Jestr

Journal of Engineering Science and Technology Review 5 (1) (2012) 72-76

JOURNAL OF
Engineering Science and Technology Review

Research Article

www.jestr.org

Visualization Techniques for Large Datasets

M. Michalos[1], P.Tselenti[1], S. L. Nalmpantis[2]

[1]*School of Computing, Information Systems & Mathematics, Kingston University, London, United Kingdom*
[2]*Dept. of Electrical Engineering, Kavala Institute of Technology, Kavala, Greece*

Received 18 February 2012; Accepted 14 July 2012

Abstract

In order to improve understanding and working with data, visualizing information is without a doubt the best method to implement. Data visualization as a term unites the established field of scientific visualization and the more recent field of information visualization. The goal of data visualization is to provide the viewer an aggregated representation of available data by taking into account human's visual system and its influence to comprehension. Spotting trends, seeing patterns and identifying outliers are some of the human's visual system processes that are being manipulated in order to make data more accessible and appealing. This procedure of graphical representations creation helps engaging data exploration and even more, data extraction. Along with computer and graphical engineering, visualizations have grown and reached a very satisfactory level of variations and techniques, indulging even the most exacting data facilitators whether they are researchers, computer scientists, statisticians etc. A variety of data visualization software has been developed the last decades but Stanford University's Protovis is by far the most distinguished tool to do the job. Below, a study is presented on data visualization's purpose and prospects and how these became a necessity through time.

Keywords: Visualization techniques, datasets, data extraction.

1. Introduction

Since large datasets are here to stay, their visualization became an art, or even better, a science, which develops images or visual representations from large quantities of information and data. As Ziemkiewicz and Kosara [1] recently mentioned, "The power of information visualization arises from its ability to apply perception and visual thinking to understanding complex data and solving difficult analytical problems". Over the past decades, [2] and specifically from the early days of computer graphics back in the 1950s, graphic design has developed along with computers in such way that there is a large choice in which you can decide according to specific needs what kind of graphic to create. This specific desire to manipulate objects on a computer's monitor has driven force behind many popular users to design new interface paradigms [3]. Whether data refers to time, meters, population or whatever else, visual representations can be creatively constructed to meet the needs of even the most pretentious receiver.

2. Why bother visualize data?

In order to understand the importance of visualizing data we should take under consideration how John Tukey [4] put it: "There is nothing better than a picture for making you think of questions you had forgotten to ask". What's hidden in this quote is that when looking at a figure it should make the receiver understand it better than he already does. Exploring figures, images and representations is an easy thing to do, but when having a closer look, one should get a message, a graphical visual imagery, which is eventually the purpose of visualization. In fact, visualization is not an exploratory process but it should be taken under consideration from a humanistic perspective as a way to communicate because it can deliver messages through data presentation, rather than just a way to present information. It is also a fact [5] that people use visualization to support comprehension and further reasoning by having a look at interactive abstract visual imageries as natural scenes with sets of implied dynamics between objects. As Ziemkiewicz and Kosara [1] state inferences are then easier to be reached and to acquire a meaning, while all those objects as well as implied dynamics are being transferred from a metaphoric level to a substantial, through human senses of course.

Clearing and exploring data is another purpose of visualization and that is possible through the capabilities of human's visual system. We can process large amounts of information because our eyes can translate a figure into the message given through visualizations, as mentioned above. For example, it is easier when visiting a toll post to see in which category your vehicle belongs to by a sign which represents figures of vehicles, than having to pay attention to a large sign with letters and detailed description of every vehicle category. As a result of the above, visualization's power is placed in the ability to apply perception in order to understand and solve complex data.

Apart from understanding the essentiality of visualizing,

* E-mail address: k1065769@kingston.ac.uk
ISSN: 1791-2377

what should be mentioned is the handling and the explosive growth of generated data. When referred to a dataset there must be a simple rank in which the amount of information given should be placed. Furthermore, it is a fact that over the time, this kind of rank changes since the data generated some years ago are way much smaller in digital quantity than today. For example, Google's CEO Eric Schmidt recently mentioned [6] "Every two days now we create as much information as we did from the dawn of civilization up until 2003, that's something like five exabytes of data". This data generation is clearly taking place due to the constant evolution of sensing, networking and data management.

