Digital Technologies, 2015, Vol. 1, No. 1, 33-38
Available online at http://pubs.sciepub.com/dt/1/1/7
© Science and Education Publishing
DOI:10.12691/dt-1-1-7

Big Data and Visualization: Methods, Challenges and Technology Progress

Lidong Wang[1,*], Guanghui Wang[2], Cheryl Ann Alexander[3]

[1]Department of Engineering Technology, Mississippi Valley State University, USA
[2]State Key Laboratory of Severe Weather, Chinese Academy of Meteorological Sciences, China
[3]Technology and Healthcare Solutions, Inc., USA
*Corresponding author: lwang22@students.tntech.edu

Received June 27, 2015; Revised July 16, 2015; Accepted July 20, 2015

Abstract Big Data analytics plays a key role through reducing the data size and complexity in Big Data applications. Visualization is an important approach to helping Big Data get a complete view of data and discover data values. Big Data analytics and visualization should be integrated seamlessly so that they work best in Big Data applications. Conventional data visualization methods as well as the extension of some conventional methods to Big Data applications are introduced in this paper. The challenges of Big Data visualization are discussed. New methods, applications, and technology progress of Big Data visualization are presented.

Keywords: *Big Data, visualization, interactive visualization, virtual reality, networks, cloud computing, information technology, telecommunication systems*

Cite This Article: Lidong Wang, Guanghui Wang, and Cheryl Ann Alexander, “Big Data and Visualization: Methods, Challenges and Technology Progress.” *Digital Technologies*, vol. 1, no. 1 (2015): 33-38. doi: 10.12691/dt-1-1-7.

1. Introduction

Data visualization is representing data in some systematic form including attributes and variables for the unit of information [1]. Visualization-based data discovery methods allow business users to mash up disparate data sources to create custom analytical views. Advanced analytics can be integrated in the methods to support creation of interactive and animated graphics on desktops, laptops, or mobile devices such as tablets and smartphones [2]. Table 1 [3] shows the benefits of data visualization according to the respondent percentages of a survey.

Table 1. Benefits of data visualization tools

Benefits	Percentages (%)
Improved decision-making	77
Better ad-hoc data analysis	43
Improved collaboration/information sharing	41
Provide self-service capabilities to end users	36
Increased return on investment (ROI)	34
Time savings	20
Reduced burden on IT	15

There are some points of advice for visualization [4]: (1) Do not forget the metadata. Data about data can be very revealing. (2) Participation matters. Visualization tools should be interactive, and user engagement is very important. (3) Encourage interactivity. Static data tools don’t lead to discovery as well as interactive tools do.

Big data are high volume, high velocity, and/or high variety datasets that require new forms of processing to enable enhanced process optimization, insight discovery and decision making. Challenges of Big Data lie in data capture, storage, analysis, sharing, searching, and visualization [5]. Visualization can be thought of as the “front end” of big data. There are following data visualization myths [4]:

- All data must be visualized: It is important not to overly rely on visualization; some data does not need visualization methods to uncover its messages.
- Only good data should be visualized: A simple and quick visualization can highlight something wrong with data just as it helps uncover interesting trends.
- Visualization will always manifest the right decision or action: Visualization cannot replace critical thinking.
- Visualization will lead to certainty: Data is visualized doesn’t mean it shows an accurate picture of what is important. Visualization can be manipulated with different effects.

Visualization approaches are used to create tables, diagrams, images, and other intuitive display ways to represent data. Big Data visualization is not as easy as traditional small data sets. The extension of traditional visualization approaches have already been emerged but far from enough. In large-scale data visualization, many researchers use feature extraction and geometric modeling to greatly reduce data size before actual data rendering. Choosing proper data representation is also very important when visualizing big data [5].

The goal and the objectives of this paper are to present new methods and advances of Big Data visualization through introducing conventional visualization methods and the extension of some them to handling big data, discussing the challenges of big data visualization, and analyzing technology progress in big data visualization.

In this study, authors first searched for papers that are related to data visualization and were published in recent years through the university library system. At this stage, authors mainly summarized traditional data visualization methods and new progress in this area. Next, authors searched for papers that are related to big data visualization. Most of these papers were published in the past three years because big data is a newer area. At this stage, authors found that most conventional data visualization methods do not apply to big data. The extension of some conventional visualization approaches to handling big data is far from enough in functions. The authors focused on big data visualization challenges as well as new methods, technology progress, and developed tools for big data visualization.

