Abstract

This article discusses the field of information visualization. It is a sort of introduction. My research project is discussed briefly but the bulk of the text deals directly with the field of information visualization. First it is defined. A brief summary of it's history is then given followed by my attempt to explain how a visualization helps accelerate cognition. Some well established methods of visualization are then briefly presented. Finlay evaluation criteria are defined.

Implementing a Horizontal Size of KDGEs Information Classification and Visualization System

Steven Diviney

November 22, 2012

1 Introduction

Visualization has come to have a wide range of uses. It is defined as an activity in which human beings are engaged as an internal construction in the mind. It is a cognitive activity facilitated by external visual representations from which people build an internal mental representation (Spence, 2001). In this article the term visualization will refer to the creation of a pictorial representation of data unless explicitly stated as the cognitive process of understanding an image.

For our purposes we will define three subfields of visualization. Data visualization, information visualization and scientific visualization. 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 broadest 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. However, 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.

1.2 Data Visualization

Data Visualization is the science of visual representation of data, defined as information which has been abstracted in some schematic form, including attributes or variables for the units of information (Friendly & Denis, 2001). The distinction between data and information visualization is not very distinct. In researching for this paper I have come across several examples of one labeled as the other. Here we will simply define data visualization as the visual representation of numerical data and information visualization as the representation of non-numerical data. The former should conjure images of scatter-plots and bar-charts while the latter generally results in somewhat more elaborate visuals, such as time-lines or treemaps. This distinction seems rather arbitrary, I see the two as largely interchangeable.

1.3 Scientific Visualization

Lastly there is scientific visualization. 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 Visualization Synthesis

In general the creation of a visual art follows three successive stages(Mazza, 2009).

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

1.4.1 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.

1.4.2 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. 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.

1.4.3 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.

1.5 KDEGs Information Classification and Visualization System

KDEG, the Knowledge and Data Engineering Group in Trinity College Dublin, are developing a system to automate this process. Up until now the development of a visualization has been a somewhat manual process. There are an abundance of tools that, given a dataset, produce one or several types of visual representation. In the vast majority of cases these tools expect data in a specific format. Additionally, there a large number of tools to assist with the preprocessing and transformation of data. Both of these tools, particularly the latter, rely on the user to contribute some knowledge about the nature of the data. KDEG are attempting to build a system that will take a data-source, or multiple sources, classify it, determine the form of the data and associate it with an appropriate visualization. This visualization is then rendered to produce the result. They system provides an interface a novice user can use to explore the

data. It also provides an interface for an expert to provide context where it is need to refine the process.

My task has been initially defined as implementing a horizontal slice of this system. I will take a number of distinct data-sets that share similar properties and attempt to visualize them by building a restricted version of the system above. Parts of the system have already been implemented by KDEG and they will be used where appropriate. The point of developing such as system is to automate the manual transition between steps. As such a key consideration of my project is to develop a system appropriate to the nature of the data, not to any one data-set. The classification and visualization of data are two very distinct task, which we will get too, and I suspect one will receive more treatment than the other. Already I admit the classification task has not been given much consideration and has been completely omitted from this essay. I present two excuses; first, the classification engine has already been implemented by researches in KDEG so I doubt I will spend much time working on it and second is that I am simply more interested in using visualizations to help accelerate cognition, I don't particularity care what that cognition relates too as of yet.

2 A Brief History of Information Visualization

Information visualization has very deep roots. The earliest examples arise in geometric diagrams and maps to aid in navigation. The earliest know map has several claimants but is generally regarded to be map of "Konya", a town in modern-day Turkey. It is thought to date back to 6200 BC (Bagrow & Skelton, 2009). There are many more examples in the subsequent centuries, largely that of maps but also of diagrams used to illustrate concepts, although these appear much later.

It was not until the 16th century that techniques and instruments for precise observation and measurement of physical quantities were well developed. These inventions were born in the fields of astronomy, surveying and cartography. It was also during this time a great new growth occurred in theory and practice. The birth of probability theory, the rise of analytic geometry and theories of estimation and measurement to name a few. This is when humanity began to amass data, along with theories to make sense of it and visual representations.

It was not until the 19th century that modern data graphics began to emerge. The steady fertilization provided by the previous centuries resulted in an explosion of statistical graphics (Friendly & Denis, 2001). Bar charts, histograms, line-graphs and many others were invented. By the middle of the century many states had established statistical offices in recognition of the importance of numerical information in social planning, commerce, transportation and other areas. This is when statistics began to gain wide-spread social significance and visualization was used to make sense of large bodies of data.

