ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 3, pp: (105-113), Month: July - September 2016, Available at: www.researchpublish.com

An Overview of Big Data Visualization Techniques in Data Mining

Samuel Soma Ajibade[1], Anthonia Adediran[2]

[1]Universiti Teknologi Malaysia, Faculty of Computing, Department of Computer Science, Skudia, Johor Bahru, Johor, Malaysia
[2] The Federal Polytechnic, SES, Department of Estate Management, Ado Ekiti, Ekiti State, Nigeria

Abstract: **The analytics of data holds an important function by the reduction of the size and complicated nature of data in data mining. Data Visualization is a major method which aids big data to get an absolute data perspective and as well the discovery of data values. Since the visualization techniques are so enormous, it can at times be difficult to know what appropriate technique to be used to visualize data. The purpose of representing data visually is to basically give an interpretation to what is insight without difficulties. At different point in time, the visualization techniques are being used to carry out different task which as well communicates different level of understanding. In this paper, we examined various data mining visualization techniques and how they can be well understood and utilized and then we made discussed our contributions in making research about the adequacy and inadequacy of data visualization technique in handling big data.**

***Keywords*: Data Visualization, data mining, Big Data, visualization technique.**

I. INTRODUCTION

Data as a raw fact are gathered together through different fields of study. Therefore there calls for an immediate need for the new generation of different tools and techniques which will help to derive useful and authentic knowledge from the fast growing data volumes, hence these techniques forms the subject matter of the newly independent knowledge discovery field in database (KDD). Hence, the need for Big data visualization; in section two of this study we discussed about the concept of Data Mining (KDD), Big Data and data visualization and also in section four we talked about the plot map of data visualization in data mining, in section four we discussed the various data visualization techniques in the discovery of knowledge and how they can be applied. Section five contains our discussions and contributions in carrying out a research about the adequacy of how the data visualization technique can handle big data and this was represented in a table and then in section six we concluded our study.

II. CONCEPT OF KDD, BIG DATA AND DATA VISUALIZATION

A. Data Mining (Knowledge discovery in Database):

Knowledge discovery in database makes use of various fields which includes business, statistics, computer science and also a structure of methodologies of which there are still more to the following: Artificial Intelligence (AI), machine learning, acquisition of knowledge for expert systems, machine learning, etc. Knowledge discovery in database recurs round examining and inventing or creating knowledge, the algorithms involved, processes to be used and as well the necessary mechanism that will be used to tackle the recovery of potential knowledge. A particular part of this activity that is very vital is identifying the direction and patterns from metadata through, which involves the semantic level that indicates the relationship of an entity. The technique of knowledge discovery in database has turned out to be a success with the huge scientific databases, which is notable in astronomy used to classify sky objects. Furthermore, business and

industrial oriented databases in finance, internet agents, marketing and manufacturing are combined with some applications [16].

Knowledge discovery in database is the removal of absolute, unknown information from data that has the great possibility of being useful. Assuming we have some facts (raw data) F, and a language (L) and a measure of certainty (C), then we can describe a pattern as being a statement (S) in language (L) in which relationships are being defined amongst a subset (F_S) of fact (F), which has certainty (C), whereby statement (S) is easier than the list of facts in (F_S). Therefore, when a pattern is interesting (according to a user-imposed interest measure) and it's adequately certain (again according to the user's criteria), then that is known as Knowledge [6].

A good example of the type of knowledge that is being sought for is the relationship that occurs between data. For instance, when a cosmetics seller get to know that a large number of his customers who buys perfumes also get to buy roll-on deodorant about 85% of the time. This is likely to have a commercial benefit such that the seller can decided to now change the location of the roll-on deodorant very close to the perfume, so as to increase the sales of both the perfume and the deodorant. We can consider an example such that we identify the current trend of a sales in a specific area or district which are either increasing or decreasing. Such information can help the decisions of the top level management also called the strategic managers though it is a bit of a constraint to get useful information from huge amounts of data that are stored.

