A Study on Big Data Visualization Techniques

P.Alagambigai
Assistant Professor
Dept. of Information Technology,
Women's Christian College, Chennai
alagambigai@yahoo.co.in

D.R.Angela Deepa
Assistant Professor
Dept. of Information Technology,
Women's Christian College, Chennai

P.Maqlin
Assistant Professor
Dept. of Information Technology,
Women's Christian College, Chennai

Abstract **–** Due to the enormous increase in the data volume, in recent years, there is a big demand for obtaining knowledge/regularity from the big data, terabytes/ petabytes of data, to create business values or make society more sophisticate and efficient. Data visualization can help to deal with this. The specific advantage of visual data exploration is that the user able to directly involved in the analysis process. There has been a wide variety of data visualization techniques which have been developed over the last decade. This paper surveys the visualization techniques which are commonly used for data exploration and mining. More specifically, this paper deals with the big data visualization and the impact of data mining through visualization**.**

.

Keywords **-** Clustering, Data Mining, Geometric Projection, Icon Based Visualization, Information Visualization, Scientific Visualization.

I. INTRODUCTION

Data mining, a synonym to "knowledge discovery in databases" is a process of analysing data from different perspectives and summarizing it into useful information [8, 10, 11]. It is a process that allows users to understand the substance of relationships between data. It reveals patterns and trends that are hidden among the data. It is often viewed as a process of extracting valid, previously unknown, non-trivial and useful information from large databases. Data mining systems can be classified according to the kinds of databases mined, the kinds of knowledge mined, the techniques used or the applications. Three important components of data mining systems are databases, data mining engine, and pattern evaluation modules. Big Data mining is the capability of extracting useful information from these large datasets or streams of data, that due to its volume, variability, and velocity, it was not possible before to do it. There are many applications of Big Data, for example the following :

- Business: costumer personalization, churn detection
- Technology: reducing process time from hours to seconds
- Health: mining DNA of each person, to discover, monitor and improve health aspects of every one
- Smart cities: cities focused on sustainable economic [24].

Visualization is defined by Ware as "a graphical representation of data or concepts", which is either an "internal construct of the mind" or an "external artifact supporting decision making". Visualization provides valuable assistance to the humans by representing information visually. This assistance may be called cognitive support. Visualization can provide cognitive support through a number of mechanisms such as grouping related information for easy search and access, representing large volumes of data in a small space and imposing structure on data and tasks that can reduce time complexity, allowing interactive exploration through manipulation of parameter values [6].

More recently there has been a lot of discussion on using visualization for data mining. Considering visualization as a supporting technology in data mining, four possible approaches are stated in [5]. The first approach is the usage of visualization technique to present the results that are obtained from mining the data in the database. Second approach is applying data mining technique to visualization by capturing essential semantics visually. The third approach is to use visualization techniques to complement the data mining techniques. The fourth approach uses visualization technique to steer mining process. This paper focuses on the survey of various visualization techniques used for data mining.

The rest of the paper is organized as follows. The overview of data visualization is discussed in section 2. The usage of visualization in data mining is presented in section 2. The Knowledge assisted visualization is discussed in section 4. The big data visualization is described in section 5 and. Section 6 concludes the paper with directions for future research work.

II. DATA VISUALIZATION

Visualization techniques could enhance the current knowledge and data discovery methods by increasing the user involvement in the interactive process. To incorporate visualization techniques, the existing clustering algorithms use the result of clustering algorithm as the input for visualization system. The drawback of such approach is that it can be costly and inefficient. The better solution is to combine two processes together, which means to use the same model in clustering and visualization. Interactive clustering allows the user to be involved into the clustering and visualizing process via interactive visualization [2, 3, 19].

A. *Taxonomy of Visualization*

Visualization has been categorized into major areas [73, 81, 82]:

- Scientific visualization which involves scientific data with an inherent physical component.
- Information visualization – which involves abstract nonspatial data.

Scientific Visualization

Scientific visualization focuses primarily on physical data such as human body, the earth, molecules and so on. It also deals with multidimensional data, but most of the datasets used in this field use the spatial attributes of the data for visualization purpose; e.g. Computer Aided Tomography (CAT) and Computer Aided Design (CAD). Also, many of the Geographical Information Systems (GIS) use either the cartesian coordinate system or some modified geographical coordinates to achieve a reasonable visualization of the data.

Information Visualization

It focuses on abstract, nonphysical data such as text, hierarchies and statistical data. Data mining techniques are primarily oriented toward information visualization. The challenge for nonphysical data is in designing a visual representation of multidimensional samples. Multidimensional information visualization presents data that are not primarily plenary or spatial.

