Asian Journal of Computer Science and Technology
ISSN 2249-0701 Vol. 4 No. 1, 2015, pp.29-33

A Survey on Data Visualization Applications Challenges Techniques and Technologies

J.Jabanjalin Hilda and C.Srimathi
School of Computing Science and Engineering, Vellore Institute of Technology, Vellore, India
Email: jabanjalin.hilda@vit.ac.in

Abstract -**Trillions of dollars are spent each year on health care. The Department of Health Research (DHR) keeps track of a variety of health care indicators across the country, resulting in a large geospatially multivariate data set. This paper presents the various techniques, tools, technology and algorithms for the representation of large scale data which aims to provide good overviews of complete structures and the content of the data in one display space. The ability to visualize multiple variables on the map and compare them using a table and charts at the same time can provide valuable insights which might not be possible to obtain from current tools. . There are large numbers of data visualization techniques which have been developed over the last decade to support the exploration of large data sets. The techniques and tools discussed in this paper are based on the data type to be visualized, the visualization technique and the interaction and distortion technique.**

***Keywords:*Geovisualization, Rectangle packing, Community Health map, ArcGIS**

I.INTRODUCTION

The progress made in hardware technology allows today's computer systems to store very large amounts of data. Researchers from different Universities estimate that Internet Traffic will reach 1.3 Zettabytes by 2016.Yotabytes of data are generated, of which a large portion is available in digital form. Geographical analysis of such large amount of data is often a difficult task and searching for patterns is particularly a difficult challenge. Geovisualization research is increasingly dealing with the exploration of patterns and dealing with the exploration of patterns and relationships in such large data sets for understanding underlying geographical processes.One of the attempts has been to use Artificial Neural Networks as a technology especially useful in situations where the numbers are vast and the relationships are often unclear or even hidden.The alternate and different views on the data can help stimulate the visual thinking process that is characteristic of visual exploration. Four goals of the exploration are emphasized:

a. Patterns discovery through similarity representations
b. Correlations and relationships exploration of the hypothesis generation
c. Exploration of the distribution of the dataset on the map
d. And the detection of the irregularities in the data.

Complex correlations in this kind of statistical data can be portrayed using the Self-Organizing Map [1]. The proposed approach offers a number of visualizations to show the clustering structure and similarity. These techniques use a distance matrix to show distances between neighboring SOM network units.The most widely distance matrix technique used is the U-matrix. It contains the distances for each unit center to all of its neighbors. Unlike chorpleth maps the position of the map units is determined during the training of the network, according to the characteristics of the data samples.

II.BACKGROUND

The visual analysis of human health data describes the number of cases of different diagnoses in a spatial and temporal frame of reference.Visualization is a promising tool to analyze larger volumes of data. If visualization is done properly, relevant information can be perceived intuitively and the underlying data can be understood more easily.By proper visualization it is meant that a visual representation has to be effective,expressive and appropriate. Effectiveness depends on the degree to which visualization supports easy and intuitive interpretation of the visualized facts.Expressiveness relates to the requirement that all relevant information must be expressed in visualization.

Theoretically, the visualization process is implemented in for main steps- data analysis, filtering, mapping and rendering(see Figure 1). They make up the visualization pipeline.

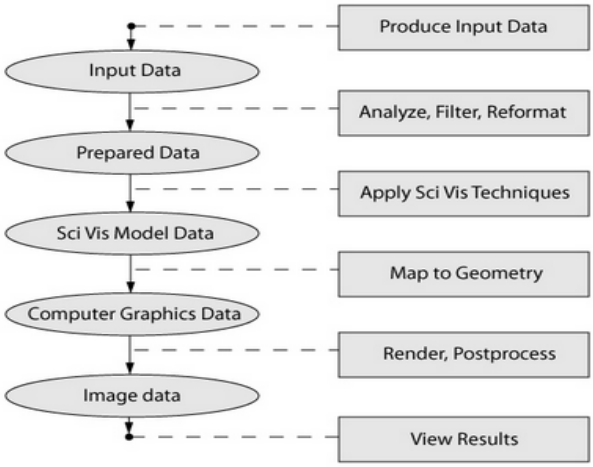

Fig.1 The visualization pipeline

In the data analysis step, data are prepared for visualization, by applying a smoothing filter, interpolating missing values, or correcting erroneous measurements. The filtering step selects the data portions to be visualized .In the mapping step, the focused data are mapped to geometric primitives(e.g. points, lines) and their attributes(e.g., color ,position ,size). The mapping step is most critical one for achieving expressiveness and effectiveness, and hence it is the most interesting one to visualization designers. Finally, geometric data are transformed to visual representations (e.g., images or animations).