B. Big Data:

The idea of big data is being confined to the field of computer science since the inception of computing. Initially big data is defined as the volume of data is difficult to adequately process because of how "big" the data is by the traditional database tools and techniques. Whenever a new medium of storing data was created, the amount of data that accesses it blows up all because it is being accessed without difficulty. Initially, structured data was being focused on immensely, but a lot of researchers and professionals later understood that a larger portion of information of the world is being vested in very huge and unstructured information, enormously in image and text form. We can say big data is the amount of data that is just too enormous beyond the ability to store, manage and process the data in an efficient and effective manner. Hence, these imitations have been recognized with the help of a robust analysis of the big data, its definite processing needs and how capable the tools are (software, hardware and methods) used to analyze it [14].

Big data also means making equivalent chances available for businesses to accomplish faster and greater perception that makes decision making stronger, also increase the experiences of customers and speed up the innovation pace. Unfortunately, nowadays majority of the big data does not yield value or meanings and businesses are over powered by the huge and various amount of data flowing into the day to day operations, and by this, such businesses find it so difficult to store up the data let alone interpret, present and examine the data in ways that are much significant [4].

C. Data Visualization:

When we talking about data visualization, then we mean act of data presentation in a graphical or pictorial layout. Data visualizations helps top management who are the decision makers to view analytics being visually represented, so it makes them to easily understand the complex ideas and identify the new structures or patterns. When visualization becomes interactive, then we are able to push the concept a little further thereby using technological tools to grasp more details from graphs and charts, therefore making changes to the data that is being seen and how such data is being processed.

It also means putting data forward and representing them in a particular methodical layout which contains some variables and attributes for bringing about information [8]. Visualization-based data discovery techniques gives room to business owners to make up sources of completely different data so as to invent custom analytical views [4].

III. A PLOT MAP OF DATA VISUALIZATION IN DATA MINING

There are different ways in which data visualization can be accessed, it solely depends on the area of interest of the researcher that is involved. The main thing here is data visualization being able to make knowledge known about the data that is being visualized. A number of times, professionals in the area of graphics and animation are much bothered about multi-dimensional data, all their work is focused new and different methods of presenting data in a graphical form and also the issues that surrounds its implementation. Also professionals in the area of Human Computer Interaction (HCI) get

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 3, pp: (105-113), Month: July - September 2016, Available at: www.researchpublish.com

bothered with the way multidimensional data is being visualized, but along the way, they may end up making use of a visualization method that is already in existence and their focus is now based on the way the user relates with it.

In a quest to adequately understand the part that data visualization in discovery of knowledge and data mining plays, a few techniques for the representation of multidimensional data is being discussed. Data visualization techniques is being used to represent data regardless of any mathematical analysis or any other one. Hence if it is being used in this particular way, then the method or technique could be considered as a data mining technique. Furthermore, there are visualization techniques that are used to depict knowledge that were recognized by some data mining techniques [9].

When we want to begin with selecting some data in knowledge discovery process, the mathematical or probably some more techniques are being used to obtain knowledge from that data, so therefore the point of origin for the process could be a visual representation of the data. When the kind of scenario occurs, the representation then acts as an investigation tool. More so, the knowledge that is now found can be represented by a data visualization technique which is being used at the end of the process. Take for example, for us to monitor a progress or even to represent a chosen subset of data, some data visualization tools can act either at intermediate steps or during knowledge discovery of data.

IV. DATA VISUALIZATION METHODS

Over the years, lots of visualization methods have been developed, so as to be able to represent large information and as well as examine them, and these methods comprises of characteristics like usability, interactivity, interface features etc., and because of this, they are not difficult to use. These methods are being used to visualize data because they have an evaluation mechanism. These methods are: histogram, line chart, table, pie chart, bar chart, scatter plot, bubble plot, area chart, flow chart, Venn diagram, data flow diagram, time line, multiple data series, entity relationship diagram, cone tree, semantic network, tree map and parallel coordinates. etc. [8]. In the section below, we are going to discuss some of these data visualization techniques.

