Problems, Challenges, and Opportunities Visualization on Big Data

Tri Wahyuningsih *

Graduated Program Informatics, Rahaja University, Indonesia
triwahyuningsih@rahaja.info
* corresponding author

(Received: July 15, 2020; Revised: July 21 2020; Accepted: August 29, 2020; Available online: September 1, 2020)

Abstract

Today, almost everything is recorded digitally, from our browsing history to our health records in hospitals, we make and process billions of data every day. In this era of big data, large amounts of data are continuously obtained for different purposes. But, just processing and analyzing the data isn't enough. If data is displayed visually, humans always search for patterns more effectively. Visualization and interpretation of the data are very critical tasks in making choices in various industries. This also guides us to new ways to find innovative ideas through visualization to solve big-data problems. In this paper, we will discuss the problems, challenges, and potential of Visualization in Big Data.

Keywords: Big data, Visualization, Data, Data science, Technology

1. Introduction

Every year the amount of data created by people, machines and corporations around the world is increasing. Big data technology has become an attraction in recent years for many sectors such as education, IT businesses and government[1]. In recent years , data growth rates have risen rapidly due to factors such as the Internet of Things and digitization of all of the offline information such as medical and education records. Big data has shown how important he is in this world, this technology has been used by almost all sectors and industries to store their data. However, The analysis of big data isn't a simple job and it needs different methods and approaches.

Big Data technology's biggest challenge lies in data collection , processing, sharing, searching, and visualization. The key feature of the Big Data research is that in very large data sets we can find interesting patterns, but in general the results of the analysis are ordinary raw numbers and it is very difficult to interpret something with those numbers. Given these challenges, data reduction and reducing latency should be applied to optimize the visualization of big data.

Visualization seems to be a very simple and logical type of data processing for a human brain, so it could be proven that the representation of visual information is an effective method, allowing raw data to be eased and providing adequate support for decision making. if the figures are represented visually, our brain will be much easier to find meaningful patterns and make appropriate decisions.

2. Literature Review

Big Data is data that is really extensive and complicated that traditional data processing systems are unable to manage it. As data size and varieties expand it becomes challenging to process and consume the insights produced. However there are many different definitions around the world, in 2011, an IDC study[2] described big data as "big data technologies describe a new generation of technologies and architectures designed to economically extract value from very large quantities of a wide variety of data by enabling high-speed collection, discovery, and/or analysis."

Big data characteristics could be summarised as 4V Which is *Volume (Scale of data)*, *Velocity (Speed of Data), Veracity (Certainly of Data), Variety (Diversity of Data).* 4Vs characteristics indicates the big data most critical issue, which is how to identify values from massive scale datasets, different types, and fast generation.

On the other hand, NIST[3] describes big data as 'Big data means data of which data volume, processing speed or representation limit the ability to use traditional relational methods to perform successful analysis or data that could be processed effectively using significant horizontal zoom technologies,' focusing on the technological aspect of big data.It leads to the need to build and use effective methods or technologies to analyze and process big data.

Additionally, querying large data disrupts the fluent interaction. It is also critical that large volume data is presented in real-time. Such drawbacks pose problems such as scalability in real time, digital scalability, visual scalability. As it comes to big data today, how it looks will help to convey information but it has to be more than just simplistic and elegant. It should operate, this should display several dimensions, it must be useful. Big data visualization also provides opportunities to present better ways of viewing big data, such as data reduction, latency reduction, etc.

2.1. Challenge

Visualizing large data sets is a daunting task. Traditional ways of presenting data will reach some limitations and the data will grow very large. Visualization tools and techniques should be able to assist users in identifying missing, error and duplicated values. It is challenging to overcome limitations such as perceptual scalability, real-time scalability, and interactive scalability. In this section, we will discuss this challenge

2.1.1 Perceptual Scalability

Human perception: Human eye has difficulty extracting important or meaningful information when data becomes larger. Not many visualization systems today are designed to present meaningful and quality information for human perception

Limited Screen: Data gets bigger and stays bigger, very challenging when visualization displays too much data or features on a limited screen, especially datasets with very many entries. If there is too much data to display on a limited screen the visualization results are too dense to be useful for the user [4]. The limitations of screen resolution force us to explore new ways to display and visualize information using various abstraction techniques. It is even more difficult to present big data on a mobile device (Mobile) because of the smaller screen and resolution.

