Implementing a Horizontal Size of KDGEs Information Classification and Visualization System

Steven Diviney
08462267

January 26, 2013

Abstract

This article discusses the field of Information Visualization. My research project is discussed briefly to provide context, but the bulk of the text deals directly with the field of information visualization. First it is defined. A brief summary of its history is then given followed by my attempt to explain how a visualization helps accelerate cognition and the challenges encountered in attempting to do so. Some well established methods of visualization are then briefly presented and analyzed. Finlay evaluation criteria for the project are defined.

1 Introduction

Information visualization is defined as an internal construction in the mind. The field of information visualization is concerned with creating visual artifacts in order to facilitate individuals in building an internal representation of a dataset (Spence, 2001). In this paper the term visualization refers to the creation of these visual artifacts unless explicitly stated as the cognitive process of understanding an image.

There are three fields subfields of information 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 Information Visualization

Information visualization is perhaps the most broadly used and can be thought to encompass all of the fields of visualization. After all, almost anything if sufficiently organized, is information of a sort (Friendly & Denis, 2001). Friendly notes that tables, graphs, maps and even text, whether static or dynamic, provide some means to see what lies within, determine the answer to a question, find relations, and perhaps apprehend things which could not be seen so readily in other forms.

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.

Agile pro
gram or

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. In researching for this paper I have come across several examples of one labeled as the other. As far as I can tell there have been no calls to address this issue and the two are used interchangeably to a large extent.

Interestingly the majority of literature I have come across states it is concerned with information visualization, and then goes on to outline steps to transform data into a visual artifact in order to help the synthesis of information. From this it would seem quite clear that the distinction is not regarded as important but I would like to point out my topic of study is concerned with data visualization. The synthesis of new information through the creation of visual artifacts.

1.3 Scientific Visualization

Fortunately scientific visualization is very well defined. It 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.

1.4 KDEG's Information and Visualization System

KDEG, the Knowledge and Data Engineering Group in Trinity College Dublin are developing a system to automate the process of data visualization. The creation of a visualization is typically quite a manual process. There are an abundance of tools that, given a dataset, produce specific types of visualizations. These tools expect the data to be formatted in a certain way and for the user to manually specify certain attributes of the data to the visualization tool. They also require a domain expert to contribute some knowledge about the nature of the data in order to create information. Typically the domain expert requires the assistance of a knowledge engineer to encode this knowledge into a format a machine can use (Hampson, 2011).

KDEG are attempting to build a system that can take data from multiplier sources, classify it, determine the form of the data and associate it with appropriate visualizations. Information about the data can be added by a domain expert without the aid of a knowledge engineer. What they lack at the moment is a method to take this data and associate it with appropriate visualizations.

My task is to implement a complete horizontal slice of this system. Given a dataset that has been augmented with expert knowledge my system must determine the nature of the data to represent, its dimensionality, the data structures needed and associate it with an appropriate visualization. The input data has been restricted to few sets that share certain attributes. Implementing a compete system is not feasible in the time frame. The goal of developing such a system is to document the process of data visualization which is poorly understood. The steps outlines above, determining dimensionality, data structures etc. are sometimes done automatically in certain cases but are typically linked together manually. An overview of the steps needed to create a visualization is presented below.

1.5 Visualization Synthesis

The creation of a visual artifact follows there successive stages (Mazza, 2009). These steps are by no means definitive. One of the major problems within the field of data visualization is the lack of consensus on a single process.

- Preprocessing and data transformations.
- Visual mapping.
- View creation.

1.5.1 Preprocessing and data transformations

The initial phase is concerned with taking raw data, that is data supplied by the world, or datasets, and organizing them them into a logical structure suitable for machine processing. Additional information can be added in this preliminary step. Filtering operations can be used to eliminate unnecessary data and calculations can be used to obtain new data, i.e. the summation of particular record instances.

1.5.2 Visual mapping

This stage determines the form of the data. Again, abstract data does not always have a location in physical space associated with it, so a visual mapping is needed. Mazza notes "The spatial substrate defines the dimensions in physical space where this visual representation is created" (Mazza, 2009). It is defined in terms of axis. The nature of the data determines this mapping. Quantitative, ordinal and nominal data are all reported in different ways along an axis. The graphical elements used to display the data and their properties are also defined at this stage.

1.5.3 View creation

The view renders the visual representation on the display of the computer. The main issue here arises when the quantity of data is too large for the display. There is a large body of research surrounding this problem as it is very common. My thesis will attempt to tackle some of these issues. In order to utilize the available screen space efficiently the data must be rendered in order of importance. This will be explained further in a later section.

2 History of Visualization

Visualization is one of the oldest fields of human inquiry. The earliest examples are of geometric diagrams and maps to aid navigation. The earliest known map is generally regarded to be a map of "Konya", a town in Turkey, dating back to 6200BC (?). Humans have perhaps been using graphical representation of the world around them to convey information for far longer, up to 14,000 years ago (?). Diagrams used to illustrate concepts rather than spatial features did not appear until much later.

