Indian Journal of Science and Technology, *Vol 9(27), DOI: 10.17485/ijst/2016/v9i27/88692, July 2016*
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Review on the Development of Big Data Analytics and Effective Data Visualization Techniques in the Context of Massive and Multidimensional Data

J. Jabanjalin Hilda*, C. Srimathi and Bhulakshmi Bonthu

School of Computing and Engineering, Vellore Institute of Technology, Katpadi - 632014, Vellore, Tamil Nadu, India;
jabanjalin.hilda@vit.ac.in, csrimathi@vit.ac.in, bhulakshmi.b@vit.ac.in

Abstract

Objectives: Data visualization, the use of images to represent information, is now becoming properly appreciated due to the benefits it can bring to business. This paper focuses on the general background of data visualization and visualization techniques. **Methods**: Data visualization has the prospective to assist humans in analysing and comprehending large volumes of data, and to detect patterns, clusters and outliers that are not obvious using non-graphical forms of presentation. For this reason, data visualizations have an important role to play in a diverse range of applied problems, including data exploration and mining, Information retrieval and intelligence analysis. In real time various techniques have been used of which Geometric projection techniques, Iconographic display techniques, Pixel-oriented, Hierarchical techniques, Graph-based techniques are discussed. **Findings**: The major difficultly in big data visualization is to preserve any of the original dimensional information. The taxonomy detailed here show that the local and global structure of the data can be visualized in an interactive manner and has a massive advantage.

Keywords: Parallel Coordinates, Star Coordinates, Visual Analytics, Visualization Techniques

1. Introduction

Big data is categorised by 5 V's, which are volume, variety (different forms of data sources) velocity (speed of change), veracity (uncertainty of data and incompleteness) and value. Reports say data from the U.S. healthcare system alone reached, in 2011, 150 exabytes. At this rate of growth, big data for U.S. healthcare will soon reach the zettabyte (10^{21} gigabytes) scale and, not long after, the yottabyte (10^{24} gigabytes) in the near future. The real issue is not that you are acquiring large amounts of data. It's what you do with the data that counts. The hopeful vision is that organizations will be able to take data from any source, harness relevant data and analyse it to find answers that enable the following 1. Reduction in the cost, 2. Reduction in the time, 3. Development of new product and optimized offerings, and 4. decision making in smarter business. For instance, by combining big data and high-powered analytics, it is possible to:

- Determine root causes of failures, issues and defects in near-real time, potentially saving billions of dollars annually.
- Optimize routes for many thousands of package delivery vehicles while they are on the road.
- Generate retail coupons at the point of sale based on the customer's current and past purchases.
- Send tailored recommendations to mobile devices while customers are in the right area to take advantage of offers.
- Recalculate entire risk portfolios in minutes.
- Quickly identify customers who matter the most.
- Use clickstream analysis and data mining to detect fraudulent behaviour

2. Big Data: Challenges

The challenges that researchers face across the globe and as well as in India are related to data inundation

**Author for correspondence*

pertaining to Fundamental Science, Computational Biology, Bioinformatics, Astrophysics, Materials Science, Atmospheric and Earth observations, Energy and Medicine, Engineering and Technology, GIS and Remote Sensing, Cognitive Science and Statistical data. These challenges require development of innovative algorithms, visualization techniques, visualization tools, data streaming methodologies and analytics. The overall constraints that community facing are

- The IT Challenge: Computational , loading and storage power
- The computer science: Design of algorithms, visualization, scalability (Data mining, network and Graph analysis, streaming of data and text mining), distributed data, architectures, data dimension reduction and implementation.
- The mathematical science: Statistics, Optimisation, uncertainty quantification, model development analysis and systems theory.
- The multi-disciplinary approach: Appropriate problem solving.

3. Why Data Visualization?

Data visualization is the presentation of data in a pictorial or graphical format. For centuries, people have depended on visual representations such as charts and maps to understand information more easily and quickly.

As more and more data is collected and analysed, decision makers at all levels welcome data visualization software that enables them to see analytical results presented visually, find relevance among the millions of variables, communicate concepts and hypotheses to others, and even predict the future.

3.1 Visualization Pipeline

Figure 1 describes the step-wise process of creating visual representations of data.

- Data Analysis: data are pre-processed and prepared for visualization (e.g., by applying filter, interposing missing values, or correcting erroneous measurements) -- usually computer-centered, little or no user interaction.
- Filtering: Portions of data selected to be visualized -- usually user-centered.
- Mapping: emphasised data are mapped to geometric primitives (e.g., points, lines) and their attributes (e.g., color, position, size); this is the most critical step for achieving expressiveness and effectiveness.
- Rendering: Geometric data are converted into image data.

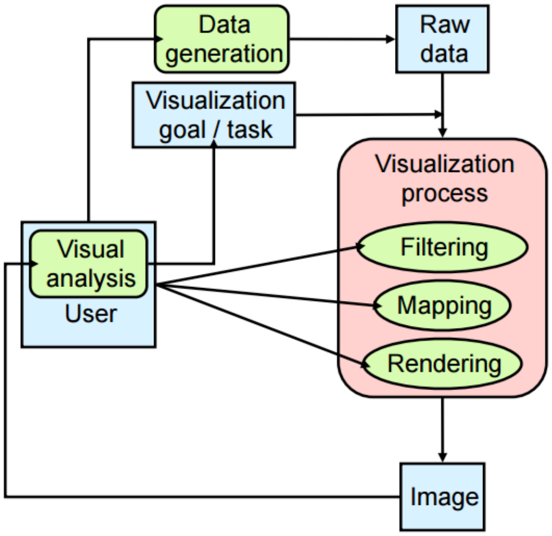

Figure 1. Visualization pipeline.

Visualization tools have become indispensable for exploring and identifying trends hidden within data. The data sources, such as satellite imagery for brushfires or real-time traffic information from Google, are critical to making good analysis and driving good decision-making.

3.2 What is Visual Analytics?

Visual Analytics is a subset of Data Visualisation which deals with actually making inferences using visual interfaces[1]. Visual Analytics is the discipline of analytical reasoning supported by interactive visual interfaces. Nowadays, data is produced at an unbelievable rate and the ability to collect and store the data is increasing at a faster rate than the ability to analyse it. Visual Analytics approaches allow decision makers to combine their human flexibility, imagination, and background knowledge with the enormous storage and processing capacities of today's computers to gain insight into complex problems.