Eventually, why bother visualize data? There are many reasons, and as the time goes by, new ones come up. As mentioned above, enormous sizes of data are constantly generated throughout the web making data management a difficult mission to accomplish or even to comprehend. The most significant is that important messages can be delivered in a simplified but also attractive imagery and the fact that large information datasets can be represented in a simple way in order to display results.

3. Data extraction through visualization

After meeting the necessity of visualization and before having a look on the top visualization techniques it is important to analyze an intermediate phase that is called data extraction. Data extraction is the ability to retrieve data from an information representation (in our case, dataset visualizations) in order to further elaborate or just store all necessary information. For example, this process could be used for gathering data in order to develop a business plan that is based in sales, costs, employee salaries etc.

Data extraction became indispensable because [7] visual representations were massively developed along with the corporal and scientific growth from 1975 until today and results had to come to a close. These results were impressed in visualizations, organized in a variety of fields so that they could be manipulated for further elaboration, as mentioned above, specifically for corporal and scientific purposes and even for teaching material in the academic area.

Apart from extracting data through a visual system, many software companies (like Microsoft[1], Senchalabs[2], QlikTech[3], and Tibco[4]) have developed programs to cope with automatic information extraction from datasets for producing visualizations. This information afterwards is being automatically developed by the software and then represented and organized in categories to meet the needs of what the receiver is specifically looking to extract. To sum up, data visualization could be considered as the most common data extraction method and apart from that, data extraction could be held subsequently even from the visualizations themselves!

4. State of the art visualization techniques

In most of the cases of displaying information and data that consist of hundreds maybe thousands of bytes a simple graphical representation is not enough. Over the past decades and because of the massive generation of data, techniques have grown up and reached a quite mature level of graphical display. Furthermore, many new ways and variations of data visualization constantly appear and this testifies that in design the only existing border is the creator's imagination.

4.1. Time series data techniques

In time series data techniques a specific value is examined for its course through time. It's the most common technique as it appeared in visualizations that were printed hundreds of years ago.

Stacked graphs (or stream graphs)

Stacked graphs as shown in fig. 1, have not been known since recently were in late 2006 Byron and Wattenbergin [8-9] created for last.fm[5] a visualization that looked like a river or stream of data pouring down or across the whole graphic. This is definitely a powerful visualization technique and can prove to be very useful but the problem with this kind of graphs is that they cannot represent a negative result which means that they can only grow in one direction or decrease in case they have already grown.

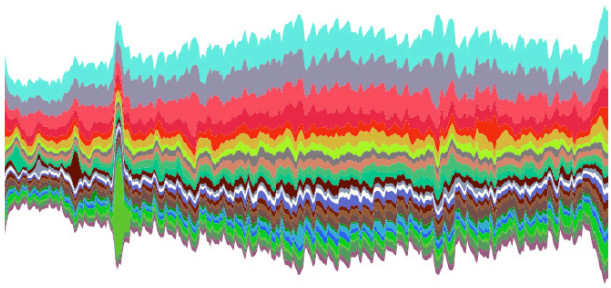

Fig. 1. Stacked graph [10]

Horizon Graphs

Horizon graph (Fig. 2) is a visualization technique [11] that helps to discern, analyze and compare multiple time datasets in a very efficient way regarding space. Visually, horizon graphs use a combination of length, like in single line graphs, and color. This kind of visualization could be used for example from a company that wants to examine in parallel some substantial pieces like stocks, product sales and other corporal data.

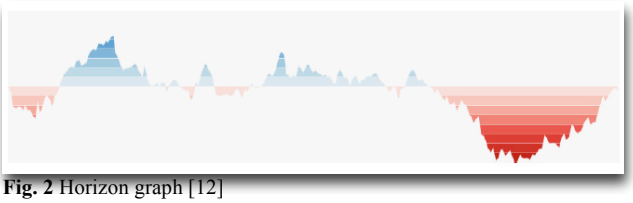

Fig. 2 Horizon graph [12]

4.2. Statistical distributions

Statistical distributions are, according to Van Hauwermeiren and Vose [13], "continuous distributions that are used to represent variables that can take any value within a defined range" and help in the comprehension of how numbers are

[1] http://www.microsoft.com/
[2] http://thejit.org/
[3] http://www.qlikview.com/
[4] http://spotfire.tibco.com/
[5] http://www.last.fm/

distributed.