Perhaps the most famous example of this 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 (Frerichs, 2007).

It might be said that this is when data visualization underwent its first revolution. Up until now the bulk of activity outside of cartography had been rather ad-hoc. The latter half of the 19th century saw the emergence of standard models that could be used to display different types of data. This was coupled with the acknowledgment of statistics as having real social value and a need to make sense of the increasing amount of data being gathered. The agreement on first principles and the influence of social circumstance are two of the most influential factors that lead to the adoption of a paradigm (Kuhn, 1996).

I am however somewhat reluctant to award information visualization the status of paradigm thus far. From my analysis there does not appear to be much competing schools of thought, nor any debate on the correct way to create a visualization. Rather individuals, or small groups of individuals, have invented novel ways of displaying information at various points in history, usually when the need arises. This reached more of a tipping point rather than a crisis point as the need for standard techniques became apparent. Whether this disqualifies a field from being considered a science is a matter for another essay. For the moment I will continue to apply the model of paradigm as it fits rather well. Later I hope to show that modern field of information visualization is indeed a science.

Progress in the field slowed dramatically around the turn of the 20th century. Statistics had become more precise and many statisticians regarded pictures incapable of stating fact, at least with no where near the of precision as their new statistical models (Friendly & Denis, 2000). Another way to look at this period is one of normal science. Graphical methods began to enter textbooks, saw use in government, education, commerce and science (Haskell, 1919) (Ayres, 1919) (Gantt, 1919). Visualization increasingly became used to help analyze data rather than just represent it.

From 1950 to 1970 information visualization underwent a seismic shift, spurred by three developments (Friendly & Denis, 2001);

- "The Future of Data Analysis" issued a call for the recognition of data analysis as a legitimate branch of statistics (Tukey, 1962)
- "Semiologie Graphique" (Bertin & Barbut, 1973) organized visual and perceptual components of graphics according to their features and relations between data. It is widely regarded as the "periodic table" information visualization.
- Computer processing of data had begun. It offered the opportunity to create new graphical forms and processes quantities of data that had before been unpractical.

From the mid seventies until the present day the field has developed at an accelerated pace. Better computers have resulted in a largery variety of highly interactive visualizations. New methods have been developed for visualizing higher dimensional data, e.g. scatter-plots, parallel coordinates. These developments have largely depended in advances in theory and tools.

The second half of the twentieth century is where, I think, we begin to see the inclusion of psychology in the field. "Semiologie Graphique" categorized various graphical properties and later studies classified them in terms of our ability to process them (Cleveland & McGill, 1984).

Today information visualization is closely intertwined with Human Computer Interaction. It can be thought of as discipline with two distinct parts. The first is how to process the data, what structures can be used to represent it while preserving important aspects. The second is concerned with how effectively we interpret these visualizations, how well they help us create our own internal visualizations. The former has already accumulated a good deal of normal scientific literature while publications surrounding the latter represent that of pre-paradigmatic psychology as we know it today: hypothesis derived from empirical studies of human behavior.

3 Information Synthesis

If the aim of information visualization is the creation of visual artifacts to amplify human cognition it would seem useful to investigate human cognition, or more specifically the process of understanding data. The process has been defined as a "continuum of understanding" (Jacobson & Wurman, 1999). It consists of four stages;

- Data: entities that lack any meaning on their own. The are the bricks from which we build information.
- Information: the transformation of data into information is accomplished by "organizing it into a meaningful form, presenting it in meaningful ways and communicating context around it" (Jacobson & Wurman, 1999).
- Knowledge: information is integrated with experience.
- Wisdom: a person has acquired levels of knowledge to such a level such that they are able to express qualified judgment on data.

The last point is up for debate. Information visualization lies somewhere between the first two so we need not concern ourselves with it. Information visualization is concerned with providing the tools to organize and represent data to produce information. The aim is to improve the human cognitive process through visual representation of data by making use of our perceptual abilities.

How does a visual representation help us to understand data? Spence (Spence, 2001) suggests we use these visualizations to build our own mental models of the data. A mental model is something cognitive psychology scholars use to describe how we build knowledge of the worlds around us. For example, if someone takes the same route to work everyday they will become familiar with it very quickly and will no longer require the use of a map. This does not mean they have memorized a copy of the map, rather they can recognize reference points associated with their mental model.

Visual representations can boost cognitive processes because they allow some inferences to be done very easily (Card, Mackinlay, & Shneiderman, 1999). They make use of our visual perceptive abilities we have gained through our evolution.