Visual analytics is more than only visualization. It can rather be seen as an integral approach combining visualization, human factors and data analysis. Figure 2 illustrates the detailed scope of visual analytics. Concerning the field of visualization, visual analytics integrates methodology from information analytics, geospatial analytics, and scientific analytics. Especially human factors (e.g., inter-

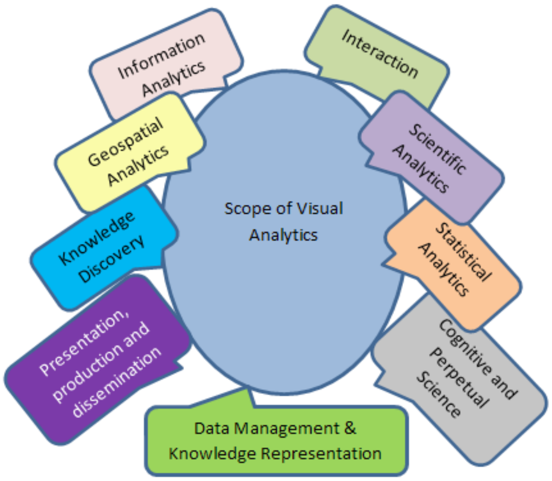
Scope of Visual Analytics
Information Analytics
Interaction
Geospatial Analytics
Scientific Analytics
Knowledge Discovery
Statistical Analytics
Presentation, production and dissemination
Cognitive and Perpetual Science
Data Management & Knowledge Representation

Figure 2. Scope of visual analytics.

action, cognition, perception, collaboration, presentation, and dissemination) play a key role in the communication between human and computer, as well as in the decision-making process.

3.3 The Visual Analytics Process

The Visual Analytics Process showed in Figure 3 combines visual analysis and automatic methods with a tight coupling through human interface in order to gain knowledge from data. The figure shows an intellectual overview of the different stages (represented through ovals) and their transitions (arrows) in the Visual Analytics Process.

The first step is often to pre-process and transform the data to derive different representations for further exploration. Other typical pre-processing tasks include data cleaning, standardization, grouping, or integration of heterogeneous data sources.

After step 1, the analyst may choose between applying visual or automatic analysis methods. If an automated analysis is used first, Machine learning techniques are applied to generate models of the original data. Once a prototype is created the analyst has to assess and improve the models, which can best be done by interacting with the data. Visualizations allow the analysts to interact with the automatic methods by modifying parameters or selecting other analysis algorithms. Model visualization can then be used to evaluate the findings of the generated models. Ambiguous results in an intermediate step can thus be discovered at an early stage, leading to better results. If a visual data exploration is performed first, the user has to confirm the generated hypotheses by an automated analysis. User interaction with the visualization is needed to reveal insightful information, for instance by shooting up on different data areas or by considering different visual opinions on the data. Findings in the visualizations can be used to direct model building in the automatic analysis. In summary, in the Visual Analytics Process knowledge can be gained from visualization, automatic analysis, as well as the preceding interactions between visualizations, models, and the human analysts.

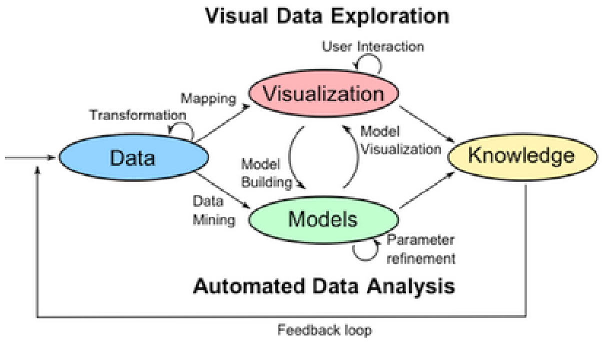

Figure 3. The visual analytics process.

4. Visualization Techniques

Many techniques used for data visualization. They are like geometric, parallel coordinate, stick figure, icon based, hierarchical, graph based (line graph) and pixel oriented visualization techniques. Some techniques are very specific to a certain application and some combine several ideas. In this section, three classifications of visualization techniques are discussed. We look at an overview on classifications by[2-4].

4.1 Classification of Visualization Techniques[2]

A multidimensional multivariate visualization technique has been ordered in the light of 2-variate displays, multivariate displays and animations[2]. 2-variate shows incorporates box plot, disseminate plot and so on. A standout amongst the most well-known multidimensional multivariate visualization techniques is the scatterplot matrix which exhibits all mix sets of all dimensions and compose them by a matrix[5]. They decide linear correlation between different variables. In a scatterplot matrix, each variety is dealt with indistinguishably. The possibility of pairwise adjacencies of variables is likewise a premise for the hyperbox[6], the hierarchical axis[7] and the hyperslide[8].

Multivariate presentations are the establishment for some as of late created multidimensional multivariate visualization techniques, the greater part of which utilize vivid design made by high-speed graphics computations. Xmdv Tool incorporates four of the current mdmv visualization tools: dimension stacking, scatterplot matrix, glyphs, and parallel coordinates into one framework with improved n-dimensional brushing. These procedures can comprehensively be classified into five sub-groups:

- Brushing permits direct control of a multidimensional multivariate visualization display. This method is depicted for scatterplot matrices[6].
- Panel matrix includes pairwise two-dimensional plots of neighbouring varieties. These strategies incorporate hyperslide[8] and hyperbox[6].
- Iconography utilizes varieties to decide estimations of parameters of small graphical objects. The mappings of data values to graphical parameters are normally produced texture patterns that ideally bring knowledge into the information. Some iconographic techniques are Chernoff face[9], stick figure icon[10], autoglyph[11] and colour icon[12].
- Hierarchical displays map a subset of variates into various hierarchical levels of the presentation. Hierarchical axis[13], dimensional stacking[14] and world within world[14] visualization techniques have a place with this gathering.
- dimension stacking
- Non-Cartesian displays map data into non-Cartesian axes. They incorporate parallel coordinates[14-16] and visdb[17] .

Animation is a capable strategy for envisioning multidimensional multivariate scientific data. Different movie animation techniques on multidimensional multivariate data and a scalar visualization animation models are introduced. The most prevalent animation technique is the grand tour technique, in which multidimensional multivariate data is projected into two dimensional planes.