2. Conventional Data Visualization Methods

Many conventional data visualization methods are often used. They are: table, histogram, scatter plot, line chart, bar chart, pie chart, area chart, flow chart, bubble chart, multiple data series or combination of charts, time line, Venn diagram, data flow diagram, and entity relationship diagram, etc. In addition, some data visualization methods have been used although they are less known compared the above methods. The additional methods are: parallel coordinates, treemap, cone tree, and semantic network, etc. [1].

Parallel coordinates is used to plot individual data elements across many dimensions. Parallel coordinate is very useful when to display multidimensional data. Figure 1 shows parallel coordinates. Treemap is an effective method for visualizing hierarchies. The size of each sub-rectangle represents one measure, while color is often used to represent another measure of data. Figure 2 shows a treemap of a collection of choices for streaming music and video tracks in a social network community. Cone tree is another method displaying hierarchical data such as organizational body in three dimensions. The branches grow in the form of cone. A semantic network is a graphical representation of logical relationship between different concepts. It generates directed graph, the combination of nodes or vertices, edges or arcs, and label over each edge [1].

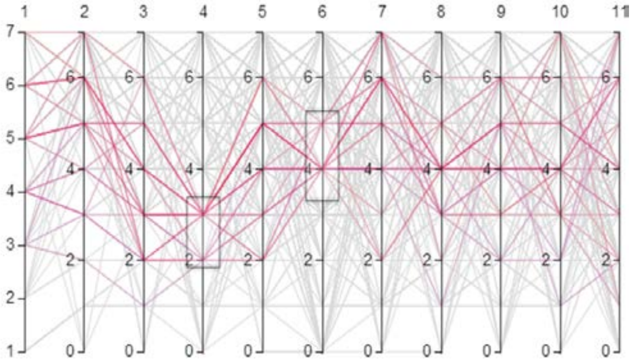

Figure 1. Parallel coordinates [6]