III.VISUALIZATION TECHNIQUES

Data visualization needs extraordinary techniques to efficiently process large volume of data within limited run times. Data visualization techniques involve a number of disciplines, including statistics, data mining, machine learning, neural networks, social network analysis, signal processing, pattern recognition, optimization methods and visualization approaches. There are many specific techniques in these disciplines, and they overlap with each other hourly.

Optimization Methods have been applied to solve quantitative problems in a lot of fields, such as physics, biology, engineering, and economics. Several computational strategies for addressing global optimization problems are discussed, such as simulated annealing, adaptive simulated annealing, quantum annealing, as well as genetic algorithm which naturally lend itself to parallelism and therefore can be highly efficient. Stochastic optimization, including genetic programming, evolutionary programming, and particle swarm optimization are useful and specific optimization techniques inspired by the process of nature. However, they often have high complexity in memory and time consumption. Real-time optimization is also required in many Big Data application, such as WSNs and ITSs. Data reduction and parallelization are also alternative approaches in optimization problems.

Statistics is the sciences to collect, organize, and interpret data. Statistical techniques are used to exploit correlationships and causal relationships between different objectives. Numerical descriptions are also provided by statistics. However, standard statistical techniques are usually not well suited to manage Big Data, and many researchers have proposed extensions of classical techniques or completely new methods. Statistical computing and statistical learning are the two hot research sub-fields.

Data mining is a set of techniques to extract valuable information (patterns) from data, including clustering analysis, classification, regression and association rule learning. It involves the methods from machine learning and statistics. Big Data mining is more challenging compared with traditional data mining algorithms. Taking clustering as an example, a natural way of clustering Big Data is to extend existing methods (such as hierarchical clustering, K-Mean, and Fuzzy CMean) so that they can cope with the huge workloads .The kind of clustering algorithms include CLARA (Clustering LARge Applications) algorithm, CLARANS (Clustering Large Applications based upon RANdomized Search), BIRCH (Balanced Iterative Reducing using Cluster Hierarchies) algorithm, and so on.

Genetic algorithms are also applied to clustering as optimization criterion to reflect the goodness. Clustering Big Data is also developing to distributed and parallel implementation [4]. Taking discriminant analysis as another example, researchers try to develop effective algorithm for large-scale discriminant analysis. The emphasis is on the reduction of computational complexity. Taking bioinformatics as another example, it becomes increasingly data-driven that leads to paradigm change from traditional

single-gene biology to the approaches that combine integrative database analysis and data mining [23]. This new paradigm enables the synthesis of large-scale portraits of genome function.

Machine learning is an important subjection of artificial intelligence which is aimed to design algorithms that allow computers to evolve behaviors based on empirical data. The most obvious characteristic of machine learning is to discovery knowledge and make intelligent decisions automatically. When Big Data is concerned, we need to scale up machine learning algorithms, both supervised learning and unsupervised learning, to cope with it. Deep machine learning has become a new research frontier in artificial intelligence . In addition, there are several frameworks, like Map/Reduce, DryadLINQ, and IBM parallel machine learning toolbox, that have capabilities to scale up machine learning. For example, Support Vector Machine (SVM), which is a very fundamental algorithm used in classification and regression problems, suffers from serious scalability problem in both memory use and computation time. Parallel SVM (PSVM) are introduced recently to reduce memory and time consumption. There are many scale machine learning algorithms , but many important specific sub-fields in large-scale machine learning, such as large-scale recommender systems, natural language processing, association rule learning, ensemble learning, still face the scalability problems.

Artificial neural network (ANN) is a mature techniques and has a wide range of application coverage. Its successful applications can be found in pattern recognition, image analysis, adaptive control, and other areas. Neural processing of large-scale data sets often leads to very large networks[2]. Then, there are two main challenges in this situation.

Visualization Approaches are the techniques used to create tables, images, diagrams and other intuitive display ways to understand data. Big Data visualization is not that easy like traditional relative small data sets because of the complexity in 3Vs or 4Vs. The extension of traditional visualization approaches are already emerged but far away from enough. When it comes to large-scale data visualization, many researchers use feature extraction and a geometric modeling to significantly reduce the data's size before the actual data rendering [2]. For more closely and intuitively data interpretation, some researchers try to run batch-mode software rendering of the data at the highest possible resolution in a parallel way .