4.2 Taxonomy of Visualization Techniques[3]

Four approaches have been introduced by to encode conceptual information, a typical event in information visualization: 1D, 2D, 3D alludes to orthogonal visualization that encodes information by positioning marks on orthogonal axes[3]. Multiple dimensions allude to the more difficult issue of multidimensional visualization where the data has such a large number of variables that an orthogonal visual structure is not adequate. Typical tasks that must be bolstered by such situations include getting knowledge from the data, such as discovering patterns, relationships, clusters, gaps and outliers, or finding specific items using interaction, such as zooming, filtering, and selection. Trees refer to utilizing association and enclosure to encode connections among cases. Networks refer to utilizing associations to encode connections among cases.

4.3 Taxonomy of Visualization Techniques[4]

Information visualization techniques have been classified by their essential visualization principle by[4]: geometric projection, iconographic, pixel-oriented, hierarchies, graph based and hybrid.

Geometric projection techniques support users in the assignment of discovering information projections of multidimensional multivariate data[18-20]. The article by[21,22] presents the results of using the parallel coordinate representation for high-dimensional data analysis. Along these lines, a high number of dimensions can be visualized. An integrated approach has been incorporated in[23,24] to preserve the original dimensional information. Typical examples here are star coordinates[25,26] and parallel coordinates and techniques are incorporated into the accompanying Table 1.

Iconographic display techniques map each multidimensional data item to an icon (or glyph) whose visual features differ contingent upon the data values[27,28]. The quantity of displayable dimensions is not restricted with this methodology.

Be that as it may, they are not utilized regularly for high-dimensional data sets, since a quick information

Table 1. Geometric projection techniques of visualization techniques

Category	Visualization technique	References
Geometric projection	Scatterplot matrices Hyperslice Parallel coordinates Andrews' plots Projection pursuit Prosection views Landscapes Radviz Star coordinates	[5] [8] [15, 16, 21] [18] [19, 20] [22] [23] [24] [25, 26]

exploration is risky. The iconographical techniques are given in the

Pixel-oriented - in pixel-based techniques, a pixel is utilized to speak to data values[29-31]. Pixels are grouped by dimension, the item it belongs to, and are organized on the screen suitable to various purposes. In general, one pixel is utilized per data value, so the quantity of displayable values is fairly high. The techniques are further sorted as "query independent" or "query dependent". In the query independent techniques, the arrangement of the pixels in the sub-windows is fixed, independently of the data values themselves. In the query dependent techniques, a query item is given and distances from the data values to the given query value are computed utilizing some metrics. The mapping of hues to pixels depends on the computed distances for every attribute and pixels in each sub-window are orchestrated according to their overall distances to the query data item. The Table 3 displays the pixel-oriented techniques.

Hierarchical techniques subdivide the m-dimensional data space and represent subspaces in a hierarchical manner. The hierarchical techniques termed as Dimensional stacking[32-34] and tree representation[35-37] are appeared in the Table 4.

Graph-based techniques envision large graphs using particular layout algorithms, query languages, and abstraction techniques to convey on their significance obviously and rapidly[38,39]. The graph-based techniques are given in the Table 5.

Table 2. Iconographic techniques of visualization techniques

Category	Visualization technique	References
Iconographic	Chenoff faces Stick figures Shape coding Color icons	[9,28] [10,27] [11] [12]

Table 3. Pixel-oriented techniques of visualization techniques

Category	Visualization technique	References
Pixel-oriented	Circle segment Spiral and axes techniques Recursive pattern	[29] [30] [31]

Table 4. Hierarchical techniques of visualization techniques

Category	Visualization technique	References
Hierarchies	Dimensional stacking Worlds within worlds (n-vision) Hierarchies Conetrees Treemap Infocube	[32] [33] [34] [35, 36] [37]

Table 5. Graph-based techniques of visualization techniques

Category	Visualization technique	References
Graph-based	Graph-based Hiernet Narcissus	[38] [39]

4.4 Star Coordinates

In star coordinates, each dimension is spoken to as a vector emanating from the centre point of a unit circle in a two-dimensional plane. At first, all axes have the same length and are consistently placed on the circle. Data points are scaled to the length of the axes, with the base being mapped to the beginning and the greatest to the next end of the axes on the unit circle.

In mathematics, the Cartesian coordinate framework is utilized to decide every point uniquely in a plane through two numbers, as a rule called the x-coordinate and the y-coordinate. A point P = (x, y) in the plane can be spoken to by a vector P = O+ xi + yj, where i = (1, 0), j = (0, 1) are the two basis vectors of the Cartesian coordinates and O = (0, 0) is the beginning. A multidimensional point is represented in a plane like the Cartesian coordinates. The 2D star coordinates system is used for representing a point in m dimensions incorporating m vectors in a plane V = {v1, . . . , vm}.Here vi = (vix, viy) = (cos2πi/m,sin2πi/m) is representing the ith dimension, i = 1, . . . ,m, and the inception is Om = (Ox,Oy). A mapping of a point (p1, . . . , pm) to a point P = (Px, Py) in two-dimensional Cartesian coordinates is controlled by the sum of basic vectors vi = (vix, viy) on every axis multiplied by the estimation of the point. Further precisely, the equation is given by:

$$P = O + \sum_{t=1}^{m} (p_t v_t)$$

Or

$$\begin{cases} Px = Ox + \sum_{t=1}^{m} (p_i v_i x) \\ Py = Oy + \sum_{t=1}^{m} (p_i v_i x) \end{cases}$$

In Figure 4, the star coordinates framework has eight axes D_1, . . . , D8 represent the eight dimensions. These axes represent for fundamental vectors of the Cartesian coordinates that uniformly put on a unit circle. The point P in the two-dimensional space is on representation of the point in eight dimensions (p1, . . . , p8). We begin at the origin O of a circle, moving along the axis D1 with length p1, keep moving parallel to the axis D2 with length p2 and so on. The end point of this procedure is the point P.