Geometric Projection Techniques

Geometric Projection techniques aim at finding “interesting” projections of multidimensional datasets. The class of geometric projection techniques include techniques for exploratory statistics such as Principal Component, Factor Analysis and multidimensional scaling, many of which are subsumed under the term “projection pursuit”. Parallel coordinate visualization techniques and Radial Visualization (RadViz) also belong to this category of visualization.

Icon based Techniques

The idea of icon based technique or iconic display is to map each multidimensional data item into an icon (or glyph) whose visual features vary depending on the data values. Some of the most commonly used iconic displays are Chernoff, Stick Figure, Star Display, Shape coding, etc.

Pixel oriented Techniques

Pixel oriented techniques map each data values to a coloured pixel and present the data values belonging to one attribute in separate window. All pixel oriented techniques partition the screen into multiple windows. For data sets with m dimension, the screen is partitioned into m windows: one for each of the dimensions.

Hierarchical and Graph based Techniques

The hierarchical techniques subdivide the m-dimensional space and present the subspaces in a hierarchical fashion. Well known representatives of hierarchical techniques are n-Vision technique, the dimensional stacking, and treemaps [19]. The basic idea of the graph based techniques is to effectively present a large graph using specific layout algorithms, query languages and abstraction techniques.

III. VISUALIZATION IN DATA MINING

Various efforts are made to visualize multidimensional datasets. With a wide range of cluster analysis in data mining, many information visualization techniques are employed in recent years [6, 12, 13, 14, 18, 19]. A short review on visual cluster analysis is as follows: Grand Tour [20] is a method for viewing multidimensional data via linear projections onto a sequence of two dimensional subspaces and then moving continuously from one projection to the next. So that, the user can looks at the high-dimensional data from different prospective.

OPTICS (Ordering Points To Identify the Clustering Structure) proposed by Mihael Ankerst, et al. [3] is a density based visual cluster analysis technique which is used to detect the cluster structure and visualize them in 'Gaussian Bumps'. 3D cluster-guided tour is an extension of Grand Tour [20] where sequences of projections are determined by cluster centroids. T. C. Sprenger et al., [17] proposed a new hierarchical clustering algorithm named Hierarchical BLOB (H-BLOB) that provides an efficient level of detail strategy and consequently becomes capable to cluster and visualize very large and complex data volumes.

The Kohonen Map (or Self-Organizing Map, or SOM)[16], is a well-known approach suited for analysis of large volumes of high-dimensional data. High-dimensional Eye (HD-Eye) [12] system is an interactive visual clustering system, visualizes the density-plot of the fascinating projection of any two of the m dimensions. It uses icons to represent both the clusters and relationships between the clusters.

Star coordinate based visual cluster analysis system is proposed by Kandogan [13] to visualize and analyze the clusters interactively. iVIBRATE [14, 15] is an interactive-visualization based three-phase framework for clustering large datasets. The two main components of iVIBRATE [15] are its VISTA [14] visual cluster rendering subsystem, which invites human into the large-scale iterative clustering process through interactive visualization, and its Adaptive Cluster Map Labeling subsystem, which offers visualization-guided disk-labeling solutions that are effective in dealing with outliers, irregular clusters, and cluster boundary extension.

To overcome the arbitrary and random adjustments of Star coordinates and its extensions, HOV^3 (Hypothesis Oriented Verification and Validation by Visualization) approach is proposed by Zhang, et al. [21]. Zhang, et al. [21] also proposed an external validation method using HOV3. In this approach, a clustered subset from a dataset is chosen as a visual

model to verify the similarity of cluster structures between the model and other same sized non-cluster subsets from the dataset by projecting them together in HOV^3.

Computational visualization techniques are used to explore, in an immerse fashion, inherent data structure in both an unsupervised and supervised manner. Supervision is provided (i) Domain knowledge contained in the data (ii) Unsupervised data mining procedures such as k-means[1] and rough k-means.

Even though visualization techniques have advantages over automatic methods, it brings up some specific problems such as limitation in visibility, visual bias due to mapping of dataset to 2D/ 3D representation, easy-to-use visual interface operations and reliable human-computer interaction[15].