3.1 Perception

If a visualization aims to enhance cognition the attributes of the data under consideration should be mapped to graphical elements in a certain way. Patterns

in the data must be easily perceivable. The "trick" to this is to make certain visual forms stand out. Our knowledge of human memory is what helps us to do this. Cognitive psychology identifies several types of memory. We will consider three based on the duration of their retention.

- Sensory or preattentive memory. This memory has very short retention, only a few hundred milliseconds. It stores the visual information from the eyes and is independent of conscious control. During preattentive processing a limited set of visual attributes are detected. These include colour, size, orientation etc. We have evolved this preattentive processing to help prioritize the massive amount of data coming from our vision system. We exploit the characteristics of this memory to make certain elements stand-out from others.
- Short term memory. This is where we begin an attentive process of perception. It can store about 7 items. An example of its usage for outr purposes would be remembering the labeling of a certain visual element while observing its distribution on screen.
- Long term. This is less of a concern as the way to "store" something in long term memory is periodic rehearsal or meaningful association.

A number of preattentive properties have been identified (Treisman, 1985). They can roughly be categorized as colour, form, movement and position. Each of these categories have certain attributes onto which we can map attributes of data. The effectiveness of each mapping is somewhat understood but there exist no explicit rules, only various sets of guidelines (Few, 2004)

3.2 Visual Structures

As stated earlier there already exists a good deal of literature around different visualization techniques. These techniques lie firmly within normal science. The reasoning behind them is either absent or lies in the field of psychology. They can be split into two categories, Multivariate Analysis and Networks.

Multivariate Analysis deals with representing data with 4 or more dimensions, that is data that cannot be easily represented with "standard" techniques, i.e graphs and bar charts.

Networks are used to represent relational data, i.e. flow charts. Again, there are already many well understood methods to represent network data.

Both of these techniques have interactive components.

4 Evaluation

Visual representations generally have some sort of specific task, usually to help some type of user make sense of some sort of data. They should therefore have evaluation criteria. The discipline of Human Computer Interaction uses analytic and empirical techniques to evaluate the effectiveness of human computer interaction. This evaluation process generally uses the following criteria (Dix, 2004)

- Assess the functionality of the system, or verify the system does what the user wants.

- Analise the effect of the system on users.
- Identify possible problems with users.

There are several criteria used to measure these goals. Given a set of initial requirements the can be evaluated with respect to their functionality, effectiveness, efficiency, usability and usefulness.

There are two methods to carry out such evaluations; analytic methods and empirical methods. Analytic methods employ experts to evaluate a visualization against a set of heuristics, or perform tasks they think a user may perform. Empirical methods use controlled experiments and qualitative methods to conduct evaluation.

The methods of evaluating a visualization first struck me as being rather soft. I later discovered that is because they are soft. They are the same methods employed by psychology. As such the principles of visualization are more merely guidelines. They are well understood but there usually are many variants. Happily visualization also contains a large section of problems subject to normal science, i.e. building the visualizations after their structure has been decided.

References

Ayres, L. (1919). *The war with germany: a statistical summary*. Govt. print. off.

Bagrow, L., & Skelton, R. (2009). *History of cartography*. Transaction Pub.

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

Card, S., Mackinlay, J., & Shneiderman, B. (1999). *Readings in information visualization: using vision to think*. Morgan Kaufmann.

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.

Dix, A. (2004). *Human-computer interaction*. Prentice hall.

Few, S. (2004). Show me the numbers: Designing tables and graphs to enlighten.

Frerichs, R. (2007). The ghost map. *Emerging Infectious Diseases*, 1134.

Friendly, M., & Denis, D. (2000). Discussion and comments. approche graphique en analyse des données. the roots and branches of modern statistical graphics. *Journal de la société française de statistique*, *141*(4), 51–60.

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

Gantt, H. (1919). *Organization for work*. London: Allen and Unwin.

Haskell, A. (1919). *How to make and use graphic charts*. Codex book company inc.

Jacobson, R., & Wurman, R. (1999). *Information design*. Mit Press London.

Kuhn, T. (1996). *The structure of scientific revolutions* (Vol. 2). University of Chicago press.

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

Spence, R. (2001). *Information visualization*. Addison-Wesley. Available from `http://books.google.ie/books?id=gYQoAQAAMAAJ`

Treisman, A. (1985). Preattentive processing in vision. *Computer vision, graphics, and image processing*, *31*(2), 156–177.

Tukey, J. (1962). The future of data analysis. *The Annals of Mathematical Statistics*, *33*(1), 1–67.