Quantile Quantile Plots

Quantile Quantile Plots or Q-Q Plots (Fig. 3) is a visualization technique that [14] uses a distribution of values, the quantiles, and its basic purpose is to compare two distributions. The first one forms a probable diagonal line and the other is formed from the quantiles that are basically compared to the line and theoretically they must lie along.

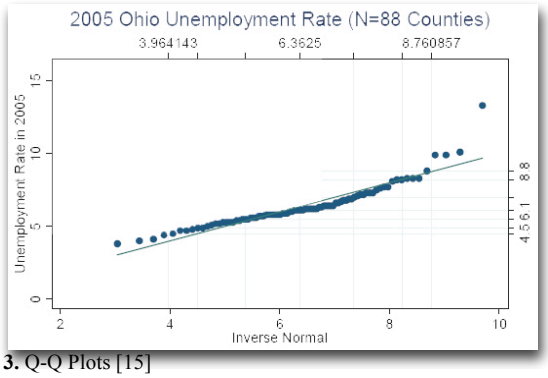

Fig. 3. Q-Q Plots [15]

Parallel coordinates

Parallel coordinates (Fig. 4), as Inselberg Dimsdale [16] have mentioned, is "a system for doing and visualizing analytic and synthetic multi-dimensional geometry". This visualization technique is appropriate for multivariate data with every variable declared on parallel axes. Although fitting models to complex data is a very difficult procedure, there are mathematical processes which are available for finding an optimal model within specific classes even for a single criterion [17]. The procedure then is much simplified, values are dispensed on the axes and then the axes are interconnected through the course of the values given from the dataset.

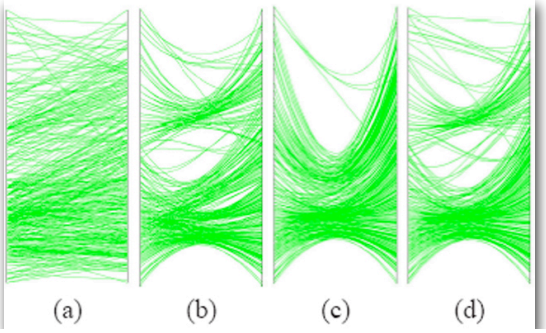

Fig. 4. Parallel coordinates [18]

4.3. Maps

Maps are often used for representing topographical and geographical information but they have been manipulated in such way that can exhibit very useful data.

Flow maps

In circumstances where [19] movements of people or objects take place on a map that have to be graphically represented, flow map (Fig. 5) is the best visualization technique to implement. Although it is visually simple to comprehend the imagery it is difficult to produce this data visualization because it exhibits high complexity due to the transition edges represented to correspond to the values required. Developing network flows and topologies [20] is a challenging procedure because the display of a large number of connections with lines may result into a visual of a clutter. Cartographers solved this problem through flow maps making their work easier as they can now illustrate distributions with lines of varying width representing the number of objects being transferred over the map.

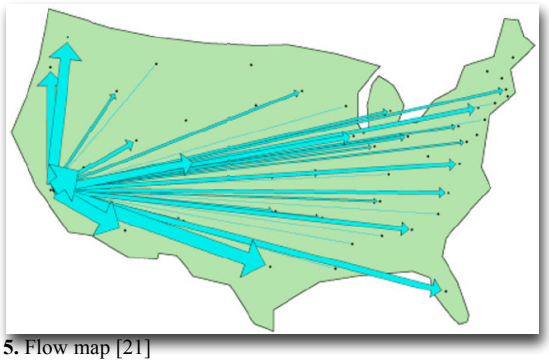

Fig. 5. Flow map [21]

4.4 Hierarchies

Data and information is not only about plain numbers but there are cases where all these figures are organized in a hierarchical system structure. This kind of visualization technique represents mostly tree-like imageries that form unfolding node graphics from one to many others.