Social Network Analysis (SNA) which has emerged as a key technique in modern sociology, views social relationships in terms of network theory, and it consists of nodes and ties. It has also gained a significant following in anthropology, biology, communication studies, economics, geography, history, information science, organizational studies, social psychology, development studies, and sociolinguistics and is now commonly available as a consumer tool. SNA include social system design , human behavior modeling , social network visualization , social networks evolution analysis , and graph query and mining . Recently, online social networks and Social media analysis have become popular . One of the main obstacles regarding SNA is the vastness of Big Data. Analysis of a network consisting of millions or billions of connected objects is usually computationally costly. Two hot research frontiers, social computing and cloud computing, are in favor of SNA to some degree. Higher level Big Data technologies include distributed file systems , distributed computational systems, massively parallel-processing (MPP) systems , data mining based on grid computing , cloud-based storage and computing resources, as well as granular computing and biological computing. Linear mapping methods, such as principal component analysis (PCA) and factor analysis, are popular linear dimension reduction techniques. Non-linear techniques include kernel PCA, manifold learning techniques such as Isomap, locally linear embedding (LLE), Hessian LLE, Laplacian eigenmaps, and LTSA . Recently, a generative deep networks, called auto encoder , perform very well as non-linear dimensionality reduction. Random projection in dimensionality reduction also have been well-developed .

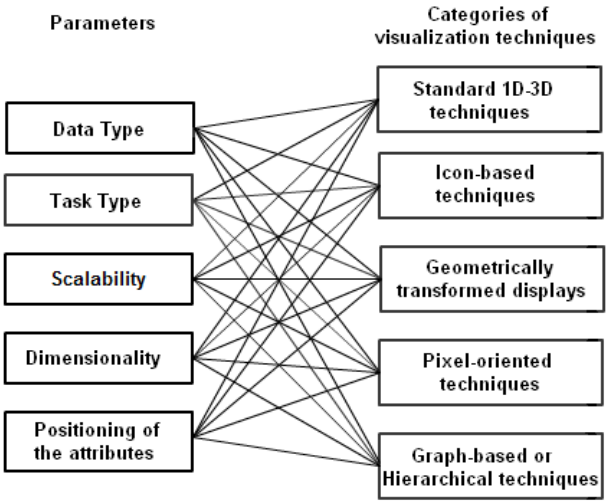

Fig.2 Visualization Techniques

The above figure shows the association among the identified parameters and the categories of visualization techniques. To visualize temporal aspects, Pencil icons are used for linear time axes and helix icons are applied for cycle time [3].

Pencil icons[3]: It is very useful visual metaphor. The natural shape of a pencil provides multiple faces that evolve from a common tip. This tip excellently serves as a 3D icon for visualizing multiple diagnoses with dependencies on linear time. To map data onto a pencil icon, linear time is encoded along the pencil faces starting from the common tip. Each face of a pencil is used to color-code one diagnosis. The number of faces of a pencil can be adjusted according to the number of diagnoses to be visualized.

Helix icons[3]: A spiral helix provides a geometric shape that allows an emphasis of the cyclic character of diagnoses. In order to construct a helix icon, a ribbon is created ;For each time step the ribbon extends in angle and height depending on the number of temporal primitives per cycle and the number of cyclic passes. Again color coding is used to encode data values along the ribbon.In order to represent multiple diagnoses, the ribbon can be divided into narrow sub-ribbons.

TreeMap is a wellknown space filling visualization technique that places all parts of the data onto one display space.

The described pencil and helix icons can be easily embedded into a 3D map display.Howeer, the embedding in 3D involves new problems compared to a 2D representation:

a.Undesired changes of the icon view upon user interaction, and

b.Loss of information due to icon occlusion and hidden surfaces.

IV. DATA VISUALIZATION TOOLS

The tools can be classified as follows:

a. Data Visualization Tools for Search Engine Optimization
b. Data Visualization Tools for Government
c. Data Visualization Tools for CRM data
d. Data Visualization Tools for Hurricane Tracking

Community Health Map[5] is a web based tool designed to visualize health care data that can be of county level or Hospital referral regions level granularity.

Dartmouth Atlas is a Web application which target health care specialists and researchers. It allows only one variable to be visualized at a time.

Health Statistics Map is a desktop application which target on Epidemiologists doing research on cancer.This map has sliders that controls the map animation and temporal changes in a geographical region can be visualized.

Hospital Compareis a web based tool which allows comparisons of upto three hospitals in a given country and three medical conditions, using a drop down menu.But here comparisons are not allowed for different geographical regions.

HealthVisPCPis a desktop application which as well as scatter plots for visualizing multivariate data[5].

Pennsylvania Cancer Atlasis a web application which mapand scatter plots.Only one variable can be viewed at a time.

TABLE I DATA VISUALIZATION TOOLS

Microsoft adCenter	SEObook rank checker	Stack
Google Adwords	Tableau	Visual Thesaurus
Dryad	Fidg't	Taglines
SEO Chat	Flickrvision	Quintura
Wordze	Digg Labs	KartOO
SEO for Firefox	BigSpy	walk2web
Digg RADAR	Swarm	CrazyEgg's

V. ALGORITHMS

Various algorithms are discussed for placing hierchical data onto display spaces.