2.1.2 Real-Time Scalability

Real-time scalability It is important to provide users with real-time visual information and it is also important to make real-time decisions based on available data [5]. However, large amounts of data will become too large to be processed in real time. Almost all visualization systems are only designed to handle data under a certain size because many data sets are too large to be loaded into memory and large data requests can cause high latency. Very challenging to overcome limitations such as data connectivity, limited storage and data processing capabilities in real time.

2.1.3 Interactive Scalability

Interactivity increases the benefits of data visualization. Interactive data visualization can help us understand insights into data faster and better. However, it takes time to process and analyze data before visualization, especially large amounts of data. and, the visualization system may even freeze for some time or even fail when trying to present large amounts of data. Measuring complex queries to terabytes of query processing techniques together with enabling interactive response times is a major issue of open research today [6].

2.2 Problem

When encountered with very large data sets, traditional visualization techniques have approached their limits, and these data change continuously. While there are several alternatives to traditional method to visualization yet they lag so far behind. It is difficult to visualize because of it's large volume and the high magnitude of the big data. The majority of the existing visualization tool has weak scalability, flexibility, and response time. However, data visualization has a variety of existing and coming problems:

2.2.1 Data that too simple

One of it's advantages of visualization is the ability to take vast quantities of data and reduce it into terms that are much more simple and understandable. It's easy to step with this though. Looking to take vast amounts of data and then limiting their assumptions to multiple representations can lead to absurd conclusions, or totally disregard some important change that can completely change the assumptions you leave behind. Consider , for example, common real-world tests, such as tests for alcohol poisoning, which aim to simplify complicated processes to clear "yes" or "no" results. Such assessments can be both inaccurate and imprecise.

2.2.2. Human limitations in the algorithm

This would be the biggest possible problem, and perhaps the most difficult one. Every algorithm used to reduce data into visual representations is focused on feedback from humans, and human involvement can be inaccurate. For instance , a person who develops an algorithm can identify various pieces of information that are "most" critical to analyze and discard the other pieces. It doesn't take all companies or all circumstances into consideration, particularly when there are international data or specific situations that need alternative approaches. The problem can be exacerbated by the fact that most of the data visualization systems are implemented on a regional scale, transforming into a one-size solution for everyone, and failing to meet individual unique needs.

2.2.3. Excessive dependence on visuals

It is a concern for users rather than developers, but it hurts the future effect of the visualization in general. Once users start relying on visuals to view data that they can use at a glance, users are going to over-rely on this input mode. For starters, they would find their findings as absolute facts, and never look deeper into the actual data set for the visuals to create. The assumptions you draw from it may be broadly relevant however they didn't reveal you anything about your audience or program.

2.2.4. Visualization that inevitable

There are hundreds of resources available today to help people identify complex data sets such as graphs , charts and visual representations, so the visualization of data is too common to be passed on. Humans are in time where the simulation takes over in different fields, and at the moment nothing is actual. It might not sound like a issue for some people but remember some of the consequences. Companies compete to create products for visualization and customers are only searching for things that provide great visualization. These outcomes can affect users who rely too much on visuals, and intensify the limits of human error in the creation of algorithms.

2.3. Opportunity

Challenges certainly brings an opportunity. This section discusses several opportunities and strategies to overcome the challenges of the big data visualization above. Using data reduction techniques include sampling, filtering, and aggregation to reduce big data into smaller data that can be received before visualization. They can overcome several problems such as: large-scale data by reducing the resolution of the results requests and displaying all the information needed in a limited display (limited display). We will also mention reducing latency techniques, which parallelize data processing and rendering, hide disk latency, and compile multivariate data tiles. we will discuss this approach in more detail in the following section.

2.3.1. Data Reduction

Data reduction strategies include sampling, filtering, and binned aggregation, which reduces bid data to smaller data that can be received before visualization [7].

- **Sampling:** The reduction technique is based on sampling. Each dataset is a sample. When given a probability value, it returns roughly a small portion of the data as a result.
- **Filtering:** Filtering techniques are needed to explore and request big data sets. With a set of conditions desired for querying the data, return the element that satisfies this condition.
- **Binned Aggregation:** Data is grouped into a subset, and a summary of the subset is returned as a result. Binning aggregates data by counting the number of data points included in each specified bin.

2.3.2. Reducing Latency

Pre-computed data: to increase interactive scalability such as shifting, zooming in and pulling. Pre-computational data is a popular strategy for visualization. This can support fast exploration rather than producing tile images intended for direct viewing. For example: Google Maps and Data Cubes.

