Cover Story

Dr. KVSN Rama Rao*, Mr. Surya Putchala and Midhun Thaduru*****

*Prof., Dept. of CSE, MLR Institute of Technology,Hyderabad
**CEO and Chairman, Zettamine Technologies
***Associate Consultant, Zettamine Technologies

Visualization-Techniques, Methods and Tools

Why Visualization?

It is a well known fact that "A picture is worth a thousand words." To extract and analyze the massive amount of generated data, visualization plays an incredible role. Further it amplifies the cognition by helping in pattern detection and enhancing visual insight of a large quantity of data. It helps us to see data in context, analyze and discover knowledge. For companies across different industries – retail, logistics, banking and Finance, Insurance, energy etc., data visualization offer terrific opportunities to identify new products or uncover customer propensities that can provide insights of tremendous value to Businesses. For example, in a retail industry, increased use of geo-spatial visualization and analysis, the location of store, the difference in market size according to region, price and compensation studies in regard to specific regions, etc. reflect more clearly their potential advantage.

Data Visualization techniques often comes handy while representing large quantities of data and help making sense of big data and thus provide an exploratory platform for gaining deeper and clear insights. Some of the key functions of visualization are highlighted below.

- To present large volumes of data (structured or unstructured) effectively and elegantly.
- Provide a platform for exploring various facets of information aesthetically and interactively.
- Promote a deeper level of understanding of the data under investigation and assists us in drawing conclusions.
- Share information effectively to persuade, collaborate and emphasize important aspects of data.

A Simple Process for Visualization

As the data and the number of sources of information keeps growing, extraction of suitable information and presenting in a human consumable form becomes a great challenge. However, a data analyst can systemically increase the value of data through two major processes – one, Data Cleansing/Pre Processing and the second, Visualization process. Visualization is a six step process.

- *Mapping :* is encoding of data into visual form. It is used for achieving accurate relationship between data points and visual objects that are to be described.
- *Selection :* of attributes from the data which aims for right pictorial representation.
- *Presentation :* is effective management and organization of information in the available screen space.
- *Interactivity :* is providing facilities to organize explore and rearrange the visualization.
- *Human factors :* are easy readability and accessibility of information for end user.
- *Evaluation:* is finding out effectiveness in the created visual, if we have succeeded in reaching our goal in creating lucid and easy to understand graphics.

A Brief Survey of Graphical Techniques

Charting or *Graphing* is small subset of visualization where the data in question is explained with the help of bar charts, line charts or pie charts.

Information graphics or *infographics* are graphic visual representations of information, data or knowledge intended to present complex information quickly and clearly. They can improve cognition by utilizing graphics to enhance the human visual system's ability to see patterns and trends. *Infographics* are used to communicate a message, to simplify presentation of large amount of data, often different facets arranged in a thematic way.

A *Scorecard* is a tabular visualization of measures and their respective targets with visual indicators to see how each measure is performing against their targets at a glance. Scorecards may contain columns that show trends in spark lines. It measures performance against goals. It displays the graphic indicators that visually convey the overall success or failure of an organization in its efforts to achieve a particular goal.

A *report* is the presentation of data transformed into formatted and organized according to specific business requirement. Reports contain detailed data in a tabular format and typically display numbers and text only, but they can use visualization to highlight key data.

Dashboarding takes visualization a step further by aggregating several different pieces of visual information in a single location. As quoted by Stephen Few, "dashboarding is a visual display of the most information needed to achieve one or more objectives which fits entirely on a single computer screen so that it can be monitored at a glance." A typical dashboard might contain a scorecard, an analytical report and an analytical chart. Digital dashboards are laid out to track the flow inherent in the business process that they monitor. Dashboard is a user interface that is used to organize and present information in a way that is easy to read. A good dashboard presents information about important data, with fewer graphs and time overview.

To generate visualization, data plays a major role. In order to visualize, data need to be in one of the format.

Following are various data formats that are available.

- Spreadsheets are electronic document in which the attributes are stored in columns and the objects are stored in rows.
- JSON (Java Script Object Notation) is a simple human readable file format with data objects consisting of key value pairs.
- XML (extensible Markup language) is a flexible way to create common information formats and share both the format and the data on the world wide web
- Delimited Separated Values is format of data that is used to store various two dimensional arrays of data by separating the values in each row with specific delimiter characters. Various types of delimiters are comma, space, tab, and semi-colon.
- RDF (Resource Description Framework) is a general framework for describing website metadata, or information about information.
- HTML (Hyper Text Mark-up Language) is a language used for describing web-pages using ordinary text.