A. Line Chart:

A line chart displays the relationship between each variable on the chart. Line charts are frequently used to make comparison between lots of items at the same time. The stacking lines are being used to also make comparison between the trends for multiple variables. One might decide to make use of the line charts when a variable change needs to be displayed. Take for example, there are 12 data points to plot or show, the best way to make those points understandable is to just display them in an order using a table [10]. The fact that one has some data points to plot or display doesn't mean that line graph is the best to pick, but you should consider the number of data points that you want to display which will tell the best visual method to pick. Data points are mostly being connected by a straight line, and line chart is actually an extension of Scatter plot. Some specific symbols and icons are being used to represent data points in a line chart [3].

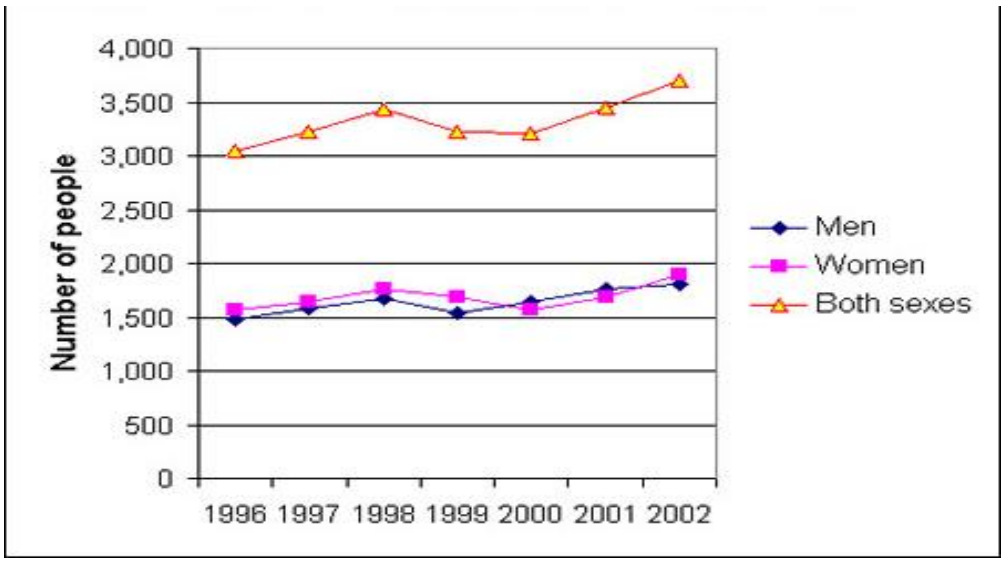

Fig.1 Line charts displaying the relationship of one variable to another. Shown here, different category line charts makes comparison of multiple items over the same time period [29].

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 3, pp: (105-113), Month: July - September 2016, Available at: www.researchpublish.com

B. Bar Chart:

Bar chart is as well referred to as column chart and they are used to for comparison of items of different groups. The bars are used to represent the various values of a group and the bar chart makes use of both horizontal bars and vertical bars [10]. When the values to be represented are clearly different and such differences in the bar are been seen by human eye, then one can decide to make use of a bar chart, but when there are very huge numbers of value to be displayed, then it might be a bit more hard to make comparison between the bars. Most times, bar chart is used to represent discrete data and it is as well used to present single data series while the data points that are related are often being grouped in a series.

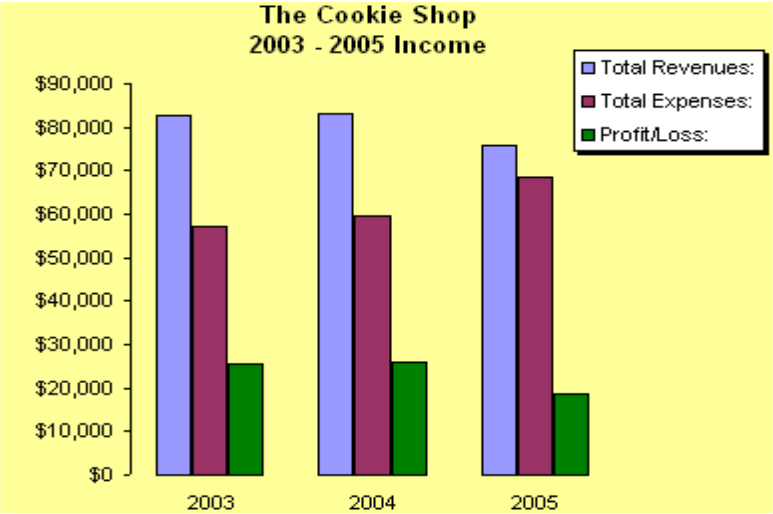

Fig.2. Displaying a simple Bar chart [21]