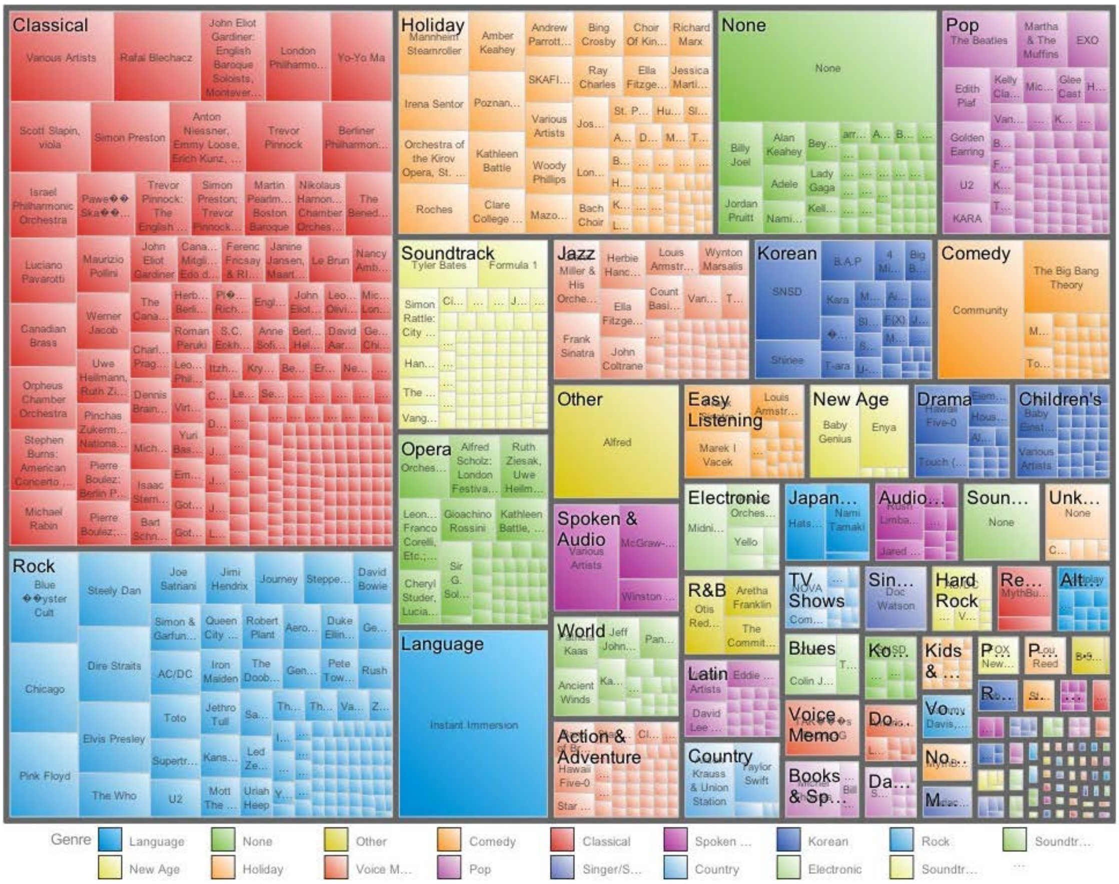

Figure 2. Treemap view of a social network's track selections from a streaming media service [7]

Visualizations are not only static; they can be interactive. Interactive visualization can be performed through approaches such as zooming (zoom in and zoom out), overview and detail, zoom and pan, and focus and context or fish eye [1]. The steps for interactive visualization are as follows [1]:

1. *Selecting:* Interactive selection of data entities or subset or part of whole data or whole data set according to the user interest.
2. *Linking:* It is useful for relating information among multiple views. An example is shown in Figure 3.
3. *Filtering:* It helps users adjust the amount of information for display. It decreases information quantity and focuses on information of interest.
4. *Rearranging or Remapping:* Because the spatial layout is the most important visual mapping, rearranging the spatial layout of the information is very effective in producing different insights.

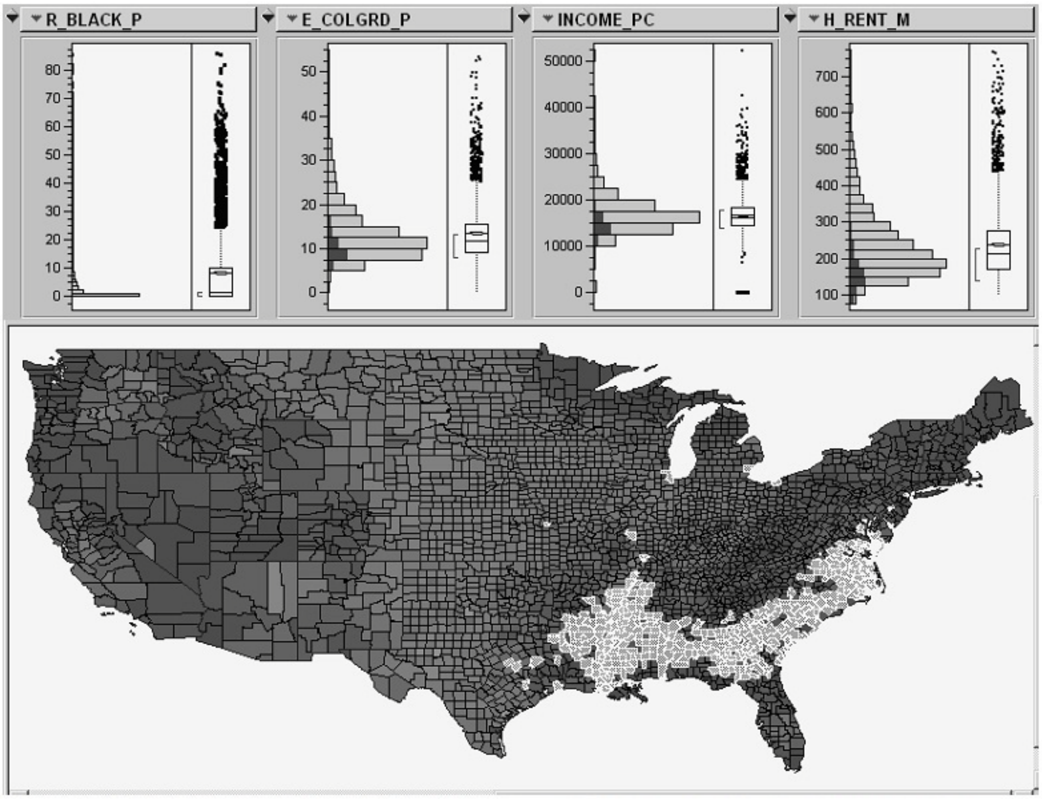

Figure 3. Interactive brushing and linking between histogram plots (top) and geographic map (bottom) of datasets [1]

New database technologies and promising Web-based visualization approaches may be vital for reducing the cost of visualization generation and allowing it to help improve the scientific process. Because of Web-based linking technologies, visualizations change as data change, which greatly reduces the effort to keep the visualizations timely and up to date. These “low-end” visualizations have been often used in business analytics and open government data systems, but they have generally not been used in the scientific process. Many visualization tools that are available to scientists do not allow live linking as do these Web-based tools [8].