In many cases, there will be a need to use more than one data format (variety

of data). In such situations, we need to integrate different varieties of data. To create visualizations for such complex data, there are several popular approaches and methods which are discussed below.

Popular Approaches/Methods for Data Visualizations

There have been some conventional ways to visualize data in the form of tables, histograms, pie charts and bar graphs. However to convey a message effectively there are some exciting visual techniques that are available.

In addition to the above methods, there are certain popular open source tools.

Technique	Description
Choropleth	Thematic map where each spatial unit is filled with pattern or color which are scaled and normalized. Choropleth used to show spatial variation of one or two variables at time by using color, shades and/or patterns. E.g. population density of each state in a country.
Contour Heat Map	used to display density from the vector point of data. Contour heat maps are used to plot when there are large number of clustered and continuous data points which can also take categorical variables. These maps do not actually plot the data but designs a surface fit to the data. E.g. density of population.
Chord Diagram	used to display inter-relationship between group of entities. The data is arranged around the circle and the relationships are displayed by arcs.
Collapsible Tree	Used for hierarchical data of long nested lists on the web pages can be difficult to understand. The tree view, a user interface widget that displays hierarchical data as nested lists, solves this problem by making lists collapsible and expandable; a list can be opened by or closed by clicking on its parent list item. where the parent node and children nodes are are joined by arcs or line. E.g. government divisions and sub-divisions.
Tree Map	Used for hierarchical data that shows attributes of leaf nodes using size and color coding. Tree maps enable users to compare nodes and sub-trees even at varying depth in the tree and help them spot the pattern. E.g. visualization of continents, countries, population and area
Calendar Heat Map	Chart time series onto vector of dates. E.g. Twitter activity of a celebrity on each day.
Method	**Description**
Word cloud	is a method of visualizing unstructured text data. Word cloud expresses the occurrence of words in the text form with size of word or phrase directly related to the frequency of occurrence with which the word has occurred in the text document. Word cloud helps quickly analyze the main focus or topics of discussion and can also help us in sentiment analysis.
Association Trees	are used for understanding word association in large quantities of text. Association is most commonly used in social media text, News analysis or customer feedback. Latent Semantic Analysis (LSA) is a statistical computation used to identify relationships between a set of documents and the terms contain by producing a set of concepts related to the documents and terms. The main idea behind LSA is that all the word combinations in which a given word and word does not appear can determine the similarity of word meanings. Association trees are a way defining such similarities.
Cubism Horizon graphs	is used to analyze time series, or streaming content. Cubism Horizon graphs is an intuitive way of project real-time time series plot. Horizon charts reduce vertical space without losing resolution.
Self-Organizing maps or Topological Analysis	is used for relationships and gaining insights from multidimensional data from multiple sources. A topological network represents the data by grouping similar data points into nodes and connecting those nodes by an edge if the corresponding collections have data points in common. The visualization techniques comes under the heading of scatter plot methods, where the data points are projected on to 2D or 3D dimensional space, then plotting projections on the coordinates in usual way.
Network Graphs	is used to study meaning and relationships between large contextual data. These graphs are used to quantify relationships between different vertices of data. These graphs can be directional or non-directional based on requirement. A network is a collection of points; called vertices with lines connection these points are called arcs.

Open Source Visualization Tools

R is a programming language used for Statistical Analysis, Data Visualization and Predictive Modelling. R is an implementation of S language combined with lexical scoping semantics. R is a scripting and an interpreted language i.e. a programming language for which most of the implementations execute instructions directly without compiling the program into machine language. R is an open source software with great contribution from R-community towards R-programming in the form of packages which are available on Comprehensive R Archive Network (CRAN).

The *base graphs* in R are used most commonly and are a very powerful system for creating 2-D graphics. The main function for base graphic is plot(). The base graphs are loaded by default into R.

Grid is an alternative graphics system added to R that allows for the creation of multiple regions on a single graphics page. The grid package needs to be loaded before it can be used by using library function.

lattice graphics is a powerful Implementation and elegant high-level data visualization system with an emphasis on multivariate data. The lattice package is an implementation of Trellis graphics for R originally developed for the S-Language. The lattice consists of high-level generic functions each designed to create a particular type of display by default. Lattice gives advantage of high user controllable settings.

ggplot2 is a plotting system which takes the best from the base and lattice graphics. The plot can be split into scales and layers which gives the added advantage over base plot.

rShiny is an interactive web application framework for r which helps us do our analysis in dynamic fashion. Shiny combines the computational power of R with interactivity of modern web. rShiny has its own capabilities which doesn't require HTML, CSS or JavaScript Knowledge.

