Visual statistic preparation and interactive graph mining:in R using packages iplots

Shahin Zafar
Dept. of computer sciences and Engineering Jamia Hamdard New Delhi
shahin_zafar2009@yahoo.com

Suraiya Parween
Asst. Prof. Dept. of computer sciences and engineering Jamia Hamdard, New Delhi
husainsuraiya@gmail.com

***Abstract*—Due to the high availability of visualization techniques it is so important to select the right one, by which researcher can convey the maximum benefits. We can say data visualization is the graphical representation of image data, which tells about the acquisition or knowledge of those image data. In the context of big data, visualization is one of the challenging problem faced by the researcher and analyst but the right and propitious selection of data for visualization is considerable for monitoring the experimental result this paper is using various terminology and techniques focusing about the visual statistic process and how visualization technique is implemented in R using packages "iplots" and there is also some pros. and cons. between different statistical tools and comparison.**

Keywords: data visualization, visual data preparation, interactive graphs and techniques (iplots)

I. INTRODUCTION

The goal of doing analytical research on the topic of visual statistic preparation and graphs mining which provides good understanding concept about the data visualization, visual statistic preparation and graphical data presentation and how the different visualization techniques work with packages in R [1].

Now information is very essential for generating and visualizing the desire result of data according to Berkeley university every year approx one Exabyte data is generated hence we are living in information age where everyone is generating every kind of data either audio, video, image, documentary data etc. at some place role of human is encountered such as for making large transaction of data and at some place it is automatically monitored by the sensor [2].

In general term data visualization help people to understand the importance of generated contextual data and which is done by using different techniques and tools. The main objectives of this paper is how data is refined from several different process and also describing about the R tools/programming language and their packages, R is programming convention which provide an environment for statistical computing and graphical distribution of data for new users understanding and visualizing the data in R a bit programming knowledge and concepts is required it is freely available on the internet with different version with different platforms, users or programmer can install any version according to their operating system, in it rich set of packages library are available for graphical mining and representation.

Example: arules(mining association rules and frequent itemset), arulesViz(visualizing association rules and frequent itemset) [3] [4]. iplots(interactive graphical representation) in Section III.

II. PROPOSED WORK

A. Visual statistic preparation:

1) Data analysis: for the purpose of visual statistic preparation the role of data analysis is very vital and important, before going any further steps the very first and initial steps is data analysis, in data analysis data is preprocessed where we eliminate some sort of errors (incorrect data, redundant data, missing values variables, incomplete data and inconsistence data etc.) from the data sets and hence it can be achieved in R [5] [6]

2) Data assortment: In this steps researcher have to choose the most appropriate and relevant data which is available according to their work or task which then minimize the risk of integrity issues once the data is assorted and processed it will filter out inconsistence or bad outlier's, and can be possible using different algorithm and packages in R [7]

3) Data depiction: When the assortment work is completed, data depiction takes place it is responsible for visualizing the chosen information or "how the encoded information is converted into visual form" shows that how a data from one model system depict to the data from another model system depiction can be applied in many ways using procedural codes, Extensible Style Sheet Language Transformation and other existing graphical interfaces, numerous potential issues (how data is stored, data lost or mis-matched) can be resolved by using data depiction

4) Data exploration and presentation: In these steps analyst commonly used data visualization tools and statistical software and application, how effectively available information is managed, organized and summarized where it facilitates user's to quickly understand relevant feature of their data set.

5) End-user factors :

- Human factors : when group of usability are performing research to build standards and user-centered measurement
- Human system intention: can enhance by using the usability and accessibility principal &user-design.

6) Effectiveness measure: it will measure system performance in terms of "speed, payload, range, time, and frequencies" and another measures suitability. [1] [8]

B. R introduction:

R is a statistic tools and software which facilitates data analysis, manipulation, calculation and graphical representation on wide range it is used by the programmer, data analyst and statisticians an effective and efficient data handling and storage facility it has for different platforms (like windows, Linux, Mac) different set up is available to compile and run the software, to download R we have to select the most preferred CRAN (comprehensive R archive network) mirror which means choose a location which is nearest to you [9].

It is used as GNU *(g, noo) project* which has rich set of application, libraries, tools and games. In R mostly the work is easily accessible by using packages, a package is bundle of codes/commands together and rich set of libraries is implemented. In R there is more than six thousand packages is available on CRAN by their name or by their date of issuing now the huge verity of packages is available and this is the reason why R is successful and on demand.