3. Challenges of Big Data Visualization

Scalability and dynamics are two major challenges in visual analytics. Table 2 shows the research status for static data and dynamic data according to the data size. For big dynamic data, solutions for type A problems or type B problems often do not work for A and B problems [9].

Table 2. The research status and challenge of visual analytics

Data type	Small, mid-sized	Big-sized
Static data	Well studied	Open issues type A
Dynamic data	Open issues type B	Highly challenging (A and B) >> A+B

The visualization-based methods take the challenges presented by the “four Vs” of big data and turn them into following opportunities [2].

- *Volume*: The methods are developed to work with an immense number of datasets and enable to derive meaning from large volumes of data.
- *Variety*: The methods are developed to combine as many data sources as needed.
- *Velocity*: With the methods, businesses can replace batch processing with real-time stream processing.
- *Value*: The methods not only enable users to create attractive infographics and heatmaps, but also create business value by gaining insights from big data.

Visualization of big data with diversity and heterogeneity (structured, semi-structured, and unstructured) is a big problem. Speed is the desired factor for the big data analysis. Designing a new visualization tool with efficient indexing is not easy in big data. Cloud computing and advanced graphical user interface can be merged with the big data for the better management of big data scalability [3].

Visualization systems must contend with unstructured data forms such as graphs, tables, text, trees, and other metadata. Big data often has unstructured formats. Due to bandwidth limitations and power requirements, visualization should move closer to the data to extract meaningful information efficiently. Visualization software should be run in an in situ manner. Because of the big data size, the need for massive parallelization is a challenge in visualization. The challenge in parallel visualization algorithms is decomposing a problem into independent tasks that can be run concurrently [10].

Effective data visualization is a key part of the discovery process in the era of big data. For the challenges of high complexity and high dimensionality in big data, there are different dimensionality reduction methods. However, they may not always be applicable. The more dimensions are visualized effectively, the higher are the chances of recognizing potentially interesting patterns, correlations, or outliers [11].

There are also following problems for big data visualization [12]:

- *Visual noise:* Most of the objects in dataset are too relative to each other. Users cannot divide them as separate objects on the screen.
- *Information loss:* Reduction of visible data sets can be used, but leads to information loss.
- *Large image perception:* Data visualization methods are not only limited by aspect ratio and resolution of device, but also by physical perception limits.
- *High rate of image change: Users* observe data and cannot react to the number of data change or its intensity on display.
- *High performance requirements:* It can be hardly noticed in static visualization because of lower visualization speed requirements--high performance requirement.

Perceptual and interactive scalability are also challenges of big data visualization. Visualizing every data point can lead to over-plotting and may overwhelm users' perceptual and cognitive capacities; reducing the data through sampling or filtering can elide interesting structures or outliers. Querying large data stores can result in high latency, disrupting fluent interaction [13].

In Big Data applications, it is difficult to conduct data visualization because of the large size and high dimension of big data. Most of current Big Data visualization tools have poor performances in scalability, functionalities, and response time. Uncertainty can result in a great challenge to effective uncertainty-aware visualization and arise during a visual analytics process [5].

Potential solutions to some challenges or problems about visualization and big data were presented [14]:

1. Meeting the need for speed: One possible solution is hardware. Increased memory and powerful parallel processing can be used. Another method is putting data in-memory but using a grid computing approach, where many machines are used.
2. Understanding the data: One solution is to have the proper domain expertise in place.
3. Addressing data quality: It is necessary to ensure the data is clean through the process of data governance or information management.
4. Displaying meaningful results: One way is to cluster data into a higher-level view where smaller groups of data are visible and the data can be effectively visualized.
5. Dealing with outliers: Possible solutions are to remove the outliers from the data or create a separate chart for the outliers.