IV. KNOWLEDGE ASSISTED VISUALIZATION

Researchers have attempted to clarify the taxonomy of terms used in the visualization community. Whereas Min Chen et. al [22] attempts to offer a different taxonomy for visualization, that provides a new dimension on visualization processes. From a systems perspective, however, data is referred to as bits and bytes stored on or communicated via a digital medium. So any computerized representations, including knowledge representations, are types of data. On the other hand, from the perspective of knowledge-based systems, data is a simpler form of knowledge. "Information visualization" is for "data mining and knowledge discovery." In other cases, these three terms indicate data types, for instance, as adjectives in noun phrases, such as data visualization, information visualization, and knowledge visualization. Min Chen et. al [22] defines three types of visualization: Data, Information and Knowledge visualization. Table 1. shows the definitions presented by Russell Ackoff's [23] and Min Chen et. al.

Table 1. Definitions of data, information and knowledge by Russell Ackoff and Min Chen et. al

Category	*Min Chen et. al*	*Russell Ackoff*
	Definition	
Data	Computerized representations of models and attributes of real or simulated entities.	Symbol
Information	Data that represents the results of a computational process, such as statistical analysis, for assigning meanings to the data, or the transcripts of some meanings assigned by human beings.	Data that are processed to be useful, providing answers to "who," "what," "where," and "when" questions
Knowledge	Data that represents the results of a computer-simulated cognitive process, such as perception, learning, association, and reasoning, or the transcripts of some knowledge acquired by human beings.	Application of data and information, providing answers to "how" questions

Information Assisted Visualization

In information-assisted visualization, the system provides the user with a second visualization pipeline, which typically displays the information about the input data set. But it can also present attributes of the visualization process, the properties of the results, or characteristics of the user's perceptual behaviors. The user uses such information to reduce

the search space for optimal control parameters, hence making the interaction much more cost effective.

Knowledge Assisted Visualization

Knowledge assisted visualization include sharing domain knowledge among different users and reducing the burden on users to acquire knowledge about complex visualization techniques. It also enables the visualization community to learn and model the best practice, so that powerful visualization infrastructures can develop and evolve. Both information and knowledge assisted visualization plays a major role in the process of data mining. They not only used for visualizing the results, but also complement and steer the mining process.

V. BIG DATA VISUALIZATION

Big data are varied from traditional data whose size is beyond the ability of commonly used algorithms and computing systems to capture, manage, and process the data within a reasonable time. Big Data Mining and Analytics discovers hidden patterns, correlations, insights and knowledge through mining and analyzing large amounts of data obtained from various applications. Big data holds out big promises to drive deeper analysis and valuable insights from significantly increased repositories of information. While these have the potential to deliver real business value and change the way an organisation operates, achieving those goals requires a lens to make data easy to interrogate and clearly show the insight. That is where data visualization tools come in.

Issues in Big Data Visualization

Big Data visualization deals with presentation the data in different type of graphical format, which helps us to understand the data and interpret the results easily. Due to complexity in the Big data, the traditional and well known visualization tools are unable to express the underlying structure of data which leads to the requirement of complex representations such as heat maps, fever charts, symbol maps and connectivity charts The primary challenges stem from what are commonly termed the "three Vs" of big data: volume, variety, and velocity. Most traditional reporting and data mining tools cannot handle the vast volume of big data—although the variety and velocity of the data often present even greater challenges. Another key challenge in analyzing big data relates to its velocity. The rapid generation of big data can lead to significant business insights and predictions, but only if real-time data can be analyzed quickly in hours rather than weeks or months. There has been a wide variety of researchers [25, 26, 27] focused on big data visualization and its challenges.

VI. CONCLUSION

Visualization is known to be most intuitive method for validating the results obtained from various data mining approaches, highly effective in explaining the underlying patterns of the data and also powerful in revealing trends. In this paper, we address a brief study on visualization techniques and the emerging trend in big data visualization. There are a number of directions in which research on Knowledge assisted visualization can be continued.

REFERENCES

[1] Ahmad A., Dey L., "A K-Means Clustering Algorithm for Mixed Numeric and Categorical Data", Data and Knowledge Engineering, Vol. 63, pp. 503-507, 2007.

[2] P. Alagambigai, K. Thangavel, N. Karthikeyani Visalakshi, "Improved Visual Cluster Rendering System", in: K. Thangavel.(ed.), Intelligent and Computing Model, Narosa Publishing House, New Delhi, pp. 16-23, 2009.