1) R vs SAS : R is statistical way of computation with full access of packages what make R's unique? there's no simple answer of this question consider an example of windows and Linux while windows always dominant because it's user friendly nature and provide ease of access where as Linux provide better security and virus protection but third party/user still prefer windows likewise r and sas. Some pros. and cons. R is much flexible, low computation power, open source where as SAS is paid software, syntax in sas support high level language where as data handling process is hazardous in R it loads all the data set or data table in RAM at one go by which data handling and memory allocation process deemed. Sas handle large data set process in few minutes but the same process R will take time more than few minutes for new users it is bit difficult to learn to learn R and throw many out of memory errors resolving these online help is available.

2) R vs D3.js: Interactive graphics is done by the packages in R not all but some packages required rjava means jdk environment to run their packages successfully like iplots which is used in (section III) and in other hand data driven documents need full support of JavaScript to run applications data visualization and all related process such as summarisation, analysis, assortment etc. can be done easily because it has various graphical representation techniques but for d3 user's need to learn JavaScript and then do so, d3 work as driver for making connection between data and documents and provide quality by using JavaScript. For static graphs mining R is suitable but for interactive plotting package required where as d3 done both statistic interactive and dynamic graphs at one go [10].

III. EXPERIMENTAL WORK

A. Model interface of iplots:

Iplots is a packages which is used for the R statistical environment and provide high interactive statistical graphical interface. To use iplots we have to first load the packages in Rstudio IDE, the iplots packages will get loaded and it will also load the other required packages automatically

*B. Preparation of data for the packages **iplots:***

```
> library(iplots)
```

loading required package: rJava now the packages **iplots get** uploaded into the console

```
> data (iris)
```

Data iris is uploaded into the global environment in Rstudio, data iris is default data set inside the iplots packages.

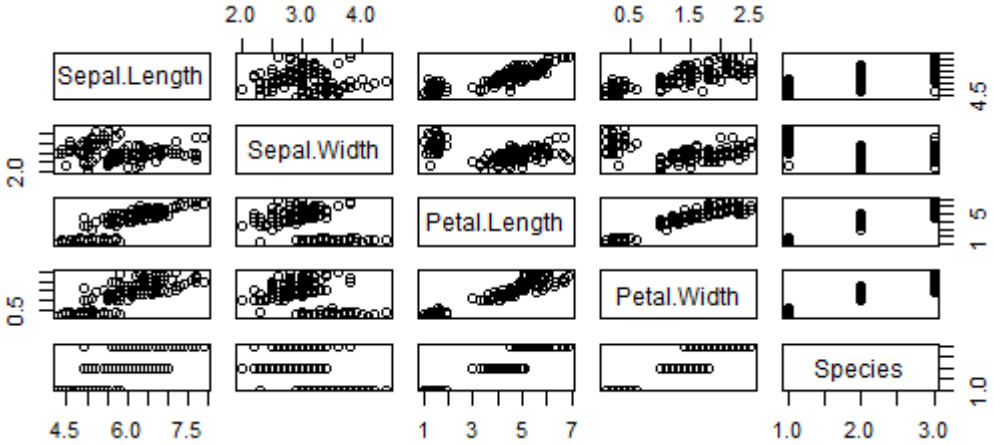
```
> summary (iris)
Sepal.Length    Sepal.Width     Petal.Length    Petal.Width          Species
Min.   :4.300   Min.   :2.000   Min.   :1.000   Min.   :0.100   setosa     :50
1st Qu.:5.100   1st Qu.:2.800   1st Qu.:1.600   1st Qu.:0.300   versicolor:50
Median :5.800   Median :3.000   Median :4.350   Median :1.300   virginica:50
Mean   :5.843   Mean   :3.057   Mean   :3.758   Mean   :1.199
```

3rd Qu.:6.400 3rd Qu.:3.300 3rd Qu.:5.100 3rd Qu.:1.800

Max. :7.900 Max. :4.400 Max. :6.900 Max. :2.500

Iris description: The famous (Fisher's or Andreson's) iris data set gives the measurements in centimeters of the variable sepal length and width, petal length and width, respectively, for 50 flowers from each of 3species of iris, the species *iris setosa, versicolor and virginica*

> plot (iris)



The default variable "x" consists of 101 cases, but the current iSet consist of 150 cases, now we have additional provision to create a new iSet.