Node-link diagrams

Node-link diagrams are the most popular tree layout visualization techniques and have many variations. The Cartesian Node-Link diagram (Fig. 6) is a very powerful visualization diagram which is represented by a circle where the first node is situated right in the center of it and the other nodes unfold to the perimeter of the circle. There is also the Radial Node-Link diagram where the head node is on top and all other nodes unfold vertically and horizontally in a symmetric order.

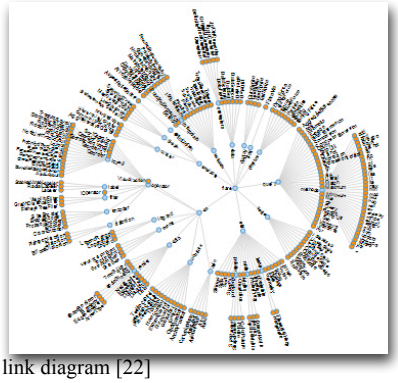

Fig. 6. Node link diagram [22]

4.5 Networks

Network is a category of visualizing data which copes with information that are interiorly related. For example [23] if someone wishes to explore in a social network who is friend with whom, then the representation to be created is situated right in this category.

Arc Diagrams

Arc diagrams (Fig. 7) are [24] made of one-dimensional layout and can be created though a pattern-matching algorithm which are used in order to find repeated substrings, and then represent them visually as translucent arcs. This technique was invented so that all possible pairs of matching substrings that are part of a subset could be graphically represented. This is eventually the main idea and usability of this visualization technique. Although this kind of visualization looks more appalling rather than delightful, it is a very common technique that is being used for its ability to provide representation of sequences. In some cases this diagram is used both ways, meaning that arcs are being designed on top but also at the bottom of the node row.

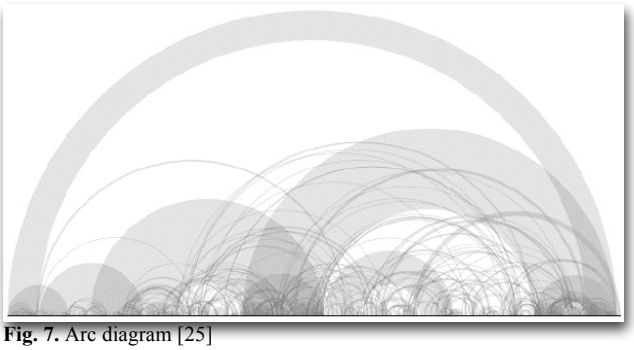

Fig. 7. Arc diagram [25]

Matrix views

A Matrix view (Fig. 8) could be conceived as a simple visualization technique but it is in fact a complex representation of values connected to each other. Graphically there are two columns with values; the point is to match the values from one column to another on a diagram composed by small boxes. The result is to create a neat pattern and if there are any sub-categories in the values then the outcome would be a lot more visually smashing with the use of colors [23]. In this last case with the use of color, one can easily spot clusters and bridges allowing in that way an interactive grouping of the diagrams' sub-categories and making it easier for the receiver to make a deeper exploration of the network structure.

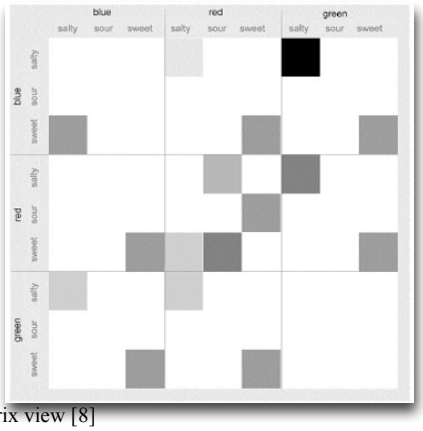

Fig. 8. Matrix view [8]

5. Protovis, a quick and simple solution to visualize data

When computers entered the world of science and education as a solution for better data exploitation, software companies started developing programs to match the calls of individuals that were in need of visualizing all the information they deal with. Over the past years more and more software solutions appeared in many variations of features and price.

