limit the number of attributes or combination of attributes of a visualization. Connectedness and enclosure are also given as attributes.

These attributes are then arranged in table 2.2 to categorize various visualizations. The version shown here is simplified.

Name	D	F	D'	X	Y	R	$--$ []	CP
Year	Q	$>$	Q	P				
Quality	Q	$>$	Q		P			
Type	N	$>$	N			C		

Table 2.2: Bar chart classifications. R, retinal properties. C, colour. $--$ Connection. [], Enclosure. CP, Control Processing.

Structure of the Information Visualization Design Space presents a good overview of the state of visualization research. The classification method is useful but is continuously amended to account for different types of visualizations.

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 chapter 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 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.

Should this just go in design?

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

A Presentation Tool, henceforth referred to as APT, is an application-independent presentation tool capable of automatically designing effective visualizations of relational information. It attempts to create a wide variety of designs and encode graphic design criteria in a useful form. It achieves this by defining graphical presentations as sentences in a graphical language. An expressiveness and effectiveness criteria are used to codify the graphic design criteria. The "sentences" produced by these two systems are then combined using artificial intelligence techniques to create designs.

APT is divided into three parts, expressiveness, effectiveness and composition. Expressiveness determines how the input information can be expressed in a graphical language. It is encoded into a formal language. This languages main purpose is to ensure that any graphic encodes all of the input data and only the input data. It is not used to select entire graphics, rather it is used to test individual graphical mappings, i.e. to a single axis or type of mark.

Double check expressiveness and effectiveness synchronous w.r.t each other.

Effectiveness is is used to determine which of the sentences generated by the expressiveness phase is most effective at displaying the data in a useful manner It is dependent on the capabilities of the user. There is a difficulty in that no verified theory of human perceptual capabilities exists so one is purposed. It is based off Bertins retinal variables (Bertin & Barbut, 1973) which have been ranked in order of effectiveness by Cleveland and McGill (Cleveland & McGill, 1984). This ranking has been extended to include rankings ordinal and nominal types of data although this extension has not been empirically verified .

Does Mazza match up with this ranking?

The system generates multi-dimensional graphics by combining the effectively one dimensional or single data series sentences generated in the expressiveness and effectiveness stages. Designs are merged on common data. The types of data to merge on are ordered in terms of effectiveness.

- Mark composition.
- Double axis composition.

- Single axis composition.

APT combines these three stages using artificial intelligence techniques. The algorithm has three steps: partition, selection and composition. Each step contains various choices. If these choices do not lead to design backtracking is used to consider others.

- Partitioning: Each set of relations, or columns in the database, are partitioned to match the criteria of of one of the sentences defined in the expressiveness stage.
- Selection: A list of candidate designs for each partition is ordered in terms of effectiveness.
- Composition: The individual designs are composed into unified presentations of all the input data using the composition algebra.

Overall APT is a very robust and complex system. However it is perhaps somewhat dated. It is only capable of generating static graphics. It can also factor in the output media into the design choices, i.e. if the screen is monochrome colour is omitted for the effectiveness stage.

2.3 Analysis

Presently there is a good amount of literature surrounding individual visualization as demonstrated by (Mazza, 2009), and large degree of consensus on how to go about creating a visualization. However there have only been a few attempts to automate such a system. Such an application would not only allow users to quickly produce useful and accurate visualizations but would also server to validate the process that has become agreed upon. Such an evaluation is frequently omitted from these publications. Mazza outlines a set of criteria to evaluate individual visualizations (Mazza, 2009) but the focus of this thesis is testing if the entire process is valid.

Works such as Polaris and APT suggest that this process is indeed valid. However Polaris does rely on the user for some tasks. APT appears to be a very robust and complex system but is somewhat dated and cannot

Replace the system with relevant name

Example in notes of bar-chart rule.

More bad things because APT is pretty cool

Published in 1986, perhaps too old?

Cite Structure of information

citation

citation

generate visualizations with any interactivity. Given the ubiquity of powerful computing platforms this seams to be quite a large deficiency. The project presented in this thesis aims to create an automated version of the visualization process.

This thesis will attempt to determine to what extent can suitable visualization be dynamically and automatically be generated from arbitrary data using browser based technologies. At no point should the user have to provide input. Browser based technologies have in recent years become very powerful platforms and are perhaps the most ubiquitous software platforms in existence.

2.4 Conclusion

Replace section with chapter in the text

Chapter 2 introduced a state of the art automated and manual visualization processes and how they can be used to simplify the creation of visualizations from arbitrary data. By making the process of creating a visualization easier more data can visualized and thus comprehended more readily. Many existing systems are limited by requiring human intervention. If the user is untrained they can introduce errors which may lead to confusing or even mis-representative visualizations.

Chapter 3

Design

3.1 Technologies used

3.2 Overview of Data Visualization Process

This will probably be broken down

3.3 Architecture

chapter may be somewhat redundant due to chapter 2.2

This will probably be broken down

Chapter 4

Implementation

Chapter 5

Evaluation and Discussion

Chapter 6

Future Work and Conclusions

References

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

Cleveland, W., & McGill, R. (1984). Graphical perception: Theory, experimentation, and application to the development of graphical methods. Journal of the American Statistical Association, 79(387), 531–554.

Mackinlay, J. (1986). Automating the design of graphical presentations of relational information. ACM Transactions on Graphics (TOG), 5(2), 110–141.

Mazza, R. (2009). Introduction to information visualization. Springer.

Miller, G. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. The Psychological Review, 63, 81-97.

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

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

Appendix A

Appendix A