Parallel Data Processing and Rendering: Data tiles can be very large in the aggregation process. That depends on the binning resolution. If the data plot has more than millions of values, it increases aggregation latency. To speed things up, visualization can use a solid indexing scheme that simplifies the processing of parallel queries.

Predictive middleware: predictive middleware that will be between the frontend visualization interface and backend data storage and will predict, pre-retrieve, and store relevant data. Predictive middleware increases real-time scalability and interactive scalability. This hides latency backend data storage by retrieving and storing data needed in the near future.

3. Research Methods

3.1. TreeMap

There is a strict restriction on this method for data objects which need to be connected hierarchically since this method is based on the space-filling visualization of hierarchical data. The Treemap is defined by a root rectangle, separated into groups, represented also by the smaller rectangles corresponding to data objects from a dataset. This method can be applied into large quantities of data, describing iteratively data layers for each hierarchical level. In this event of excessive device resolution, the analyst can always move to the next block to continue his research into more detailed data on lower hierarchy level.

The benefits of the method:

- Hierarchical classification clearly indicates the relationship to data.
- While using special color, extreme outliers are immediately visible.

The disadvantages of the method:

- Not effective for reviewing historical tendencies and time patterns.
- Negative values can not be the factor used for size calculation.

3.2. Circle Packing

This method is alternative to treemap, besides the fact that it uses circles as its primitive form, which can also be included from a higher level of hierarchy into circles. The main benefit of this method is that by using classical Treemap[8] we can possibly place and perceive a larger amount of objects. Because the method of circle packing is based on the Tree-map method, they have the same properties. So, researchers can infer that this method meets only the criterion for large volumes of data.

The benefits of the method:

- Type of spatially effective visualisation compared to Treemap.

The disadvantages of the method:

- It has the same disadvantages just like Treemap method.

3.3. Sunburst

This method uses Treemap visualization that converted into polar coordinate system, so this method can be defined as Treemap alternative. The Variable parameters on this method isn’t width and height, it’s a arc and radius length. This difference helps us not to remodel the entire diagram after data change. But only one sector which contains new data by adjusting its radius. This method very effective to show data dynamics with animation.

The benefit of the method:

- Easy to understand for almost everybody [9].

The disadvantages of the method:

- Have the same disadvantages just like treemap.

3.4. Parallel Coordinates

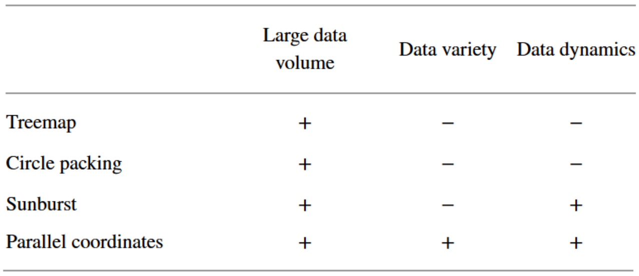
Here on this method researchers allowed visual analysis to be extended for different objects with multiple data factors. All data factors to be evaluated are put on one axis, and on the other, the corresponding data object values are put in relative scale. Each data object is represented by a series of interconnected traverse lines, showing its place in other object context. This method also allows us to use just a thick line up on the screen to represent individual data objects and this approach enables it to meet the first criterion of large volumes of data[10].

The whole method can manage multiple factors for a large number of objects for every single screen, therefore it meets the criterion of data variety. Since this method is based on relative values, the estimation of the minimum and maximum values for each factor is necessary. While values change between the minimum and maximum values of each factor, there is no need to repaint all images, but for a case where value exceeds this limit, we need to repaint the image in order to display adequate visualization. That approach can be used in dynamic data visualization. The alternative to view a data in time requires using three-dimensional extensions for the method of polar coordinates[11].

The benefit of the method: The ordering factors should not impact overall perceptions of the diagram, This method helps us to evaluate both entire object data collection and individual data objects simultaneously

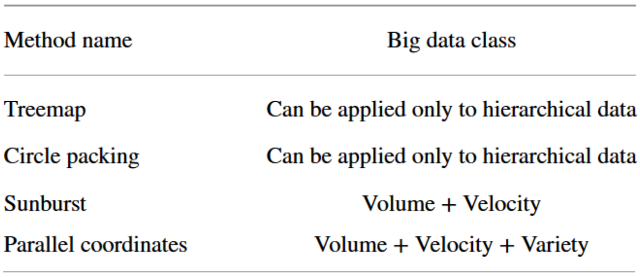
The disadvantages of the method: Dynamic data visualization ends up changing representation of whole data [12]

Table. 1. Visualization Method Attributes



	Large data volume	Data variety	Data dynamics
Treemap	+	–	–
Circle packing	+	–	–
Sunburst	+	–	+
Parallel coordinates	+	+	+

Table. 2. Visualization Techniques Classifications



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

After describing various method for visualization, Table 1 showing which method that can handle large volumes data, data variety and data dynamics. in the end, it can be proven that Treemap & Circle packing method can't be applied into Big Data classes since it only passes one criteria. And Table 2 Describes a few of the common visualization techniques when and how to use them or not.