4. Some Progress of Big Data Visualization

As for how visualization should be designed in the era of big data, visualization approaches should provide an overview first, then allow zooming and filtering, and provide deep details on demand [15]. Visualization can play an important role in using big data to get a complete view of customers. Relationships are an important aspect of many big data scenarios. Social networks are perhaps the most prominent example and are very difficult to understand in text or tabular format; however, visualization can make emerging network trends and patterns apparent [7]. A cloud-based visualization method was proposed to visualize an inherence relationship of users on social network. The method can intuitionally present the users' social relationship based on the correlation matrix to represent a hierarchical relationship of user nodes of social network. In addition, the method uses Hadoop based on cloud for the distributed parallel processing of visualization, which helps expedite the big data of social network [16].

Big data visualization can be performed through a number of approaches such as more than one view per representation display, dynamical changes in number of factors, and filtering (dynamic query filters, star-field display, and tight coupling), etc. [12]. Several visualization methods were analyzed and classified [12] according to data criteria: (1) large data volume, (2) data variety, and (3) data dynamics.

Treemap: It is based on space-filling visualization of hierarchical data.

Circle Packing: It is a direct alternative to treemap. Besides the fact that as primitive shape it uses circles, which also can be included into circles from a higher hierarchy level.

Sunburst: It uses treemap visualization and is converted to polar coordinate system. The main difference is that the variable parameters are not width and height, but a radius and arc length.

Parallel Coordinates: It allows visual analysis to be extended with multiple data factors for different objects.

Streamgraph: It is a type of a stacked area graph that is displaced around a central axis resulting in flowing and organic shape.

Circular Network Diagram: Data object are placed around a circle and linked by curves based on the rate of their relativeness. The different line width or color saturation is usually used to measure object relativeness.

Table 3 and Table 4 [12] show the classifications. Table 3 indicates which method can process large volume data, various data, and changing data with time. According to Table 4, visualization methods can be classified according to Big Data classes.

Table 3. Properties of visualization methods

Method name	Large data volume	Data variety	Data dynamics
Treemap	+	-	-
Circle packing	+	-	-
Sunburst	+	-	+
Parallel coordinates	+	+	+
Streamgraph	+	-	+
Circular network diagram	+	+	-

Table 4. Classification of visualization methods

Method name	Big data class
Treemap	Can be applied only to hierarchical data
Circle packing	Can be applied only to hierarchical data
Sunburst	Volume + Velocity
Parallel coordinates	Volume + Velocity + Variety
Streamgraph	Volume + Velocity
Circular network diagram	Volume + Variety

Traditional data visualization tools are often inadequate to handle big data. Methods for interactive visualization of big data were presented. First, a design space of scalable visual summaries that use data reduction approaches (such as binned aggregation or sampling) was described to visualize a variety of data types. Methods were then developed for interactive querying (e.g., brushing and linking) among binned plots through a combination of multivariate data tiles and parallel query processing. The developed methods were implemented in imMens, a browser-based visual analysis system that uses WebGL for data processing and rendering on the GPU [13].

A lot of big data visualization tools run on the Hadoop platform. The common modules in Hadoop are: Hadoop Common, Hadoop Distributed File System (HDFS), Hadoop YARN, and Hadoop MapReduce. They analyze big data efficiently, but lack adequate visualization. Some software with the functions of visualization and interaction for visualizing data has been developed [3]:

Pentaho: It supports the spectrum of BI functions such as analysis, dashboard, enterprise-class reporting, and data mining.

Flare: An ActionScript library for creating data visualization that runs in Adobe Flash Player.

JasperReports: It has a novel software layer for generating reports from the big data storages.

Dygraphs: It is quick and elastic open source JavaScript charting collection that helps discover and understand opaque data sets.

Datameer Analytics Solution and Cloudera: Datameer and Cloudera have partnered to make it easier and faster to put Hadoop into production and help users to leverage the power of Hadoop.

Platfora: Platfora converts raw big data in Hadoop into interactive data processing engine. It has modular functionality of in-memory data engine.

ManyEyes: It is a visualization tool launched by IBM. Many Eyes is a public website where users can upload data and create interactive visualization.

Tableau: It is a business intelligence (BI) software tool that supports interactive and visual analysis of data. It has an in-memory data engine to accelerate visualization.