Today, apart from all the fancy software in the market, there is a totally free program which copes with all the variations of even the most demanding visualization techniques. It is called Protovis and it was developed by Michael Bostock and Jeff Heer [26], both members of the Stanford University Visualization Group.

Protovis [26] is an open-source program which is provided free for individuals under BSD Lisence and its core is based on JavaScript and SVG for producing web-native visualizations. The only requirement for producing visual representations is for the individual to have a modern web browser installed on his computer. Furthermore, [27] it comes with many graphic libraries for drawing shapes and retouching pixels in order to produce communicational imageries by serving data in the most drastic and novel way.

To sum up, Protovis is a powerful graphical toolkit that was developed to cope with visualization, maintaining low-level control of graphical systems that deal with graphical components and presenting them in an expressive and accessible way.

6. Conclusion and future work

The current paper is an introductory step in data visualization for large datasets, from their purpose and prospect, to the top techniques and even a sample software to use. Several topmost approaches of handling feature-dependent (time, weight, income, etc) visualization datasets have been explored through this research though there are a lot more, less in graphic and building process. Data and information visualization is a quite large field for scientific and corporal research and, as previously mentioned, graphical development of datasets could be a quite influential process providing professional tools and elements for further improvement or individuals a way to receive interactive messages through attractive and novel representations.

Ultimately, as Ziemkiewicz and Kosara [1] states the goal of visual representations should be the application of implied dynamics, consisted by theoretical frameworks in order to interpret information structures into elements of a conclusive and explanatory visual structure. As for the techniques themselves, a small piece of them have been visited through our small course above; however, according to Heer et al. [23] there are still more visualizations that await discovery or even more, exist in the "wild". Researchers and designers constantly formulate new and creative representations or just improve the existing ones because of the perpetual elaboration of text visualization and data complexity.

Visualization techniques are used in many research areas for interpreting or just presenting large datasets and in the next few years, more powerful practices will be available for implementation. Furthermore, what should be taken under serious consideration is the fact that browsing today is being held from even more and more smaller handheld devices (cell phones, smart phones and PDAs) as this field of technology is in a constant evolution. These bring the need of developing dashboards so that visualizations could be delivered in small screens and comply with the latest features of these tiny devices. In addition, more immersive environments of working with information visualization

should be developed. The objective of making environments where information would be more accessible and easy to edit is not far away from becoming a reality. Last but not least, as Few [28] also states, because of the large amount of data, some visualizations tend to be complex, so that finally a clutter wall arises, leaving the receiver with more questions than answers. Researchers should be focused on methods involving algorithms in order to curb data and information for their later retouch.

It is everybody's desire to be surprised from the data visualization research field in the future: a new journey will then begin with a fresh top down approach in visualization techniques for large datasets.