C. Pie Chart:

It is as well-known as a circle graph. A pie chart shows information statistics and data in a way that is not difficult to read called "pie-slice" form and the various sizes of slice shows how much of an element is in existence. When the slice is big, then it shows of the data was gathered. It is also used to compare values of data and the moment some values are represented on pie chart, then you will be able to view which of the items is the least popular or which is more popular [4]. The best and effective way to make use of a pie chart is when they contain a few components and when the percentages and texts are also involved in order to define the content. . By providing additional information, report consumers do not have to guess the meaning and value of each slice. If you choose to use a pie chart, the slices should be a percentage of the whole [10].

A wedge is used to represent a data parts that has the same characteristics and the pie chart control usually decides the data wedge size when it's being compared with the other data wedges. Pie charts consist of two popular variations called Doughnut chart and Exploding pie chart. The Doughnut chart are almost same as the standard pie chart just that it consist of hollow center and the exploding charts, the wedges are being obtained from the other wedges [28][11].

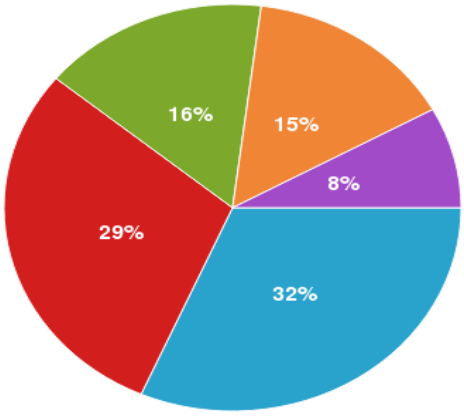

Fig.3. A simple Standard Pie chart [30]

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 3, pp: (105-113), Month: July - September 2016, Available at: www.researchpublish.com

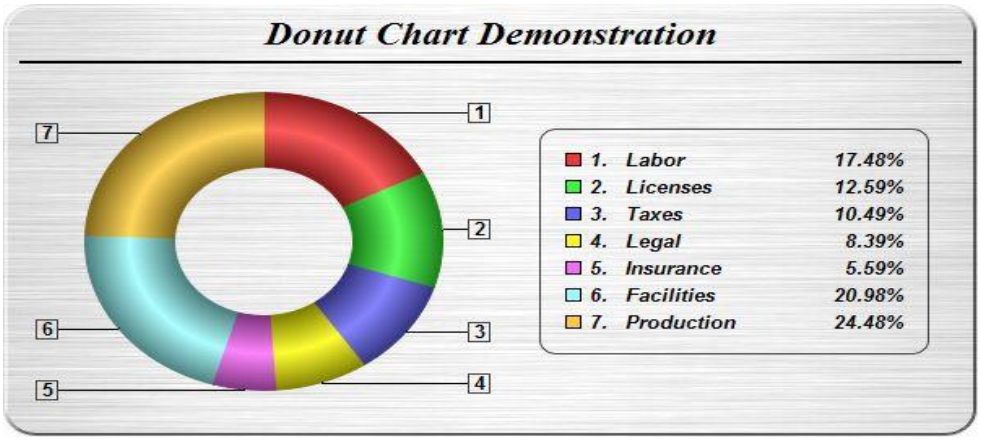

Fig.4. A simple Doughnut Pie chart [25]

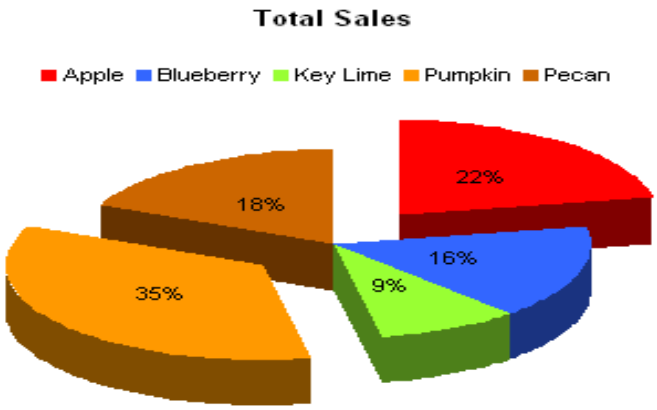

Fig.5. A simple exploding Pie chart [22]