Visualization has only recently become a field in its own right. Up until now it was simply a tool used to convey data. As such it had to wait for advancements in other fields to progress. The 16th century saw significant advancement in the techniques and instruments used for precise observation and measurement of physical quantities. Astronomers, surveyors and cartographers needed to convey greater amounts of increasingly precise information. The field of mathematics was also seeing the birth of probability theory, analytic geometry and theories of estimation and measurement.

The 19th century saw an explosion of statistical graphics (Friendly & Denis, 2001). Bar charts, histograms, line graphs. Statistics was becoming an increasingly important field and began to see use in social planning, commerce, transportation and other areas. For the first time visualization was used to explain data rather than just supplement its description with an illustration.

One of the most famous examples of the use of visualization to convey an idea is the dot map used to illustrate the source of a cholera epidemic in London. John Snow, an English physician theorized that cholera was spreading through the cities water supply, not the air as was accepted at the time. Snow's chemical examinations of the water from the Broad Street pump were not sufficient to prove it as the source. Rather his studies of the pattern of the disease were what convinced the local authorities. He plotted each case of cholera on a map. The map clearly shows the water pump at the center of the outbreak (?).

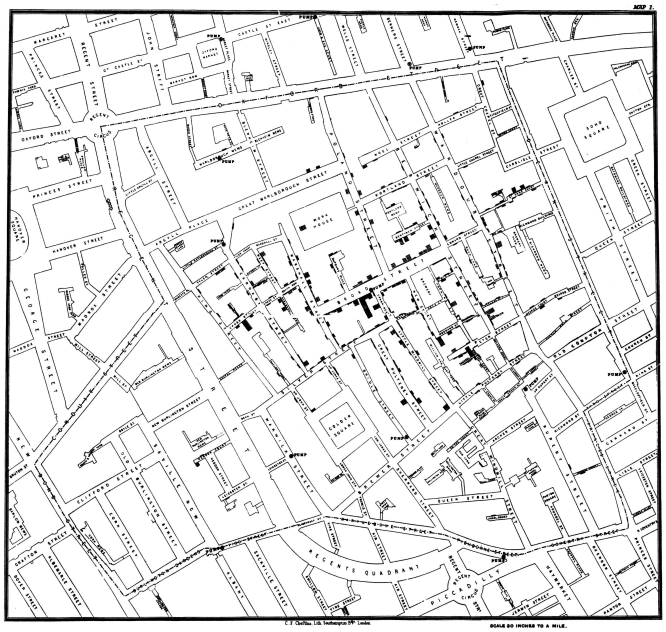

Figure 1. Dot map showing cholera cases per household.

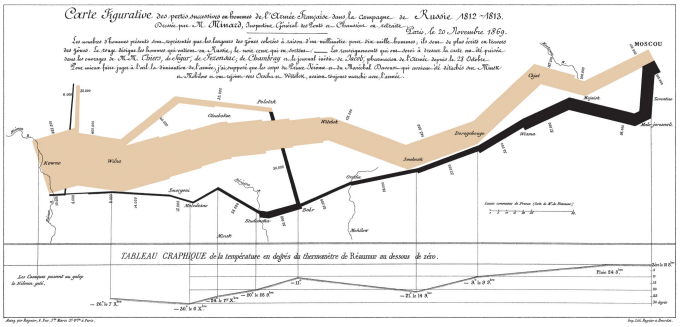

Figure 2. Napoleon's march on Moscow.

Visualization is a field heavily intertwined with human psychology. This will be examined later in greater detail but I draw attention to it here to illustrate just how concise an effective visualization can be. Human perception, particularity visual perception, is a very complex process. However certain visual attributes, such as size or colour, are given perceptual preference. They are what we focus on and process first (?). A visualization that exploits this fact effectively is arguably far more effective than any textual description. Charles Minard's flow map of Napoleon's march to Russia is perhaps one of the most famous visualizations ever created, shown again and again in academic literature. The graphic conveys several pieces of complex information at once. The size of the army corresponds to the width of the line and is the most obvious feature. It provides a strong representation of the human suffering. Rivers along the route can be identified by a sudden loss of life. The line shows the direction and route the army took. The date and temperature are also recorded along the path.

3 Information Synthesis

4 Evaluation

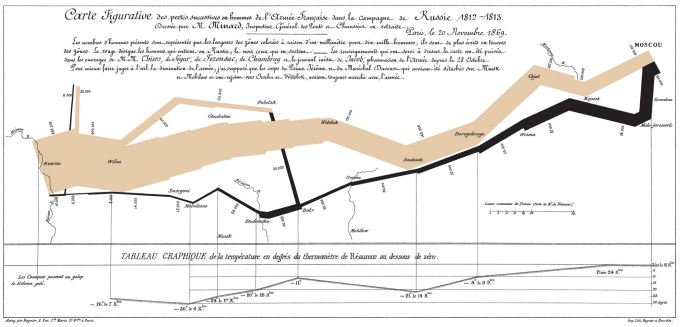

Figure 3. Napoleon's march on Moscow.

References

Friendly, M., & Denis, D. (2001). Milestones in the history of thematic cartography, statistical graphics, and data visualization. Accessed: March, 18, 2010.

Hampson, C. (2011). An expert supported approach to data exploration. Unpublished doctoral dissertation, Trinity College Dublin.

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

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

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