1) ibar(interactive bar)

Example:

> data ("iris")

> attach (iris)

> ibar(Species)

ID: 1 Name: "Barchart (Species)"

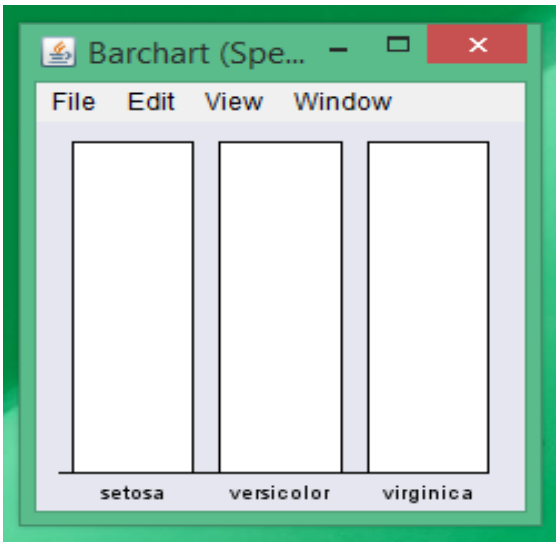


Fig.1

2) Ihist(Interactive histogram):

Histogram is used to represent the frequency distribution between data points

Ihist description: the function ihist is used to create new interactive histogram from given data points

ihist(iris$Sepal.Width)

ID: 2 Name: "Histogram (iris$Sepal.Width)"

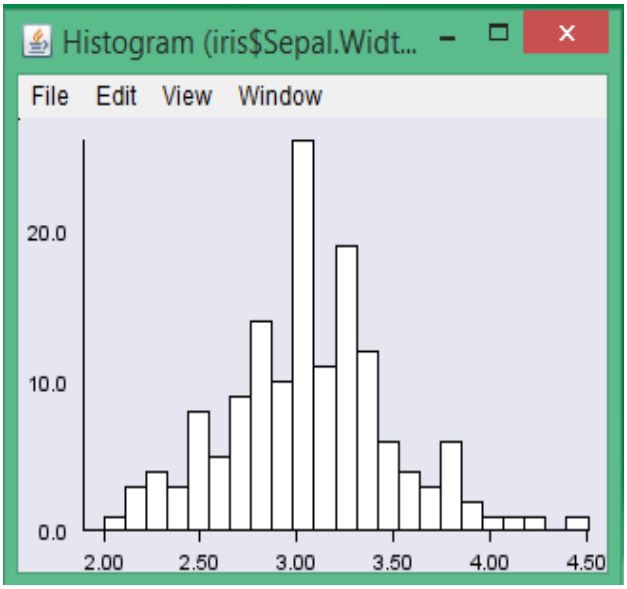


Fig.2

3) Ibox(Interactive box plot):

Interactive box plots (ibox) is useful for visualizing group of data patterns for large groups it provide an efficient way to visualize the range and other related characteristic and box plot is very helpful to explore outliers

> ibox(Sepal.Length,Species)

ID: 3Name: "Box plot (Sepal.Length) by Species"

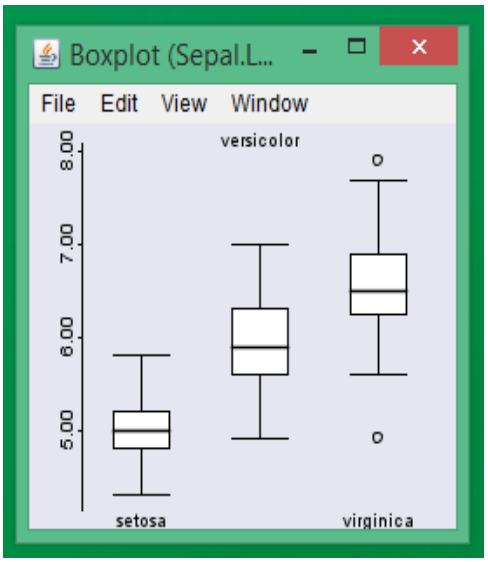


Fig.3

4) iplot(scattere plot):

> iplot(Sepal.Width/Sepal.Length, Species)

ID: 4 Name: "Scatterplot (Species vs. Sepal.Width/Sepal.Length)"