D. Table:

Table is simply the arrangement of data using rows and column. In conducting research and analysis of data, the role of table is very important. Tables are simple to understand and analyze and it's simple to interpret the method of data representation. A row is a representation of variables and column is as well a representation of records that have set of values. . At times, this order of arrangement can be changed, i.e. rows could represent records and columns represent variables [9] [7].

Average annual expenditures for selected categories of consumer spending, by housing tenure, 1986 and 2010

Expenditure	Homeowners			Renters		
	1986	2010	Percent change	1986	2010	Percent change
Annual expenditures	$56,050	$55,780	0	$33,524	$33,460	0
Food	7,907	6,820	-14	5,169	4,802	-7
Food at home	4,580	4,000	-13	2,972	2,902	-2
Food away from home	3,327	2,820	-15	2,196	1,900	-13
Housing	16,637	18,503	11	11,038	12,843	16
Apparel and apparel services	3,030	1,781	-41	2,111	1,544	-27
Transportation	11,609	9,056	-22	6,408	5,046	-21
Gasoline and motor oil	2,147	2,458	15	1,285	1,511	18
Healthcare	2,837	4,016	42	1,313	1,518	16
Health insurance	943	2,314	145	406	909	124
Entertainment	2,775	3,088	11	1,486	1,390	-6
Personal insurance and pensions	5,368	6,665	24	2,374	2,907	22

NOTE: For comparison, 1986 dollars were converted to 2010 dollars.
SOURCE: U.S. Bureau of Labor Statistics.

Fig.6. Shows a simple table [26]

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 3, pp: (105-113), Month: July - September 2016, Available at: www.researchpublish.com

D. Scatter Plot:

A scatter plot is described as a 2-dimensional plot which displays the joint variation of two data items. A scatter plot is also called a scatter chart, scatter diagram, scatter graph. For a scatter plot, observations are being represented by each marker and the marker position usually indicate value for the observations. A scatter plot shows data in Cartesian coordinate in a graphical display which displays the relationship that exist between two variables in which one is represented as a vertical distance and the other as horizontal distance [15].

The moment you have all the data plotted on a scatter plot, you will now be capable of determining in a visual form if the data points are related or not. Scatter plots can help you know how the data points are scattered or spread across the graph and also you will know they are closely related [10]. A scatter plot will display the variables and how strong they are related and you can also know how far the data are scattered [17].

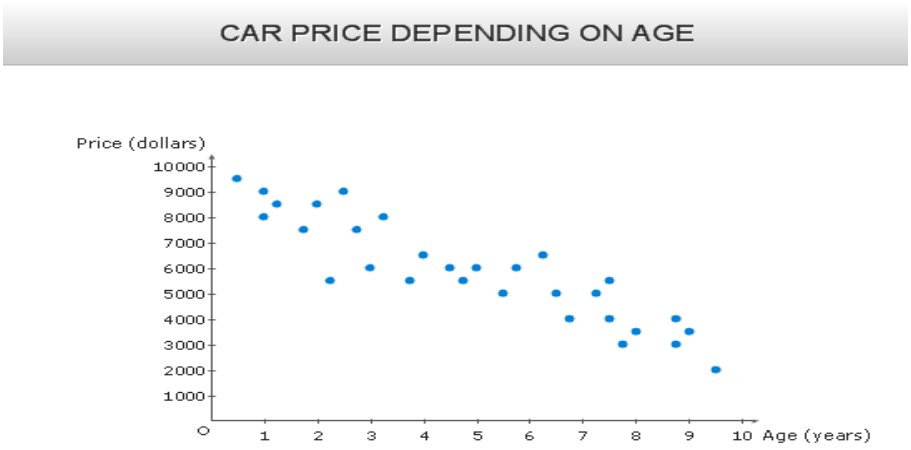

Fig.7. Basic scatter plot of two variables [27]

E. Bubble Chart:

A bubble plot is some degree of difference of a scatter plot and the markers in it are being substituted with bubbles and this is possible only we have a set of data points which has three values contained in each data item [24]. It shows the relationship that exists between the minimum of three variables. Two of them gets represented by the plot axes i.e. x-axis and y-axis, while the third one by the bubble size and each bubble is a representation of an observation.