D3.js is a Java Script Library which helps build data visualization framework. D3 stands for Data Driven Documentations. D3.js is a powerful tool for creating dynamic and interactive data visualizations. D3.js uses Scalable Vector Graphics, JavaScript, HTML5, and Cascading Style Sheets (CSS3) standards.

google charts is a simple and powerful open source which can be used to visualize simple line charts to complex hierarchical tree maps. Google charts are a specialist for geocharts. Google charts can easily connect charts and controls into interactive dashboards. It can also be used to connect to data in real time using variety of data connection tools and protocols.

Gephi is an interactive visualization and exploration platform for all kinds of network and complex systems, dynamic and hierarchical graphs. Gephi is used for exploratory data analysis, link analysis, social network analysis, and biological network analysis and poster creation.

Lumify is a open source big data analysis and visualization platform. Its intuitive web-based interface helps users discover connections and explore relationships in their data via a suite of analytic options, including 2D and 3D graph visualizations, full-text faceted search, dynamic histograms, interactive geographic maps, and collaborative workspaces shared in real-time.

Conclusion

Visualization provides great value addition for the data that is to be analyzed. There are several techniques, methods and open source tools for visualization. Once the data is ready in a particular format, visualization can be generated by using these techniques.

References

[1] M Khan and S S Khan, (2011). "Data and Information Visualization Methods, and Interactive Mechanisms: A Survey", International Journal of Computer Application (0975-8887), vol. 34 – No.1, November 2011

[2] https://www.dashingd3js.com/why-build-with-d3js

[3] S Card, J MacKinlay, and B Shneiderman, (1998). "Readings in Information Visualization: Using Vision to Think". Morgan Kaufmann.

[4] Alfredo R Teyseyre and Marcelo R Campo, (2009). "An Overview of 3D Software Visualization", IEEE Transactions on Visualization and Computer Graphics, vol.15, No.1.

[5] L Chittaro, (2006). "Visualizing Information on Mobile Devices", ACM Computer, v.39 n.3, p.40-45.

[6] Edward R Tufte, (2007). "The Visual Display of Quantitative Inforamtion", Second Edition, Graphics Press.

[7] http://whatis.techtarget.com/definition/infographics

[8] http://thenextweb.com/dd/2013/10/16/10-ways-use-infographics/

[9] http://blogs.wsj.com/cmo/2014/06/24/outside-voices-the-visual-web-is-changing-everything-in-media-and-advertising/

[10] http://spotfire.tibco.com/blog/?cat=34

[11] http://www.computer.org/portal/web/computingnow/archive/january2014

[12] http://smartdatacollective.com/jgptec/140486/3-big-trends-data-visualization

[13] http://code.stephenmorley.org/javascript/collapsible-lists/

About the Authors

Dr. KVSN Rama Rao is a Professor,Dept of CSE at MLR Institute of Technology,Hyderabad.He is a Doctrate in Computer Science with over two decades of academic experience. He has published several papers in reputed International and National journals and conferences. His research interests are Cyber Security and Big data.

Surya Putchala is CEO and Chairman for ZettaMine Technologies, a firm focuses on providing high end educational and Management consulting services to Business around the world. It also aims to become the first think tank in "Big Data" space in India. Over the last 2 decades, he provided thought leading consulting solutions in the areas of Business Intelligence, Data Warehousing, Data Management and Analytics to Fortune 500 Clients. He has architected commercial Analytical Applications such as Product MDM and Procurement Optimization. He has held senior leadership roles with firms such as GE capital, Cognizant, Accenture and HCL.

Midhun Thaduru is an Associate Consultant at ZettaMine Technologies. He is with the Data Science team and focusses on Statistical Analysis, Exploratory Data Analysis, machine learning and Visualization. He extensively uses R, python, D3.js for his day to day programming. His area of focus is in Life Sciences (Pharma) and Insurance, particularly, Payer and Provider analytics. His interests are in developing high performance algorithms for Predictive modelling. Midhun graduated from BITS, Pilani.