Mesh Edge Based Rectangle packing algorithm: This quickly packs a set of arbitrarily sized rectangles in a small display space [1]. It interactively displays large scale hierarchical data, where icons,thumbnails, or borders are treated as arbitrarily sized rectangles. It generates a Delaunay triangular mesh that connects the centers of the placed rectangles and refers to it to quickly find the gaps.

Template based rectangle packing: The templates describes he ideal positions of the nodes of input data .it places the rectangle as close as possible to the ideal positions described in the templates , while it still reduces the usage of display space and avoids overlaps among the data items.[1]

On combination of these two algorithms, the following ultimate features can be achieved.

a) Effective use of display spaces
b) No overlaps between nodes
c) Aspect ratio of subspaces
d) Flexible placement of arbitrarily shaped nodes
e) Similarity
f) Semantics of placement

VI.CONCLUSION

In this paper a brief overview on Data visualization problems, including its opportunities and challenges, current techniques and technologies are discussed. Also, several potential algorithms to solve the problem have been discussed. Although these technologies are still under development, we have confidence that in the coming future we will receive great breakthroughs in those areas. Undoubtedly, today and future's Data visualization problems benefit from those progresses. These techniques can be applied for the visualization of Web access and real time monitoring of distributed processes. Statistical and data mining methods like Principle component Analysis or clustering are helpful analytical tools to support the identification of the important in human health data. Elaborate description has been given on choosing visualizing techniques properly with respect to the characteristics of the data.

REFERENCES

[1] Itoh, Takayuki, et al. "Hierarchical data visualization using a fast rectangle-packing algorithm." *Visualization and Computer Graphics, IEEE Transactions on* 10.3 (2004): 302-313.

[2] Chen, CL Philip, and Chun-Yang Zhang. "Data-intensive applications, challenges, techniques and technologies: A survey on Big Data." *Information Sciences* 275 (2014): 314-347.

[3] Tominski, Christian, Petra Schulze-Wollgast, and Heidrun Schumann. "Visual methods for analyzing human health data." *Encyclopedia of healthcare information systems* 1 (2008): 1357-1364.

[4] Hanson III, C. William, and Bryan E. Marshall. "Artificial intelligence applications in the intensive care unit." *Critical care medicine* 29.2 (2001): 427-435.

[5] Sopan, Awalin, et al. "Community Health Map: A geospatial and multivariate data visualization tool for public health datasets." *Government Information Quarterly* 29.2 (2012): 223-234.

[6] Brewer, Cynthia A. "Color research applications in mapping and visualization."*Color and Imaging Conference*. Vol. 2004. No. 1. Society for Imaging Science and Technology, 2004.

[7] MacEachren, Alan M., Cynthia A. Brewer, and Linda W. Pickle. "Visualizing georeferenced data: representing reliability of health statistics." *Environment and Planning A* 30.9 (1998): 1547-1561.

[8] Ren, Fang, and Mei-Po Kwan. "Geovisualization of human hybrid activity-travel patterns." *Transactions in GIS* 11.5 (2007): 721-744.

[9] Yamaguchi, Yumi, and Takayuki Itoh. "Visualization of distributed processes using" data jewelry box" algorithm." *Computer Graphics International, 2003. Proceedings*. IEEE, 2003.

[10] Henry, Nathalie, J. Fekete, and Michael J. McGuffin. "NodeTrix: a hybrid visualization of social networks." *Visualization and Computer Graphics, IEEE Transactions on* 13.6 (2007): 1302-1309.

[11] Liang, Zhong, ChiTian He, and Zhang Xin. "Feature based visualization algorithm for large-scale flow data." *Computer Modeling and Simulation, 2010. ICCMS'10. Second International Conference on*. Vol. 1. IEEE, 2010.

[12] Keim, Daniel A., et al. "Visual data mining in large geospatial point sets."*Computer Graphics and Applications, IEEE* 24.5 (2004): 36-44.

[13] Ahrens, James, et al. "Large-scale data visualization using parallel data streaming." *Computer Graphics and Applications, IEEE* 21.4 (2001): 34-41.

[14] Müller, Wolfgang, Thomas Nocke, and Heidrun Schumann. "Enhancing the visualization process with principal component analysis to support the exploration of trends." *Proceedings of the 2006 Asia-Pacific Symposium on Information Visualisation-Volume 60*. Australian Computer Society, Inc., 2006.

[15] Fuchs, Georg, and Heidrun Schumann. "Intelligent icon positioning for interactive map-based information systems." *Proceedings of the International Conference of the Information Resources Management Association (IRMA)*. 2004.