All coordinates systems are given by a starting point and a few vectors. Regularly, the vectors are linearly independent, e.g., Cartesian coordinates, and a point is exceptionally spoken to. In the star coordinates system, the vectors are linearly dependent, and the representation of a point is not special. As a rule, the mapping from multidimensional space into a low-dimensional space is not one of a kind. Just with a guide of interactive dynamic transformations such as rotations and translations one can understand the data representation. Star coordinates essentially endeavours to extend this thought to higher dimensions. Clusters, patterns, and outliers in a data set are protected in the projected multidimensional data visualization and interactions affirm this. Customary star coordinates initially included rotation and scaling and later stretched to incorporate range selection, marking, histograms, footprints, and sticks.

Table 6. Visualization algorithms

Algorithm	References
Self-organizing Map	[49]
Fast Map	[50]
Sammon's mapping algorithms	[51]

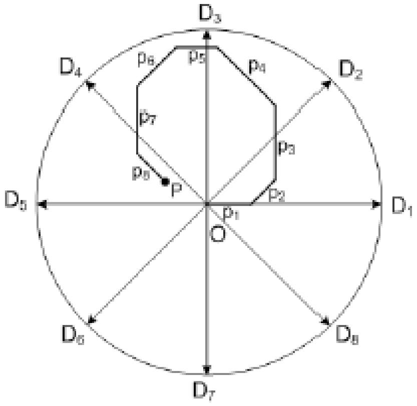

Figure 4. Calculation of data point location for an eight-dimensional data set.

Viz3D have been presented that tasks multidimensional data into a 3D display space[40]. Like star coordinates, the essential system of Viz3D is obtained from the fundamental system of star coordinates by adding 1 to the third coordinates that implies the fundamental system of Viz3D is given by:

$$vi = (\cos 2\pi i/m, \sin 2\pi i/m, 1), i = 1, \ldots m$$

what's more, the mapping from multidimensional data space into a 3D visual space is detailed as:

$$\begin{cases} Px = Ox + \frac{1}{m}\sum_{i=1}^{m} \frac{pi \cos 2\pi i}{m} \\ Py = Oy + \frac{1}{m}\sum_{i=1}^{m} \frac{pi \sin 2\pi i}{m} \\ Pz = Oz + \frac{1}{m} \cdot \sum_{i=1}^{m} pi \end{cases}$$

Axes arrangement is presented for show that keeps profoundly comparable attributes near one another, which might be accomplished by computing information on the attributes comparability from the data set[40].

Expansions of star coordinates have been proposed into three dimensions[41]. The creators add a third dimension to conventional star coordinates, which takes into consideration connection in the third dimension, yet it keeps up the two-dimensional display. Three-dimensional star coordinates expand the conventional two-dimensional star coordinates in a few ways:

Stars disseminate in a volume rather than a plane, giving clients more space to exploit.

- Depth cues permit users to incorporate more meaningful variables at the same time in an analysis.
- Transformations are stretched to three dimensions.
- System rotation is presented as a powerful new transformation.

An algorithm has been presented for a computerized way of finding the best configuration when high-dimensional data points are projected into a 3Dvisual space[42]. The best design of star coordinates is found among some arbitrary star coordinates configurations in view of self-organizing maps clustering algorithm in visual space to gauge nature of the star coordinates display.

Another algorithm has been proposed[43]for naturally finding the best configuration of star coordinates in light of the minimization of a multidimensional scaling object function (stress function).

VISTA mappings have been presented by[44]. The VISTA maps multidimensional data points into 2D visual space while giving the convenience of visual parameter change:

$$\begin{cases} Px = Ox + \frac{c}{m}\sum_{i=1}^{m} pi\ \alpha i\ cos\theta i \\ Py = Oy + \frac{c}{m}\sum_{i=1}^{m} pi\ \alpha i\ sin\theta i \end{cases}$$

where $\alpha = (\alpha_1, \ldots, \alpha_m)$ are the dimension adjustment parameters in [−1, 1], angles

$\theta = (\theta 1, \ldots, \theta m)$ are set to $\theta i = \frac{2\pi i}{m}$ at first and can be balanced, and c is the scaling of the radius of the display region. VISTA is an augmentation of conventional star coordinates that takes into for more intuitive exploration of multidimensional data.

Star class is presented[45] that permits intuitive star coordinates for visual classification.

Advanced star coordinates is presented[46] that utilize the diameter rather than the radius as the dimensions, axis, such that data points in multidimensional space are mapped into visual space saving attribute values with orthogonal distance from the visual point to the diameter. The diameters configuration procedure depends on correlations. The advanced star coordinates visualizes the clusters and structure of multidimensional data.

A technique has been proposed[47,48] for projecting multidimensional data in view of *class*-preserving projection. The authors introduced an algorithm for finding the best two-dimensional plane that preserves inter-class distances.

The mapping is a linear dimension reduction method, in which an advanced two dimensional

Sub-space is chosen keeping up the distance between means of classes.

4.5 Parallel Coordinates

Parallel coordinates is a standout amongst the most well-known visualization techniques for multidimensional multivariate data sets. Parallel coordinates are presented and are created for visualizing multidimensional geometry. Parallel coordinates depend on an arrangement of parallel coordinates, which incorporates Anon-projective mapping amongst multidimensional and two-dimensional sets.

Parallel coordinates on the plane with Cartesian coordinates, and beginning on the y-axis, m duplicates of the real line, marked X1, X2, …, Xm, are placed equidistant and perpendicular to the x-axis. Regularly, the Xi axe perpendicular to the x-axis lies at positions i−1, for i=1, …, m. They are the axes of the parallel coordinates system for the Euclidean m-dimensional space R^m all having the same positive orientation as the y-axis. A point P = (p1, …, pm) is spoken to by the polygonal line whose m vertices are at (i−1, pi) on the Xi axes for i= 1, …, m, see Figure 5. Points on the plane are represented by segments and the line contains the segment as depicted in Figure 6.

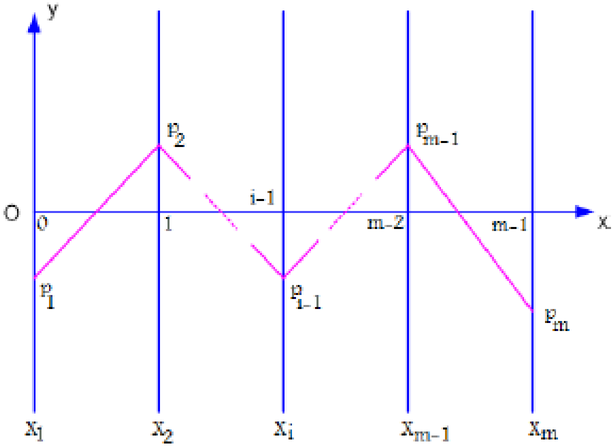

Figure 5. A polygonal line $\bar{P}\bar{P}$ represents a point P = (p1, . . . , pm).