Tableau has three main products to process large-scale datasets, including Tableau Desktop, Tableau Sever, and Tableau Public. Tableau also embed Hadoop infrastructure. It uses Hive to structure queries and cache information for in-memory analytics. Caching helps reduce the latency of a Hadoop cluster. Therefore, it can provide an interactive mechanism between users and Big Data applications [5].

Big data processing tools can process ZB (zettabytes) and PB (petabytes) data quite naturally, but they often cannot visualize ZB and PB data. At present, big data processing tools include Hadoop, High Performance Computing and Communications, Storm, Apache Drill, RapidMiner, and Pentaho BI. Data visualization tools include NodeBox, R, Weka, Gephi, Google Chart API, Flot, D3, and Visual.ly, etc. A big data visualization algorithm analysis integrated model based on RHadoop was proposed. The integrated model can process ZB and PB data and show valuable results via visualization. The model is suitable for the design of parallel algorithms for ZB and PB data [17].

Interactive visual cluster analysis is the most intuitive way for discovering clustering patterns. The most challenging step is visualizing multidimensional data and allowing users to interactively explore the data and identify clustering structures. Optimized star-coordinate visualization models for effective interactive cluster exploration on big data were developed. The star-coordinate models are probably the most scalable technique for visualizing large datasets compared with other multidimensional visualization methods such as parallel coordinates and scatter-plot matrix [18]:

- Parallel coordinates and scatter-plot matrix are often used for less than ten dimensions, while star coordinates can handle tens of dimensions.
- The star-coordinate visualization can scale up to many points with the help of density-based representation.
- Star-coordinate based cluster visualization does not try to calculate pairwise distances between records; it uses the property of the underlying mapping model to partially keep the distance relationship. This is very useful in processing big data.

Direct visualization of big data sources is often not possible or effective. Analytics plays a key role by helping reduce the size and complexity of big data. The visualization and analytics can be integrated so that they work best. IBM has embedded visualization capabilities into business analytics solutions. What makes this possible is the IBM Rapidly Adaptive Visualization Engine (RAVE). RAVE and extensible visualization capabilities help use effective visualization that provides a better understanding of big data [7]. IBM products, such as IBM® InfoSphere® BigInsights™ and IBM SPSS® Analytic Catalyst, use visualization libraries and RAVE to enable interactive visualizations that can help gain great insight from big data. InfoSphere BigInsights is the software that helps analyze and discover business insights hidden in big data. SPSS Analytic Catalyst automates big data preparation, chooses proper analytics procedures, and display results via interactive visualization [7].

The use of immersive virtual reality (VR) platforms for scientific data visualization has been in the process of exploration including software and inexpensive commodity hardware. These potentially powerful and innovative tools for multi-dimensional data visualization can provide an easy path to collaborative data visualization. Immersion provides benefits beyond traditional "desktop" visualization tools: it results in a better perception of data scape geometry and more intuitive data understanding. Immersive visualization should become one of the foundations to explore the higher dimensionality and abstraction that are attendant with big data. The intrinsic human pattern recognition (or visual discovery) skills should be maximized through using emerging technologies associated with the immersive VR [11].

The SWOT (Strengths, Weaknesses, Opportunities, and Threats) analysis is a well-known method to ensure that both positive factors and negative factors are identified. A SWOT analysis of the above software tools for big data visualization has been conducted and is shown in Table 5. In Table 5, Strengths and Opportunities are positive factors; Weaknesses and Threats are negative factors.

Table 5. The SWOT analysis of current big data visualization software tools

Strengths	Opportunities
• With the functions of visualization and interaction for visualizing data. • Able to visualize a variety of data types.	• Immersive visualization with virtual reality (VR) results in a better perception of data scape geometry and more intuitive data understanding. • The intrinsic human pattern recognition (or visual discovery) skills could be maximized.
Weaknesses	**Threats**
• There is room to improve in visualizing big data with high velocity or the combination of three high Vs (Volume + Velocity + Variety).	• Lack adequate visualization in a lot of Big Data applications.

5. Conclusions

Visualizations can be static or dynamic. Interactive visualizations often lead to discovery and do a better job than static data tools. Interactive visualizations can help gain great insight from big data. Interactive brushing and linking between visualization approaches and networks or Web-based tools can facilitate the scientific process. Web-based visualization helps get dynamic data timely and keep visualizations up to date.