Bubble plot is used with a lot of value, say hundreds of them or also used if the values are somewhat different by numerous structure of magnitude. Colors are being used to represent an additional measure and the bubbles could be subjected to animation in order to show data changes over a period of time.

The bubble plot is also very useful in project management in comparing the rate of risk and success involved in executing a project and where there are three values as net present values, then the probability of success and the total sum represent the bubble size [5].

Fig.8. Showing a simple bubble plot [19]

ISSN 2348-1196 (print)
International Journal of Computer Science and Information Technology Research ISSN 2348-120X (online)
Vol. 4, Issue 3, pp: (105-113), Month: July - September 2016, Available at: www.researchpublish.com

F. Parallel Coordinates:

The parallel coordinate technique makes use of the concept of networking a multi-dimensional point to some axes and all of these are parallel to each other. In these technique, single data elements are being plotted across many dimensions and these dimensions are connected unto a y-axis and each object of the data is shown along the axes as a series of connected points [18]. The parallel coordinate is important if you want to show a multidimensional data and a lot of these dimensions are being organized and expanded by this technique.

When there is a line that forms a single polygonal line for all the occurrences represented, then it connects the individual coordinate mappings. Therefore the number of dimensions that is being represented is not limited at all. This visualization technique is applicable in areas such as: computer vision, air traffic control, computational geometry, robotics and data mining [1].

A good advantage of this visualization technique is that it usually represents lots of dimensions without limits. Though, you can encounter a case such as the polygonal lines being overlapped which causes difficulty in identifying characteristics in the data and this caused when you have many points that are being represented when engaging the parallel coordinate approach [2].

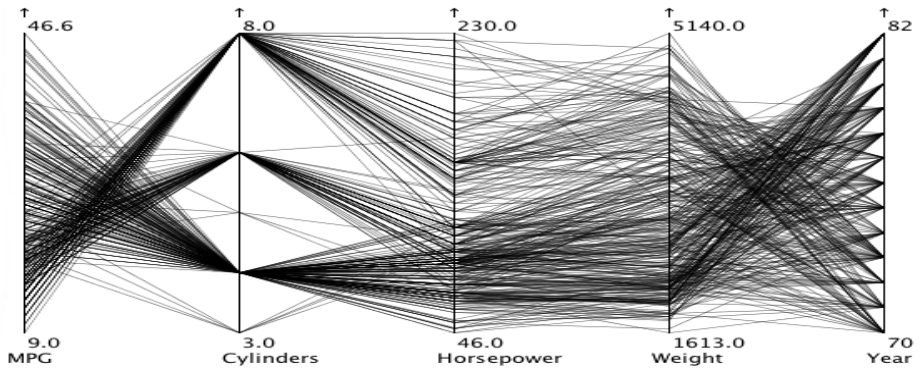

Fig.9. Showing a parallel coordinate [20]

G. Tree Map:

A tree map is a visualizing technique that has the attribute of showing data in hierarchy in a nested or layered rectangle form [12]. It is a very effective technique that is used to visualize structures of hierarchies. User are able to compare nodes and sub nodes at different depth and also they are able to identify expected results and patterns. A lot of data set have the hierarchy characteristics and the objects are thereby divided into different divisions, sub divisions, etc.