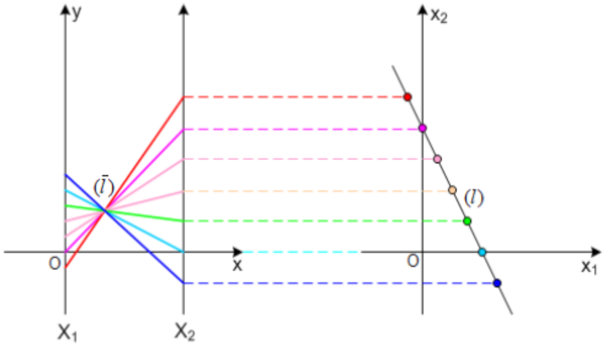

Figure 6. The dual line and point in parallel coordinates.

As a result, a one-to-one correspondence between points in R^m and planar polygonal lines with vertices on X1, …, Xmis set up.

The fundamental duality: we consider the X1X2 parallel coordinates and also the Ox1x2 Cartesian coordinates that are appeared in Figure 5. In the Cartesian coordinates Ox1x2, we draw a line (l) that is portrayed by the accompanying condition :(l) :x2 = mx1 + b. Every point (x1, x2 = mx1 + b) lying on the line (l) in the Cartesian coordinates is shown by a segment line with endpoints (0, x1) and (1, x2 = mx1 + b) in parallel coordinates. Consequently the points on (l) which are represented in parallel coordinates form an unbounded group of lines. On the off chance that m ≠ 1, the group of lines has a common point:

$$(\bar{l}):\quad \left(\frac{1}{1-m},\frac{b}{1-m}\right)$$

The point ($\bar{l}$) in parallel coordinates denotes the line (l) in Cartesian coordinates. For the circumstance m = 1, the group of lines has a typical point at boundlessness with direction (1, b). Each point in two-dimensional Cartesian coordinates is represented by a line in parallel coordinates and every point in parallel coordinates, which can be comprehended as a group of lines that cross at this point, represents a line in Cartesian coordinates. This property is known as a duality amongst line and point.

Multidimensional lines: A line (l) in R^mcan be depicted by m−1 linearly autonomous conditions of the form:(l) :xi+1 = mixi+ bi, i= 1, . . . , m−1.

The line (l) is represented in parallel coordinates by m – 1 indexed points in the XiXi+1 parallel coordinates. In Figure 7 the points $\bar{l}$ correspond to nearby variables.

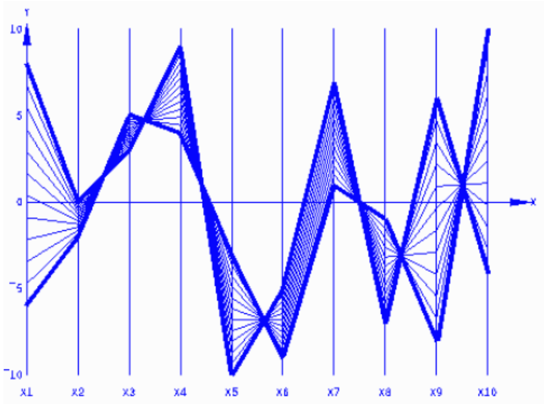

Figure 7. Parallel coordinates display an interval of a line in R10.

Parallel coordinates in data analysis: Wegman[21] presented a technique to analyse data utilizing parallel coordinates. In his paper, the author proposed two strategies called density plots and color histograms. For density plots, the algorithm depends on Scott's notion of the Average Shifted Histogram (ASH) to visualize density plots with parallel coordinates. The author utilized contours to represent the two dimensional density. Parallel coordinates density plots have the benefit of being graphical representations of data that are simultaneously high-dimensional and extensive. In colour histograms, the idea is to code the magnitude of an observation along a given axis by a colour canister. The diagram is drawn by picking an axis, and sorting the perceptions in ascending order. The author likewise presented a permutation algorithm of the axes for pairwise comparisons.

Multi-resolution view with parallel coordinates: Multi-resolution perspectives of the data have been built up by[49] through various hierarchical clustering and utilize a variation on parallel coordinates to convey aggregation information for the outcomes.

Focus + context visualization have been presented by[50] in parallel coordinates. Every pair of adjacent axes representing a pair of dimensions, in a two dimensional subspace is divided into b×b canisters, which make a frequency-based and output-oriented representation of the original data.

Frequency and density-based visualizations have been created[40].The fundamental thought of the algorithm is to make two-dimensional frequency histograms for each pair of adjacent attributes in parallel coordinates. A two-dimensional region between a couple of adjacent axes in parallel coordinates is divided into w × h bins, where w is the number of horizontal canisters and h is the number of vertical canisters. The estimation of frequency is stored in matrix F = (Fij) w×h. For every data point in multi-dimensional data sets, a line segment is drawn with the Bresenham algorithm, if the line segment goes through the (i, j)th container, they add 1 to the value of Fij. For the density plot, matrix frequencies F = (Fij) w×hare linearing scaled into [0, 255].

For the frequency plot, they utilized a 3×3 averaging filter applied to the FW×H matrix.

High-precision texture has been presented by[51,52] that can be utilized to uncover diverse sorts of cluster information. This visualization technique can be used to examine exclusive, overlapping, and hierarchical clusters. For displaying clusters in parallel coordinates, the authors utilized a transfer function on the intensity value which permits non-linear and additionally user-defined mappings.

Generalization of parallel coordinates: Perhaps the earliest multidimensional data visualization was presented, in which each multidimensional data point x = (x1, …, xm) is spoken to by a function of the form $f_{x(t)} = \frac{x1}{\sqrt{2}} + x2 \sin(t) + x3 \cos(t) + \ldots +$

What's more, this function is plotted on the extent [−π, π]. Some helpful properties of the Andrews' plots are safeguarding of means and distances. A free-form curve[53] has been represented such that the space between two neighbouring axes can be proficiently exploited to encode more information of the axes, which can assist to notice correlations among more than two dimensions. Smooth curves[54] have been used to permit users to distinguish an individual path through the curves' nodes.