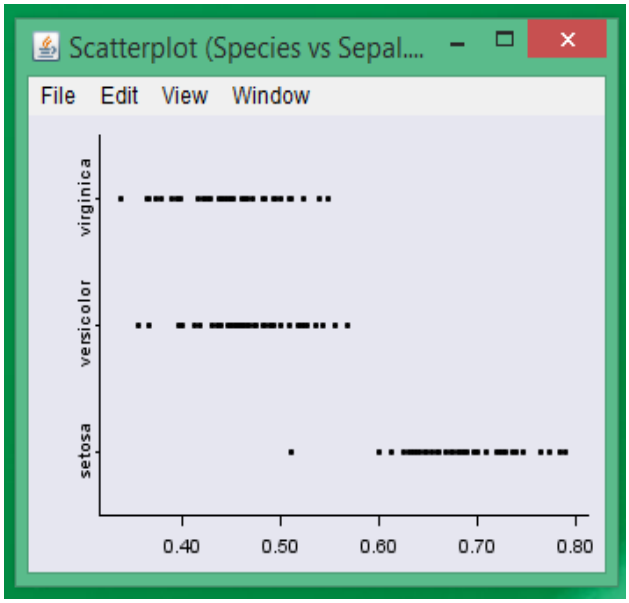


Fig.4

Interactive plot in general it is used to show how the relationship between two variable is strong, there are many additional parameter is used while creating iplot inside packages iplots.

Example of additional parameters:

spaceprop – spacing between data points.

1.0 means no spacing

1.5 means half of actual space of data points

drawAxes - to drawn an Axes

equiscale – on both axes same scale is drawn

ptDiam – used to apply Point diameter

minimalDiam - used for Minimum point diameter etc.

5) ipcp(interactive parallel coordinate plot)

> data("iris")

> ipcp(iris)

ID: 5Name: "Parallel coord. plot (default)" for data iris (species) a default ipcp is created.

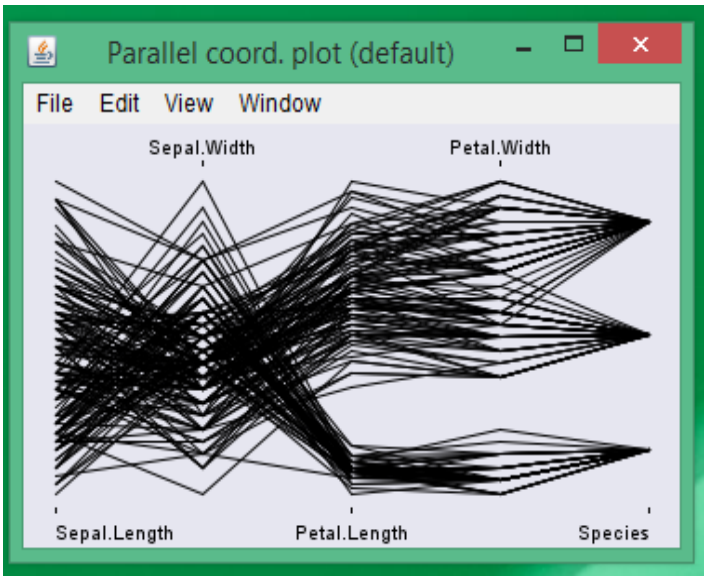


Fig.5

C. Operation performed on these visualization techniques:

When you select one attribute in your graphs, it will automatically highlight the clustered data points in another different graphs, or automatic transmission done between plots while selecting points.