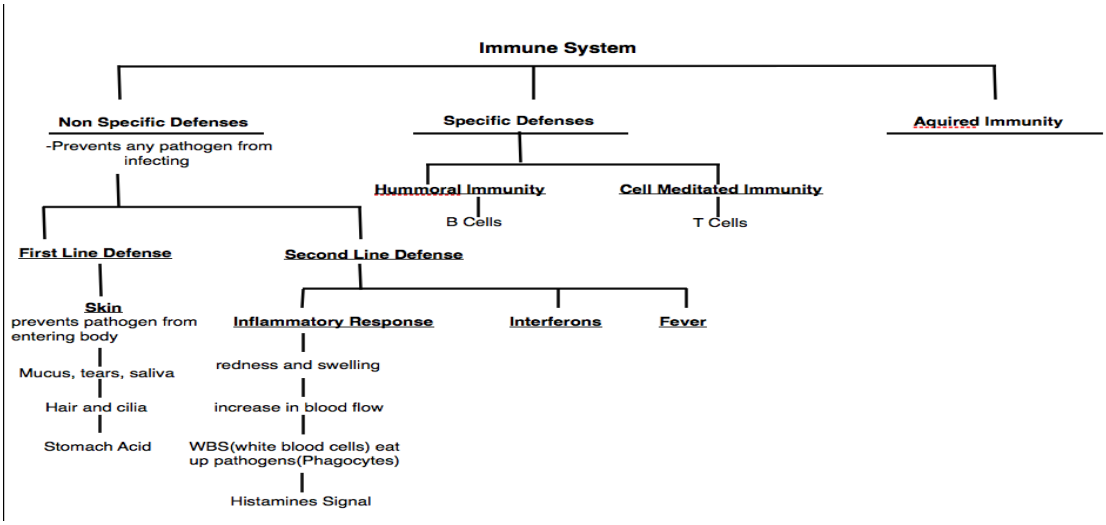

Fig.10. Tree map display hierarchical data [23]

These data visualization techniques that have been discussed are just a few of the several visualization techniques used for data mining. There are some general key features of visualization-based data discovery techniques, they are as follows [4]:

- Support online creation of dynamic, interactive presentations and reports.
- They allow end users to interact with data, often on mobile devices.
- They give room for real time data analysis.
- Gives room for users to share and collaborate securely.
- It holds data in-memory, where it is accessible to multiple users.

V. DISCUSSIONS

In this section, we will discuss our views and research contributions on the properties of the data visualization techniques, i.e. showing how adequate or inadequate the data visualization techniques can handle big data, and this will be represented by the table 1 below. Big data comprises of some characteristics that affects its operations, these characteristics are also known as the 3Vs of big data. They are: Volume of data, Variety of data and Velocity of data. Data volume is the quantity of data that is made available to a company or organization and they don't have to own all of the data as long as such data can be accessed. It majorly deals with the size of data and this can be measured in terabytes, petabytes, Exabyte, etc. Variety gives a measure of how rich the representation of data is and this can be in audio, videos, texts, images, etc. and they can be either semi structured, structured or unstructured. Velocity is the process by which the speed of data creation is measured and also the speed of streaming it and as well the speed of aggregating the data is being measured.

TABLE I: THE DATA VISUALIZATION TECHNIQUES PROPERTIES

TECHNIQUES	LINE CHART	BAR CHART	PIE CHART	TABLE	SCATTER PLOT	BUBBLE CHART	PARALLEL COORDINATE	TREEMAP
VOLUME	-	-	+	+	+	+	+	+
VARIETY	+	+	-	+	+	+	+	-
VELOCITY	+	+	+	+	+	+	+	-

VI. CONCLUSIONS

The main focus of this paper is to give a brief survey of some of the multi-dimensional visualization techniques that are used in data mining, knowing fully well that the techniques are not limited to the ones that have been discussed in this paper as there are much more to this. Data mining is a field of computer science that is fast emerging most especially in the business sector of our economy and lots of research is being done every day as pertaining this aspect and it is so unfortunate that are business sectors are still not getting the exact result they want when trying to visualize their data using any of the data visualization techniques. One main reason why this challenge is still persistent is because these techniques are being put to use wrongly, many of our business people still don't know what technique is best to use when they want to carry out a particular task and so they end up choosing the wrong visualization technique for the right data and they eventually end up getting wrong results. The use of the data visualization techniques used in data mining could be interesting and at times challenging as well, it all depends on how effective you put it to use but for you to be able to choose the best underlying visualization technique to display your data effectively, you must first of all understand the data want to visualize with its size and cardinality (the uniqueness of data value contained in a column), also you should determine what you are trying to visualize and the type of information to be communicated, also you suppose to have a good knowledge of your audience and understand how it processes the vital information and lastly, you should make use of a visual that carried the information in the best and easiest way to your audience or end users.