A smooth plot between two adjacent axes has been used[55]. While in conventional parallel coordinates, a line segment can be comprehended as a linear interpolation, the authors presented a new group of smooth functions using smooth interpolation. Curved lines[56] are used to form visual bundles for clusters in parallel coordinates. The visual clustering is enhanced by changing the shape of the edges while keeping their relative order.

Dimension ordering in parallel coordinates: Dimension ordering, spacing, and filtering can enhance the parallel coordinate's layout and ease data exploration. Data dimensions have been clustered according as per their similarity, then data dimensions are reworked such that dimensions showing a similar behaviour are situated beside each other. A hierarchical approach has been proposed[57] to deal with enhancing the intuitiveness of dimension reordering, spacing, and filtering. A visual clutter measure[58] has been characterized as the proportion of outlier points to the aggregate data points. The optimized dimension order is then processed to minimize the proposed clutter measure.

Interacting with parallel coordinates: Angular brushing[59] has been used to choose data subsets with particular patterns between nearby axes. Parallel coordinates[60] has been straight forwardly manipulated by progressively outlining an arrangement of polylines and interactively visualizing correlation between polyline subsets. These brushing and interactive strategies are viewed as very effective tools in investigating the structures within the clusters.

Integration with parallel coordinates**:** Self-organizing map[61] has been used in conjunction with parallel coordinates, in which clusters are represented rather than data points, which helps to see a review and details in parallel coordinates[62]. Proposed the tight coupling amongst radviz and parallel coordinates called spring view. In spring view, the user can choose a 2D rangeon the radviz representation getting the corresponding components highlighted in the parallel coordinates cluttering. The colour coding on the radviz (based on a 2Dcolor-map guide to a rectangular board) is automatically computed, which takes into account automatically clustering the parallel coordinate's polylines, exploiting their similarity and their distances.

5. Visualization Tools

5.1 Starfish's Visualizer

When a Map Reduce job executes in a Hadoop cluster, a lot of information is generated including logs, counters, resource utilization metrics, and profiling data. This information is organized, stored, and managed by Starfish's Metadata Manager in a catalog that can be viewed using Starfish's Visualizer. A user can employ the Visualizer to get a deep understanding of a job's behaviour during execution, and to ultimately tune the job. Broadly, the functionality of the Visualizer can be categorized into Timeline views, Data-flow views, and Profile views.

5.2 D3.js

D3.js, short for 'Data Driven Documents. It uses HTML, CSS and SVG to render some amazing charts and diagrams. It uses HTML, CSS, and SVG to render some amazing charts and diagrams. It is feature packed, interactivity rich and extremely beautiful. Most of all it's free and open-source.

5.3 SAS Visual Analytics

SAS Visual Analytics make sense of big data. SAS Visual Analytics uses intelligent autocharting to create the best possible visual based on the data that is selected. If SAS Visual Analytics determines that the data is geographic, a map frequency chart is used as shown in Figure 7.

6. Visualization Algorithm

6.1 Fast Map

Fast Map is a fast algorithm to map objects into points in some k-dimensional space (k is user-defined), such that the dis-similarities are preserved.

6.2 Self-Organizing Map (SOM)

SOM algorithm to analyse large data set and show that SOM is an excellent tool for the analysis and visualization of gene expression profiles

6.3 Sammon's Mapping Algorithms

The Sammon's mapping algorithm places data with similar, but not identical profiles in neighbouring groups creating a smooth transition of related profiles over the whole matrix.

7. Big Data Analytics and Data Visualization

The strategy for visualizing enormous data is critical and analysing is done to enhance the technique for customizing it to pick up consideration from the business examiner, information researchers and scientists. Examining time series data continuously is the testing undertaking and is accomplished for predicting the future for better comprehension of the behaviour[63]. The author gave an automated model to extricating learning from enormous information and discovering insights out of it. The data is produced at quick rate at the server side and the testing undertaking is to capture the information in the on-going and is to be put away in the database for the investigation of huge information and to visualize data in R. The broke down results are stored into the Mongodb for visualization purpose[63]. The broke down results are visualization in the R environment.

The author[64] had indicated the sources of steps, in light of the internal and external sources of upgraded decision making, insight discovery can be distinguished as 1. Acquisition of data from various sources, 2. Processing, 3. Visualize 4. Intelligence. Perception is helpful to draw the inferences and test the theories. Enterprise top management can take savvy decision from the visualization and patterns leaving big data analysis.

Big Data analytics and neural network technique helps to better know the objectives of diagnosing and treating patients in need of healthcare. The clinical data are usually typically questionable and also ambiguous with. So, in the paper[65] Neural Network is proposed to diagnose the diabetic disease from the clinical diabetic enormous data. Two scheme of Neural Network, namely BFGS quasi-Newton Back propagation and Resilient Back propagation are suggested for the diagnosis of diabetes[65]. From the result of the proposed methods it is experiential that Resilient Back propagation performs better contrasted with BFGS quasi-Newton Back propagation. Hence it is reasonable to expect a rapid boost in the understanding of Artificial Neural Network to analyse enormous Data efficiently.

The fundamental tasks of data visualisation are listed as filtering, Meta data interpretation, find extreme maximum, sort and shuffle, determination of range, clustering, data correlation and intricate computation. The paper[66] focus on the key term called data correlation and shuffle and sort sequence where job scheduling and job sequence takes a vital part. Hadoop yarn will be an advanced version of Map reduce, YARN is used to handle those general processing system beyond the Map reduce. YARN is an advanced processing engine used to run numerous jobs or applications in HADOOP by involvement those resources. One of the optimal scheduling algorithms utilized in the naïve HADOOP might have been "late" algorithm. Data correlation values for high dimensional data which surpasses the ordinary dimensionality within the attribute are correlated to attain better results in terms of accuracy, Kappa values and so forth.

8. Conclusion

Visual analytics is an emerging field of research combining strengths from information analytics, geospatial analytics, scientific analytics, statistical analytics, knowledge discovery, data management and knowledge representation, presentation, production and dissemination, cognition, perception and interaction. Its goal is to gain insight into homogeneous, contradictory and incomplete data through the combination of automatic analysis methods with human background knowledge and intuition.