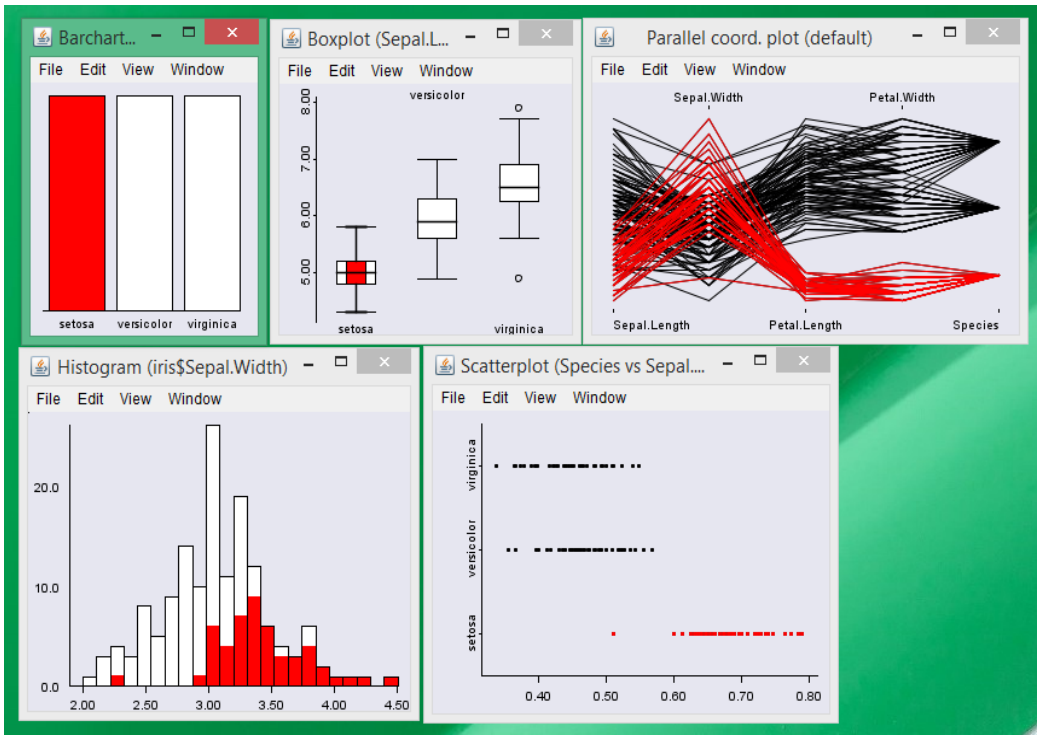


Fig.6

And in the next graphs we can see the results for outliers while selecting one outliers in one chart it will highlight all outliers in different plots readers can see the results in fig.7

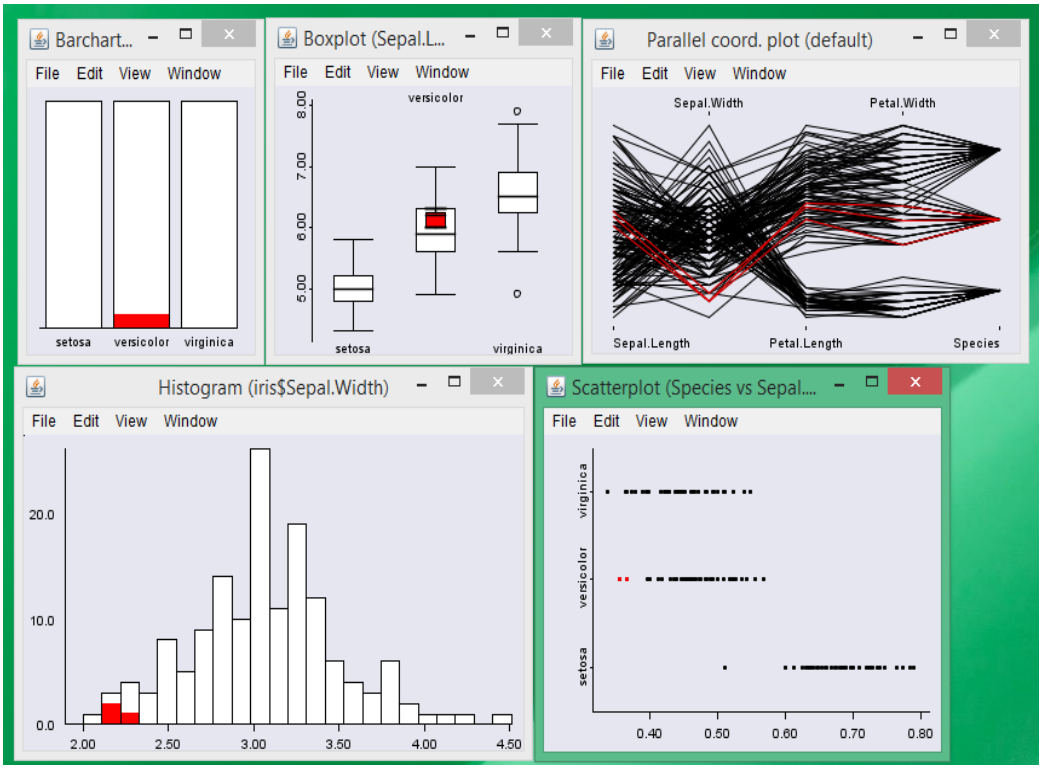


Fig.7

Various different features are available for these visualization techniques see the Table. I below

1) Comparision Analysis between plots:

Table I.