REFERENCES

[1] Inselberg, B. Dimsdale. *Parallel Coordinates: A Tool for Visualizing Multi-Dimensional Geometry*; Visualization '90, San Francisco, CA, 1990, pp 361-370.

[2] D. A. Keim, H. Kriegel. *Visualization Techniques for Mining Large Databases: A Comparison*; IEEE Transactions on Knowledge and Data Engineering, Special Issue on Data Mining; Vol. 8, No. 6, December 1996, pp 923-938.

[3] G. Burton, Andreas. "Experimental psychology", 1965, page 186

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

[5] Jeff Berman "Maximizing project value: defining, managing, and measuring for optimal return", AMACOM Div American Management Association, 2007. Page .63-64.

[6] J. *William Frawley, Gregory Piatetsky-Shapiro, and Christopher J. Matheus,* Knowledge Discovery in Databases: An Overview, AI Magazine Volume 13 Number 3, 1992.

[7] M. C. Oliveira, H. Levkowitz. "Visual Data Exploration to Visual Data Mining: A Survey", IEEE Transaction on Visualization and Computer Graphs 9(3), 2003, 378–394.

[8] 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.

[9] Robert Redpath, a Comparative Study of Visualization Techniques for Data Mining- a Thesis Submitted to the School of Computer Science and Software Engineering Monash University, 2000.

[10] SAS, Data Visualization Techniques: From Basics to Big Data with SAS® Visual Analytics, 2014.

[11] Stephen, Few. "Save the Pies for Dessert", Perceptual Edge Visual Business Intelligence Newsletter August 2007.

[12] Shneiderman, Ben, Plaisant, Catherine (2009). "Tree maps for space-constrained visualization of hierarchies".

[13] S. Card, J. MacKinlay and B. Shneiderman. "Readings in Information Visualization: Using Vision to Think". Morgan Kaufmann 1988.

[14] S. Kaisler et al. Big Data: Issues and Challenges Moving Forward, 46th Hawaii International Conference on System Sciences 2013.

[15] Utts, M. Jessica. "Seeing Through Statistics", 3rd Edition, Thomson Brooks/Cole, 2005, pp 166-167

[16] U. M. Fayyad. Et al. *Advances in Knowledge Discovery and Data Mining*; AAAI Press, Menlo Park, California, 1996.

[17] What is Scatter plot? www.psychwiki.com/wiki/What_is_a_scatterplot%3F

[18] Zach Gemignani, (2010). "Better Know a Visualization: Parallel Coordinates", www.juiceanalytics.com/writing /parallel-coordinates.

[19] http://code.tutsplus.com/tutorials/how-to-create-a-bubble-chart-in-flex--net-3542

[20] http://eagereyes.org/techniques/parallel-coordinates

[21] https://foxhugh.com/charts/describe-bar-charts/

[22] https://missom.wordpress.com/tag/exploded-donut-chart/

[23] http://p3sts1011.blogspot.my/2010/11/immune-system.html

[24] http://study.com/academy/lesson/what-is-a-pie-chart-definition-examples-quiz.html

[25] http://www.advsofteng.com/doc/cdcfdoc/donut.htm

[26] *http://www.bls.gov/opub/btn/volume-1/pdf/a-comparison-of-25-years-of-consumer-expenditures-by-homeowners-and-renters.pdf*

[27] http://www.conceptdraw.com/samples/business-charts-area-line-scatter

[28] http:// www.cs.vu.nl/eliens/multimedia/assets/flex3/datavis_flex3, Adobe Flex, Advanced Data Visualization Developer Guide, 2008.

[29] http://www.statcan.gc.ca/edu/power-pouvoir/ch9/line-lineaire/5214824-eng.html

[30] http://www.zingchart.com/docs/chart-types/pie-charts/