We systematically selected and rigorously analysed a comprehensive set of visualization techniques and tools in order to provide an evidential based knowledge about the current state of visualization and the potential areas of research. From the initially identified 2 papers through manual and automatic searches, 66 papers have been selected based on the inclusion and exclusion criteria for this review.

7. References

1. Keim DA, Mansmann F, Schneidewind J, Ziegler H. Challenges in visual data analysis. 10th International Conference on Information Visualization; London, England. 2006 Jul 5. p. 9–16.

2. Wong PC, Bergeron RD. 30 years of multidimensional multivariate visualization. Scientific Visualization. 1994 May 1; 3–33.
3. Card SK, Mackinlay JD, Shneiderman B. Readings in information visualization: Using vision to think. Morgan Kaufmann; 1999.
4. Keim DA. Information visualization and visual data mining. IEEE Transactions on Visualization and Computer Graphics. 2002 Jan; 8(1):1–8.
5. Cleveland WS. Visualizing data. Hobart Press; 1993 Sep 1.
6. Alpern B, Carter L. The hyper box. Proceedings of IEEE, Conference on Visualization (Visualization'91); CA, USA.1991 Oct 22. p. 133–9.
7. Mihalisin T, Gawlinski E, Timlin J, Schwegler J. Visualizing a scalar field on an N-dimensional lattice. Proceedings of the 1st Conference on Visualization'90; San Francisco, CA. 1990 Oct 23. p. 255–62.
8. Van Wijk JJ, Van Liere R. Hyper Slice: Visualization of scalar functions of many variables. Proceedings of the 4th Conference on Visualization'93; DC, USA.1993 Oct 25. p. 119–25.
9. Chernoff H. The use of faces to represent points in K-dimensional space graphically. Journal of the American Statistical Association. 1973 Jun 1; 68(342):361–8.
10. Pickett RM, Grinstein GG. Iconographic displays for visualizing multidimensional data. Proceedings of the IEEE Conference on Systems, Man and Cybernetics; 1988 Aug 8. p. 519.
11. Beddow J. Shape coding of multidimensional data on a microcomputer display. Proceedings of the 1st Conference on Visualization'90; San Francisco, CA. 1990 Oct 23. p. 238–46.
12. Levkowitz H. Color icons: Merging colour and texture perception for integrated visualization of multiple parameters. Proceedings of the 2nd Conference on Visualization'91; San Francisco, CA. 1991 Oct 22. p. 164–70.
13. Mihalisin T, Gawlinski E, Timlin J, Schwegler J. Visualizing a scalar field on an n-dimensional lattice. Proceedings of the 1st Conference on Visualization'90; San Francisco, CA. 1990 Oct 23. p. 255–62.
14. LeBlanc J, Ward MO, Wittels N. Exploring n-dimensional databases. Proceedings of the 1st Conference on Visualization'90; San Francisco, CA. 1990 Oct 23. p. 230–7.
15. Inselberg A. The plane with parallel coordinates. The Visual Computer. 1985 Aug 1; 1(2):69–91.
16. Inselberg A, Dimsdale B. Parallel coordinates: A tool for visualizing multi-dimensional geometry. IEEE Proceedings of the 1st Conference on Visualization; 1990. San Francisco CA. p. 361–75.

17 Keim DA, Kriegel HP. VisDB: Database exploration using multidimensional visualization. IEEE Computer Graphics and Applications. 1994 Sep; 14(5):40–9.

18. Andrews DF. Plots of high-dimensional data. Biometrics. 1972 Mar 1; 28(1):125–36.
19. Friedman JH, Tukey JW. A projection pursuit algorithm for exploratory data analysis. IEEE Transactions on Computers. 1974; 23(9):981–890.
20. Huber PJ. Projection pursuit. The annals of Statistics. 1985 Jun 1; 13(2):435–75.
21. Wegman EJ. Hyperdimensional data analysis using parallel coordinates. Journal of the American Statistical Association. 1990 Sep 1; 85(411):664–75.
22. Furnas GW, Buja A. Prosection views: Dimensional inference through sections and projections. Journal of Computational and Graphical Statistics. 1994 Dec 1; 3(4):323–53.
23. Wright W. Research report: Information animation applications in the capital markets. Proceedings on Information Visualization; 1995 Oct 30. p. 19–25.
24. Hoffman P, Grinstein G, Marx K, Grosse I, Stanley E. DNA visual and analytic data mining. Proceedings on Visualization'97; USA. 1997 Oct 24. p. 437–41.
25. Kandogan E. Star coordinates: A multi-dimensional visualization technique with uniform treatment of dimensions. Proceedings of the IEEE Information Visualization Symposium; 2000. p. 22.
26. Kandogan E. Visualizing multi-dimensional clusters, trends and outliers using star coordinates. Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining; NY, USA. 2001 Aug 26. p. 107–16.
27. Pickett RM. Visual analysis of texture in the detection and recognition of objects. Picture Processing and Psychopictorics. 1970; p. 289–308.
28. Tufte ER, Graves-Morris PR. The visual display of quantitative information. Cheshire, CT: Graphics Press; 1983 Oct.
29. Ankerst M, Keim DA, Kriegel HP. Circle segments: A technique for visually exploring large multidimensional data sets. Proceedings of Visualization; San Francisco, CA. 1996. p. 1–4.
30. Keim DA, Kriegel HP. VisDB: Database exploration using multidimensional visualization. IEEE Computer Graphics and Applications. 1994 Sep; 14(5):40–9.
31. Keim DA, Ankerst M, Kriegel HP. Recursive pattern: A technique for visualizing very large amounts of data. Proceedings of the 6th Conference on Visualization'95; GA. 1995 Oct 29. p. 279.
32. LeBlanc J, Ward MO, Wittels N. Exploring n-dimensional databases. Proceedings of the 1st Conference on Visualization'90; San Francisco, CA. 1990 Oct 23. p. 230–7.
33. Feiner SK, Beshers C. Worlds within worlds: Metaphors for exploring n-dimensional virtual worlds. Proceedings of the 3rd Annual ACM SIGGRAPH Symposium on User Interface Software and Technology; NY, USA. 1990 Aug 1. p. 76–83.