4. Discussion

There are several different methods of graphical visualization, but visualization of multidimensional data is only still a little known and a specific subject of study. Graphical visualization itself has already been used in multiple aspects of our activity[13], but the efficiency and even implementation of approaches can become a real problem with the growth of data volumes and the speed of data production. The issue identified comes from as follows:

(1) The need for artificial data slices preparation, for partial visualizing information.

(2) Visual restriction of the perceived number of data variables.

Researchers need to study current methods of data visualization and develop methods that can address these issues. These approaches need to provide more perceptible and descriptive representations of data to help the analyst identify hidden connections in big data. Most of the methods of data visualization usually do not appear from nothing, but they become an evolution of previous traditional methods [14].

Analyst tools must fulfill some qualifications, such as analysts should be able to use more than one view of data representation at a time, active user interaction and analysable view, and continuous changing of the number of variables during the visual process.

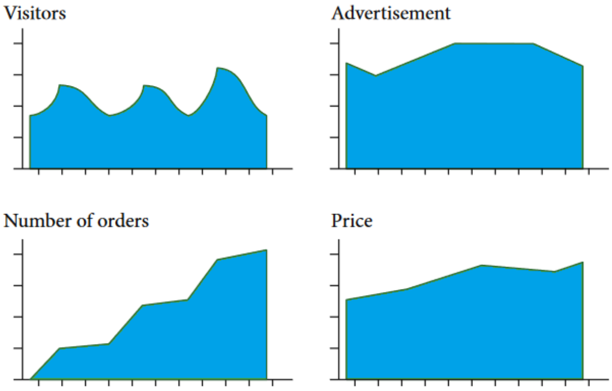
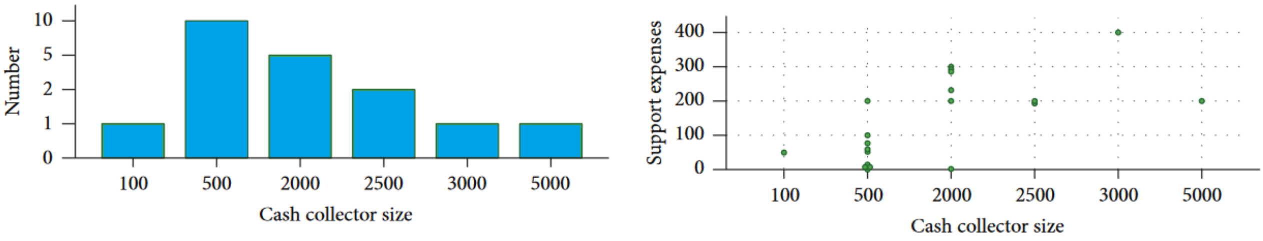


Fig. 1. Same view on multiple data representations.

Display with more than one view, In order to achieve a complete understand of data, analyst normally uses a simple approach when placing different classical views of data, which also include just a limited set-off of actors so that he could still easily obtain some relationships between these views or in a specific view[15].

Knowing the fact that any form of data visualization should be used entirely, we will also see an approach where the analyst uses only a few identical or related objects. For illustration, Fig. 1 above. The analyst could be interested in evaluating completely different visualizations of the same data, but in that case the entire process of visual analysis becomes much more difficult. Today, not only must the researcher compare related graphic objects, he must also explicitly discern different data and make a conclusion based on specific factors[16].

It can be assumed that an approach will direct analysts into the right position and provide appropriate help to make decisions at the very first analysis level. There might be some cases where this stage in the current research can become a final, driving analyst away from totally inaccurate choices. Analysts can choose whether to organize views in a number of ways: choosing objects in one view that show similar details in other views, or include selection parameters to exclude material from the other displays instead. Related navigation gives an additional method of coordination: scrolling or zooming over one view will control other views simultaneously[17].