Techniques →	Ibar	Ibox	Ihist	Iplot(scatter plot)	Ipcp
Rotate	✓	✓	✓	✓	✓
Reset zoom	×	✓	×	✓	✓
Setcolor (CB)	✓	×	✓	×	×
Setcolors Rainbow	✓	×	✓	×	×
Spineplots	✓	×	×	×	×
Sortby highlighting	✓	✓	×	×	×
Spinogram	×	×	✓	×	×
Transparent highlighting	×	✓	×	✓	✓
PCP	×	×	×	×	✓
box plot		✓			
Pcp over boxex		×			
Large points up	×	×	×	✓	×
Small_points down				✓	
Paint selection	✓	✓	✓	✓	✓

IV. CONCLUSION:

In this paper I have presented visual statistic preparation via different steps such as data analysis, data assortment, data depiction, Data exploration and presentation, end-user factors and effectiveness measure and how the processed visual data is graphically represented to form meaningful image data and which is performed by using different techniques in R tools and I did comparison analysis between these techniques now it's found R is much easy and convenient for plotting the interactive graphs using iplots.

Future work will focus on plotting the data in "ggplot2" it has theme system for brushing and polishing plot much flexible it will also provides to the users and researcher full graphics system.

ACKNOWLEDGMENT

I am very thankful to almighty who have given me patience capability and encouragement to complete my work, after that I am thankful to Tabrez Nafees and Syed Ali Mehdi (Assistance pro. of Jamia hamdard) for their kind support and guidance.

REFRENCES

[1] M. a. S. S. K. Khan, ""Data and information visualization methods, and interactive mechanisms: A survey."," International Journal of Computer Applications , vol. 34.1, pp. 1-14., (2011):.
[2] D. A. Keim, " "Information visualization and visual data mining," " IEEE transactions on Visualization and Computer Graphics 8.1 , pp. 1-8., (2002.
[3] M. a. S. C. Hahsler, ""Visualizing association rules: Introduction to the R-extension package arulesViz."," R project module, no. R project module (2011): 223-238., pp. 223-238., (2011):.
[4] M. e. a. Hahsler, ""The arules R-package ecosystem: analyzing interesting patterns from large transaction data sets."," Journal of Machine Learning Research :, pp. 2021-2025., 12.Jun (2011).
[5] M. a. P. F. Templ, ""Visualization of missing values using the R-package VIM."," Reserach report cs-2008-1, Department of Statistics and Probability Therory, Vienna University of Technology., 2008.
[6] M. M. a. A. A. J. Hamad, ". "An enhanced technique to clean data in the data warehouse." Developments in E-systems Engineering (DeSE)," 2011. IEEE,., 2011.
[7] D. A. S. a. D. P. Khachane, ""Data selection and filtration technique for tuning virtual sensor model of NOx estimation in MATLAB." Computing Communication Control and automation (ICCUBEA),," International Conference on. IEEE,2016, 2016.
[8] J. J. C. S. a. B. B. Hilda, " "A review on the development of big data analytics and effective data visualization techniques in the context of massive and multidimensional data."," Indian Journal of Science and Technology, vol. 9.27 , 2016.
[9] [Online]. Available: https://ftp.iitm.ac.in/cran/. [Accessed february 2017].
[10] L. a. H. Z. Chen, ""Research and application of dynamic and interactive data visualization based on D3." Audio, Language and Image Processing (ICALIP),," International Conference on. IEEE, 2016., 2016.
[11] R. Kohavi, " "Data mining and visualization."," in Sixth Annual Symposium on Frontiers of Engineering. National Academy Press, , DC, 2001.
[12] H. e. a. Herodotou, ""Starfish: A Self-tuning System for Big Data Analytics."," Cidr. . ., vol. Vol. 11. No, no. 2011, 2011..
[13] M. a. S. C. Hahsler, ""Visualizing association rules in hierarchical groups." 42nd Symposium on the Interface: Statistical, Machine Learning, and Visualization Algorithms (Interface)," in The Interface Foundation, of North America.20, 2011..
[14] U. G. P.-S. a. P. S. Fayyad, ""From data mining to knowledge discovery in databases."," AI magazine, vol. 17.3, p. 37., (1996).
[15] R. S. e. a. Raghav, ""A survey of data visualization tools for analyzing large volume of data in big data platform." Communication and Electronics Systems (ICCES),," International Conference on. IEEE, , 2016.
[16] M. N. S. S. a. V. F. L. Moreno, " "Association Rules: Problems, solutions and new applications."," Actas del III Taller Nacional de Mineria de Datos y Aprendizaje , Vols. 317-323., (2005):.
[17] P. R. e. a. ". U. o. I. V. T. (. Luzzardi, ""Evaluating Usability of Information Visualization Techniques."," 2002.