The extension of some conventional visualization approaches to handling big data is far from enough in functions. More new methods and tools of Big Data visualization should be developed for different Big Data applications. Advances of Big Data visualization are presented and a SWOT analysis of current visualization software tools for big data visualization has been conducted in this paper. This will help develop new methods and tools for big data visualization. Big Data analytics and visualization can be integrated tightly to work best for Big Data applications. Immersive virtual reality (VR) is a new and powerful method in handling high dimensionality and abstraction. It will facilitate Big Data visualization greatly.

Acknowledgment

This research was supported in part by Technology and Healthcare Solutions, Inc. in Mississippi, USA.

References

[1] M. Khan, S.S. Khan, Data and Information Visualization Methods and Interactive Mechanisms: A Survey, *International Journal of Computer Applications*, 34(1), 2011, pp. 1-14.

[2] Intel IT Center, Big Data Visualization: Turning Big Data Into Big Insights, White Paper, March 2013, pp.1-14.

[3] V. Sucharitha, S.R. Subash and P. Prakash , Visualization of Big Data: Its Tools and Challenges, *International Journal of Applied Engineering Research,* 9(18), 2014, pp. 5277-5290.

[4] P. Simon, The Visual Organization: Data Visualization, Big Data, and the Quest for Better Decisions, *Harvard Business Review*, June 13, 2014, pp. 1-8.

[5] C.L. P. Chen, C.-Y. Zhang, Data-intensive applications, challenges, techniques and technologies: A survey on Big Data, *Information Sciences*, 275 (10), August 2014, pp. 314-347.

[6] B. Porter, Visualizing Big Data in Drupal: Using Data Visualizations to Drive Knowledge Discovery, Report, University of Washington, October 2012, pp. 1-38.

[7] T. A. Keahey, Using visualization to understand big data, Technical Report, IBM Corporation, 2013, pp. 1-16.

[8] P. Fox and J. Hendler, Changing the Equation on Scientific Data Visualization, *Science*, 331(11), February 2011, pp. 705-708.

[9] I. B. Otjacques, UniGR Workshop: Big Data- The challenge of visualizing big data, Report, Gabriel Lippmann, 2013, pp. 1-24.

[10] H. Childs, B. Geveci, J. Meredith, K. Moreland, C. Sewell, E.W. Bethel, T. Kuhlen, W. Schroeder, Research Challenges for Visualization Software, Joint Research Report of Lawrence Berkeley National Laboratory, Oak Ridge National Laboratory, Sandia National Laboratories, Los Alamos National Laboratory, RWTH Aachen University (Germany), May 2013, pp. 1-11.

[11] C. Donalek, S.G. Djorgovski, A. Cioc, A. Wang, J. Zhang, E. Lawler, S. Yeh, A. Mahabal, M. Graham, A. Drake, S. Davidoff, J.S. Norris, G. Longo, Immersive and Collaborative Data Visualization Using Virtual Reality Platforms, 2014 IEEE International Conference on Big Data, pp. 1-6.

[12] E.Y. Gorodov and V.V. Gubarev, Analytical Review of Data Visualization Methods in Application to Big Data, *Journal of Electrical and Computer Engineering*, 013, Article ID 969458, pp. 1-7.

[13] Z. Liu, B. Jiangz and J. Heer, imMens: Real-time Visual Querying of Big Data, Eurographics Conference on Visualization (EuroVis) 2013, 32(3), 2013, pp. 421-430.

[14] SAS Institute Inc., Five big data challenges and how to overcome them with visual analytics, Report, 2013, pp. 1-2.

[15] F. Shull, Getting an Intuition for Big Data, *IEEE Software*, July/August 2013, pp. 1-5.

[16] Y. Kim, Y.-K. Ji and S. Park, Social Network Visualization Method using Inherence Relationship of User Based on Cloud, *International Journal of Multimedia and Ubiquitous Engineering*, 9(4), 2014, pp. 13-20.

[17] L. Cai, X. Guan, P. Chi, L. Chen, and J. Luo, Big Data Visualization Collaborative Filtering Algorithm Based on RHadoop, *International Journal of Distributed Sensor Networks*, Article ID 271253, pp. 1-10.

[18] K. Chen, Optimizing star-coordinate visualization models for effective interactive cluster exploration on big data, *Intelligent Data Analysis*, 18, 2014, pp. 117-136.