[3] Alagambigai, P., Thangavel, K., "Visual Clustering through Weight Entropy", International Journal on Data Mining, Modelling and Management, Vol.2,No:3,2010

[4] Ankerst M., Breunig M., Kriegel H. P., Sander J.,"OPTICS: Ordering Points To Identify the Clustering Structure," In: Proceedings of ACM SIGMOD '99, International Conference on Management of Data, Philadelphia, pp. 49-60, 1999

[5] Bhavani Thuraisingham (1999), "DataMining: Technologies, Techniques, Tools and Trends",CRC press, London,Newyork, Washington,1999.
[6] Daniel A.Keim and Hans-Peter (1996), "Visualization Techniques for Mining Large Databases: A Comparison", IEEE Transactions on Knowledge and Data Engineering, Vol. 8, No. 6, Dec 1996, pp.923-937.
[7] G., Guha., R., Rastogi and K., Shim(1998)., "CURE: An efficient clustering algorithm for large databases", Proc. of the ACM SIGMOD, 73-84,1998.
[8] Han, J., & Kamber, M. (2000). Data mining: Concepts and Techniques. Morgan Kaufman Publishers.
[9] Hoffman, P., & Grinstein, G. (2002). A survey of visualizations for high-dimensional data mining. In U. Fayyad, G. Grinstein & A. Wierse (Eds.), Information visualization in data mining and knowledge discovery. California: Morgan Kaufmann.
[10] Jain, A.K., Murty, M.N., Flynn, P.J.(1999). Data Clustering: A Review, (Ed.), ACM Computing Surveys,264-323
[11] Ravinchandra Rao I.K., "Data mining and clustering techniques," DRTC Workshop on Semantic Web, Bangalore, 8 th -10 th December 2003.
[12] A., Hinnerburg, D., Keim and M. Wawryniuk(1999), "HD-Eye: Visual Mining of High – Dimensional Data," IEEE Computer Graphics and Applications, Vol.19, No.5, 1999.
[13] E. Kandogan (2001)," Visualizing Multi-dimensional Clusters Trends and outliers using star co-ordinates, Proc of ACM KDD, 2001.
[14] Keke Chen and Liu. Lc(2004) , "VISTA: "Validating and Refining clusters via Visualization", Information Visualization 3, 4, 257-270,2004.
[15] Keke Chen and Liu.L(2004a), "iVIBRATE:" Interactive Visualization-Based Framework for Clustering Large Datasets", ACM Transactions on Information Systems, Vol. 24, April 2006, Pages 245-294.
[16] Schreck T, Bernard J., Tekusova T., Kohlhammer j., "Visual Cluster Analysis of Trajectory Data with Interactive Kohonen Maps," In: Proceedings IEEE VAST 2008, pp. 3–10, 2008.
[17] Sprenger T. C., Brunella R., Gross M. H., "H-BLOB: A Hierarchical Visual Clustering Method using Implicit Surfaces," In: Proceedings of Visualization 2000, pp. 61-68, 2000.
[18] Thangavel. K and Alagambigai. P., " EVISTA-Interactive Visual Clustering System", international Journal on Recent trends in Engineering, pp.83-37,2009.
[19] Winnie Wing-Yi Chan, "A survey on Multivariate Data visualization", June 2006.
[20] Yang. Li, "Interactive Exploration of Very Large Relational Datasets through 3D Dynamic Projections", Porc. Of SIGKDD2001.
[21] Ke-Bing Zhang, Mehmet A. Orgun and Kang Zhang, "A Prediction-Based Visual Approach for Cluster Exploration and Cluster Validation by HOV3", KNOWLEDGE DISCOVERY IN DATABASES: PKDD 2007, Lecture Notes in Computer Science, 2007, Volume 4702/2007, 336-349.
[22] M. Chen, D. Ebert, H. Hagen, R. Laramee, R. Van Liere, K. Ma, W. Ribarsky, G. Scheuermann, and D. Silver. Data, Information, and Knowledge in Visualization. IEEE Computer Graphics and Applications, 29(1):12–19, 2009.
[23] R.L. Ackoff, "From Data to Wisdom," J. Applied Systems Analysis, vol. 16, 1989, pp. 3–9.
[24] Wei Fan, Albert Bifet, "Mining Big Data: Current Status, and Forecast to the Future", SIGKDD Explorations Volume 14, Issue 2, pp.1-5, 2013.
[25] Lidong Wang, Guanghui Wang, Cheryl Ann Alexander, " Big Data and Visualization: Methods, Challenges and Technology Progress", Digital Technologies, 2015, Vol. 1, No. 1, 33-38
[26] Daniel Keim ; Huamin Qu ; Kwan-Liu Ma, "Big-Data Visualization", IEEE Computer Graphics and Applications, Volume: 33 Issue: 4 , 2013
[27] Alejandro Sanchez ; Wilson Rivera, "Big Data Analysis and Visualization for the Smart Grid" 2017 IEEE International Congress on Big Data