References

1. Ziemkiewicz Caroline and Kosara Robert, "Implied Dynamics in Information Visualization", Proceedings of the 9th International Working Conference on Advanced Visual Interfaces, May 26 – May 28, 2010, Rome, Italy. ACM, 2010.
2. Post Frits, Nielson Gregory and Bonneau Georges-Pierre, Data Visualization - The State of Art. London: Kluwer Academic Publishers 2003.
3. Chen Chaomei, Information Visualization Beyond the Horizon. 2nd Edition. London, Springer 2006.
4. Tukey John, The Collected Works of John W. Tukey II. Time Series: 1965-1984. Monterey, CA.: Wadsworth 1985.
5. Chih Christine and Douglas Parker, "The Persuasive Phase of Visualization", Proceedings of the 14th Knowledge Discovery and Data Mining Conference, August 24 – August 27, 2008, Las Vegas, USA. ACM, 2008.
6. Eric Schmidt: Every 2 Days We Create As Much Information As We Did Up To 2003. 2010. [Internet]. Techcrunch. <http://techcrunch.com/>. [Accessed 3rd October 2011].
7. Card Stuart, Mackinlay Jock and Shneiderman Ben. Readings in Information Visualization, Using Vision to Think. San Diego, CA, USA: Academic Press 1999.
8. Wattenberg Martin, Visual exploration of multivariate graphs, New York, NY, USA: ACM. p.6, fig.7. 2006.
9. Suda Brian, A Practical Guide to Designing with Data. London: Five Simple Steps. p.135, Stream graph figure 2010.
10. Suda Brian, A Practical Guide to Designing with Data. London: Five Simple Steps 2010.
11. Panopticon - Visualize Multiple datasets and Compare trends with Horizon Graphs 2010. [Internet]. Panopticon. <http://www.panopticon.com/>. [Accessed 5th October 2011].
12. Stephen Few, Time on the Horizon. Visual Business Intelligence Newsletter. p.4, fig. 1., 2008.
13. Van Hauwermeiren Michael and Vose David, 2009. A Compendium of Distributions. [Internet]. Ghent, Belgium: Vose Software. <http://www.vosesoftware.com/>. [Accessed 4th October 2011].
14. Statsoft textbook 2010. [Internet]. Statsoft. <http://statsoft.com//>. [Accessed 11th October 2011].
15. Hun Myoung Park, "Univariate Analysis and Normality Test Using SAS, Stata, and SPSS*". Working Paper. The University Information Technology Services (UITS), Statistical and Mathematical Computing, Indiana University. p.6, fig. 3. 2008.
16. Inselberg Alfred and Dimsdale Bernard, "Parallel Coordinates: A tool for visualizing multi-dimensional geometry", Proceedings of the First IEEE Conference on Visualization, October 23 – October 26, 1990, San Francisco, CA, USA. IEEE, 1990.
17. Unwin Antony, Theus Martin and Hofmann Heike. Graphics of Large Datasets, visualizing a million. Singapore: Springer 2006.
18. Hong Zhou, Xiaoru Yuan, Huamin Qu, Weiwei Cui and Baoquan Chen, Visual Clustering in Parallel Coordinates. Oxford, UK: Blackwell Publishing. p.4, fig. 1. 2008.
19. Pieke Birgit and Kruger Antonio, "Flow Maps - Automatic Generation and Visualization in GIS", Proceedings of the 2007 GI-Days Young Researchers Forum, September 10 – September 12, Munster, Germany. GI-Days, 2007.
20. Phan Doantam, Xiao Ling, Yeh Ron, Hanrahan Pat and Winograd Terry. "Flow Map Layout, Proceedings of the IEEE Symposium on Information Visualization (InfoVis '05)", October 23 – October 25, 2005, Minneapolis, MN, USA. IEEE, 2009.
21. Phan Doantam, Xiao Ling, Yeh Ron, Hanrahan Pat and Winograd Terry, Flow Map Layout, Washington, DC, USA: IEEE Computer Society. p.1, fig. b. 2005.
22. Bostock Michael and Heer Jeff, "Protovis: A Graphical Toolkit for Visualization", Proceedings of the IEEE Symposium on Information Visualization (InfoVis '09), October 11 – October 16, 2009, Atlantic City, NJ, USA. IEEE, 2009.
23. Heer Jeffrey, Bostock Michael and Ogievetsky VADIM, "A Tour through the Visualization Zoo". Communications of the ACM, Vol. 53(6), pp. 56-67, (2010).
24. Wattenberg Martin, "Arc Diagrams: Visualizing Structure in Strings, Proceedings of the IEEE Symposium on Information Visualization (InfoVis '02)", October 28 – October 29, 2002, Boston, MA, USA. IEEE, 2002.
25. Wattenberg Martin, Arc Diagrams: Visualizing Structure in Strings, Washington, DC, USA: IEEE Computer Society. p.5, fig.10. 2002.
26. Protovis, a graphical approach to visualization. 2010. [Internet]. Stanford University. <http://www.stanford.edu/> [Accessed 10th October 2011].
27. Bostock Michael and Heer Jeff, 2010. Protovis, a graphical approach to visualization. [Internet]. Stanford University Visualization Group. <http://vis.stanford.edu/protovis/>. [Accessed 4th October 2011]. Main figure.
28. Stephen Few, Data Visualization, past, present and future. Perceptional Edge 2007.