34. Robertson GG, Mackinlay JD, Card SK. Cone trees: Animated 3D visualizations of hierarchical information. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems; NY, USA. 1991 Apr 27. p. 189–94.
35. Shneiderman B. Tree visualization with tree-maps: 2-d space-filling approach. ACM Transactions on Graphics (TOG). 1992 Jan 2; 11(1):92–9.
36. Johnson BS. Treemaps: Visualizing hierarchical and categorical data [Doctoral Dissertation]. USA: University of Maryland at College Park; 1993.
37. Rekimoto J, Green M. The information cube: Using transparency in 3D information visualization. Proceedings of the 3rd Annual Workshop on Information Technologies and Systems (WITS'93); Canada. 1993 Dec 5. p. 125–32.
38. Eick SG, Wills GJ. Navigating large networks with hierarchies. Proceedings of IEEE Conference on Visualization; San Jose, CA. 1993 Oct 25. p. 204–10.
39. Rusu A, Santiago C, Jianu R. Real-time space-efficient synchronized tree-based web visualization and design. 2006.
40. Artero AO, de Oliveira MC. Viz3d: Effective exploratory visualization of large multidimensional data sets. Proceedings of 17th Brazilian Symposium on Computer Graphics and Image Processing; 2004 Oct 17. p. 340–7.
41. Cooprider ND, Burton RP. Extension of star coordinates into three dimensions. Electronic Imaging. 2007 Jan 28; 64950.
42. Shaik JS, Yeasin M. Visualization of high dimensional data using an automated 3D star co-ordinate system. International Joint Conference on Neural Networks (IJCNN'06); 2006 Jul 16. p. 1339–46.
43. Shaik J, Yeasin M. Selection of best projection from 3D star coordinate projection space using energy minimization and topology preserving mapping. International Joint Conference on Neural Networks (IJCNN). Orlando, FL. 2007 Aug 12. p. 2604–9.
44 Chen K, Liu L. Clustermap: Labeling clusters in large datasets via visualization. Proceedings of the 13th ACM International Conference on Information and Knowledge Management; GA. 2004 Nov 13. p. 285–93.
45. Teoh ST, Ma KL. StarClass: Interactive visual classification using star coordinates. SDM 2003 Jan 1. p. 178–85.
46. Sun Y, Tang J, Tang D, Xiao W. Advanced star coordinates. The Ninth International Conference on Web-Age Information Management (WAIM'08); 2008 Jul 20. p. 165–70.
47. Dhillon IS, Modha DS, Spangler WS. Visualizing class structure of multidimensional data. Computing Science and Statistics. 1998 May 13;488–93.
48. Dhillon IS, Modha DS, Spangler WS. Class visualization of high-dimensional data with applications. Computational Statistics and Data Analysis. 2002 Nov 28; 41(1):59–90.
49. Toronen P, Kolehmainen M, Wong G, Castren E. Analysis of gene expression data using self-organizing maps. FEBS Letters. 1999 May 21; 451(2):142–6.
50. Faloutsos C, Lin KI. FastMap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. ACM; 1995 Jun 1. p. 1–12.
51. Yin H. ViSOM - A novel method for multivariate data projection and structure visualization. IEEE Transactions on Neural Networks. 2002 Jan; 13(1):237–43.
52. Aftab Z, Tuaseef H. Enhancing pixel oriented visualization by merging circle view and circle segment visualization techniques. Multi-Disciplinary Trends in Artificial Intelligence. 2012 Dec 26; 7694:101–9.
53. Tatu A, Albuquerque G, Eisemann M, Schneidewind J, Theisel H, Magnor M, Keim D. Combining automated analysis and visualization techniques for effective exploration of high-dimensional data. IEEE Symposium on Visual Analytics Science and Technology (VAST); Atlantic City, NJ. 2009 Oct 12. p. 59–66.
54. Graham M, Kennedy J. Using curves to enhance parallel coordinate visualisations. IV Proceedings of 7th International Conference on Information Visualization; 2003 Jul 16. p. 10–6.
55. Moustafa R, Wegman E. Multivariate continuous data-parallel coordinates. Graphics of Large Datasets. 2006;143–55.
56. Zhou H, Yuan X, Qu H, Cui W, Chen B. Visual clustering in parallel coordinates. Computer Graphics Forum. 2008 May 1; 27(3):1047–54.
57. Yang L. Visualizing frequent item sets, association rules and sequential patterns in parallel coordinates. Computational Science and Its Applications - ICCSA; Heidelberg. 2003 May 18. p. 21–30.
58. Yang J, Peng W, Ward MO, Rundensteiner EA. Interactive hierarchical dimension ordering, spacing and filtering for exploration of high dimensional datasets. IEEE Symposium on Information Visualization (INFOVIS); DC, USA. 2003 Oct 19. p. 105–12.
59. Kosara R, Bendix F, Hauser H. Parallel sets: Interactive exploration and visual analysis of categorical data. IEEE Transactions on Visualization and Computer Graphics. 2006; 12(4):558–68.
60. Siirtola H, Laivo T, Heimonen T, Raiha KJ. Visual perception of parallel coordinate visualizations. 13th International Conference on Information Visualisation; Barcelona. 2009 Jul 15. p. 3–9.
61. Johansson J, Treloar R, Jern M. Integration of unsupervised clustering, interaction and parallel coordinates for the exploration of large multivariate data. IV Proceedings of 8th International Conference on Information Visualisation; 2004 Jul 14. p. 52–7.
62. Bertini E, Dell'Aquila L, Santucci G. Springview: Cooperation of radviz and parallel coordinates for view optimization and clutter reduction. Proceedings of 3rd International Conference on Coordinated and Multiple Views in Exploratory Visualization; DC, USA. 2005 Jul 5. p. 22–9.

63. Parthiban P, Selvakumar S. Big data architecture for capturing, storing, analysing and visualizing of web server logs. Indian Journal of Science and Technology. 2016 Jan 17; 9(4).

64. Puri GD, Haritha D. Survey Big Data Analytics, Applications and Privacy Concerns. Indian Journal of Science and Technology. 2016 May 18; 9(17).

65. Sapna S, Kumar MP. Diagnosis of disease from clinical big data using neural network. Indian Journal of Science and Technology. 2015 Sep 1; 8(24):1.

66. Koteeswaran S, Visu P, Kannan E. Enhancing JS–MR based data visualisation using YARN. Indian Journal of Science and Technology. 2015 Apr 1; 8(11).