Number of factors Dynamical changes, Maybe the most basic visual processing process is found in a specification for the data visualization. An analyst will suggest Visitors Number of Advertising Cost orders which information is to be displayed and how it can be presented to ease the perception of knowledge. Almost every graphical visualization can be implemented to every information. however, there is always an up-to-date question as to whether the method chosen is correctly applied to the data set to obtain any useful information? The researcher usually can not analyze the whole data collection for Big Data, locate patterns in it, or identify associations at the first look. A dynamic change in the number of factors is therefore another topical approach. Once the analyst have selected one factor, he is able to see a classical histogram that displays the amount of records allocated based on the form of data. For illustration look Fig. 2 above, They could see connection between the amount of cash collector units currently in operation in the payment system and the scale of each cash collector.

Since another factor has been selected by the analyst, for example supporting expense, the form of diagram has also modified into point diagram. The right section of Fig. 2 indicates the allocation of support expenses for each unit of cash collectors. Proceeding on, we will adjust the number of factors con-sequentially, reduce or increase the number of observable factors and we can see adjustments in the diagram. This cycle is iterative and can be replicated before the correct pattern is reached.

5. Conclusion

In a world of big data, where every piece of information in one way or another is very important, we rely on visual information to identify a valuable pattern. Although conventional methods of visualization do not suit data speed and volume, we need techniques that apply to all of the big data characteristics and give us results without leaving output and reacting to time. From this paper we discuss why big data visualization is so necessary and what issues and opportunities associated with this are. We also note that interactivity visualization is the most essential, and interactive visualization must be generated by a good visualization tool. We 're also studying how people are proposing new systems to meet that challenge.

References

[1] Jin, Xiaolong, et al. "Significance and challenges of big data research." *Big Data Research* 2.2 (2015): 59-64.

[2] Gantz, John, and David Reinsel. "Extracting value from chaos." *IDC iview* 1142.2011 (2011): 1-12.

[3] Naphade, Milind, et al. "Large-scale concept ontology for multimedia." *IEEE multimedia* 13.3 (2006): 86-91.

[4] Fröhlich, Bernd, and John Plate. "The cubic mouse: a new device for three-dimensional input." *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 2000.

[5] Yu, Yuen Tak, and Man Fai Lau. "A comparison of MC/DC, MUMCUT and several other coverage criteria for logical decisions." *Journal of Systems and Software* 79.5 (2006): 577-590.

[6] Ali, Syed Mohd, et al. "Big data visualization: Tools and challenges." *2016 2nd International Conference on Contemporary Computing and Informatics (IC3I)*. IEEE, 2016.

[7] Agrawal, Rajeev, et al. "Challenges and opportunities with big data visualization." *Proceedings of the 7th International Conference on Management of computational and collective intElligence in Digital EcoSystems*. 2015.

[8] Tedesco, Jon, et al. "Theius: a streaming visualization suite for hadoop clusters." *2013 IEEE International Conference on Cloud Engineering (IC2E)*. IEEE, 2013.

[9] Cawthon, Nick, and Andrew Vande Moere. "The effect of aesthetic on the usability of data visualization." *2007 11th International Conference Information Visualization (IV'07)*. IEEE, 2007.

[10] Few, Stephen. "Multivariate analysis using parallel coordinates." *Perceptual edge* (2006): 1-9.

[11] Johansson, Jimmy, et al. "Perceiving patterns in parallel coordinates: determining thresholds for identification of relationships." *Information Visualization* 7.2 (2008): 152-162.

[12] Edsall, Robert M. "The dynamic parallel coordinate plot: visualizing multivariate geographic data." *Proc. 19th International Cartographic Association Conference, Ottawa*. 1999.

[13] Helfman, Jonathan, and Joseph Goldberg. "Data visualization techniques." U.S. Patent No. 8,640,056. 28 Jan. 2014.

[14] Gorodov, Evgeniy Yur'evich, and Vasiliy Vasil'evich Gubarev. "Analytical review of data visualization methods in application to big data." *Journal of Electrical and Computer Engineering* 2013 (2013).

[15] Robertson, George, et al. "Effectiveness of animation in trend visualization." *IEEE transactions on visualization and computer graphics* 14.6 (2008): 1325-1332.

[16] Heer, Jeffrey, and Ben Shneiderman. "Interactive dynamics for visual analysis." *Queue* 10.2 (2012): 30-55.

[17] Cleveland, William S., and Robert McGill. "Graphical perception: Theory, experimentation, and application to the development of graphical methods." *Journal of the American statistical association* 79.387 (1984): 531-554.