ISSN 2229-5518

IBA Graph Selector Algorithm for Big Data Visualization using Defence Dataset

Madhu Sudhan S, Chandra J

Abstract— Data visualization is a technology generally used for better understanding of the data and relationships by representing the data in the form of graphs. Most of the business intelligence software's are embedded with data visualization tools. Other than business domains, many other domains like army, hospitals etc. are using data visualization. Defence Services is also one of the domains where the visualization techniques can be used. Any updates in the database can be viewed in the form of graphs which may help in decision making at critical situation and war planning. The main aim of the research work is to visualize the dataset efficiently by building a cross platform application, which helps the user in choosing the right graph, using the proposed Input Based Analyzer (IBA) graph selector algorithm.

Index Terms—Big Data Mapping, Choosing a Right graph, Data visualization, Data analysis, Graph selector algorithm, Graphs, Input Based Analyser.

1 INTRODUCTION

DATA Visualization is the process of visually representing the data in the form of graph, for the better understanding of the data and seeking information, which helps in decision making. Primary objective in data visualization is to gain insight into the information space [1].
Mapping the big data and visualizing it, is a difficult process. Big data is the collection of huge amount of data which are of different data type's i.e. ordinal, nominal, ratio etc. Most of the database maintained by organization and companies are huge. Defence Services is one of the domains which maintains huge amount of data which may be image, text etc.

In this paper, we are proposing a framework and Input Based Analyser (IBA) graph selector algorithm for visualizing dataset by guiding the user to choose the right graph. Based on observation made on few frameworks and as per the requirements of application, a new framework has been proposed and implemented.

2 METHODOLOGY

The main aim of proposing a framework is to build a cross platform application for visualizing the big data with appropriate graphs. Proposed framework is based on few frameworks and architecture [2],[8],[12].one of the framework is aimed on integration of diverse data types on Xmdv Tool [2],[3] and other framework is on visualizing and integration of database using self-organizing map.

- *Madhusudhan S is currently pursuing master degree program in computer science in Christ University, India, PH-09535085413.E-mail: madhusudhan.chokure@gmail.com/Madhusudhan@cs.Christuniversity.in*
- *Chandra J is currently working as Assistant Professor Computer Science department in Christ University, India, PH-09886306307. E-mail: chandra.j@ christuniversity.in*

Aim of IBA graph selector algorithm we are proposing is to help user in choosing right graph depending on the dataset and data analysis type selected by the user.

2.1 Observation Made On Framework

Referred framework [2], is a combination of two mapping module, Visual mapping and Data mapping. The main aim of the framework is to introduce another mapping process, parallel to the visual mapping process, to include the semantics of data within the pipeline of visualization tool [2].

Focused on our research, few observation were made on the referred framework [2], they are.

- Every mixed data is getting converted to nominal, which may not be required in few instances.
- Framework is not concentrating on visualization techniques, types of visualisation depending on data.
- If dataset is big and confidential, then application should be secure and more interactive.

2.2 Proposed Framework

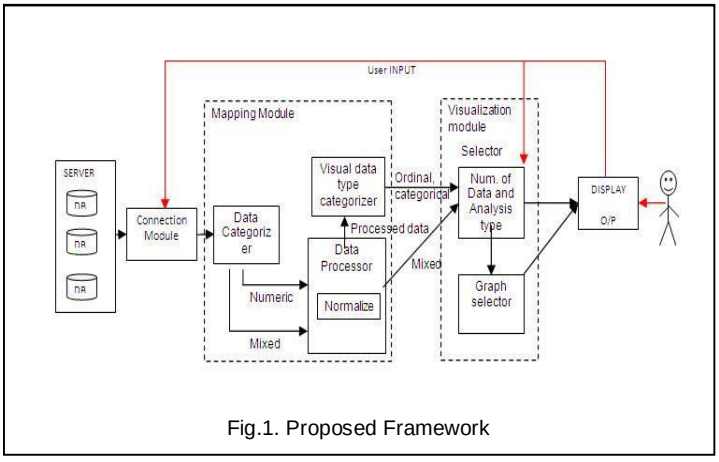

Fig.1. Proposed Framework

Proposed framework [Fig.1] is combination of three modules, they are connection, mapping and visualization module.

Connection Module

The aim of this module is to make the application a cross platform, which will work on all kind of database provider (MYSQL, SQL, PostgreSQL etc.).

User needs to know the IP address of the server to fetch the dataset. When the IP address is given to the application, it checks for valid input and then checks for the ports which server is responding. After checking, the application returns back the database providers which are active on the server.

Once the connection is established to the port, the user selects the database and the corresponding tables are fetched from the database.

Mapping Module

Mapping the data to appropriate data type is the main aim of this module. When data is big, mapping basic data type to visualization data type, i.e. categorical, ordinal, interval and ratio is a tricky task.

Mapping is required, as finding distance, mean between two nominal data is meaningless. Mapping nominal variable into numeric should be implemented by substituting numeric values to the equivalent nominal data. The numeric substitution should be meaningful, i.e. the user should be able to find distance between two data using these numeric data.

Data Categorizer

The categorizer accepts the data from data base and divides the data into nominal and numeric. The output of the data preprocessor will be mixed and nominal data. This output is then fed into data processor.

Data Processor

This block is for processing the nominal and numeric data extracted from the data categorizer. Normalization of data can also be done if needed to selected the needed data and put them into one format.

Visual data type Categorizer

The data in the data base are of integer, character etc. they don't have any distance, order or classing. After processing the data, the present data types are converted into visualization data types i.e. Categorical, Ordinal, Interval, etc. to make the process of visualization easier.

To convert the nominal or mixed data to numeric, Existing DQC model [4] can be used. First step is Distance step which identify the independent dimension to calculate the distance between the nominal variable. To calculate distance or dissimilarity of data is done using Euclidean, Minkowski or Supermum distance formula [5]. Second step is Quantification step. Using the distance information and independent dimension, order and spacing is assigned among the nominal variable. Classing is the final step. The Previous two step results are used in this step to find the similarity and make grouping.

Mapping module is not implemented in our application, as dataset was already processed and normalized and mapping was not needed.

Visualization Module

Visualizing the mapped data is not an easy task, because graph generated is dependent on analysis type and number of data to be visualized.

When data values are more, and a pie chart is used to visualize it, the chart will be overfilled and least informative.

TABLE 1
DATA ANALYSIS TYPE

	Comparison	Distribution	Composition	Trends	Relationship	Table
Line	*			*		
Bar	*	*	*			
Stacked Bar	*		*			
Bullet Bar	*					
Column	*	*		*		
Stacked Column			*	*		
Pie chart			*			
Pie with highlights			*			
Scatter plot					*	
Bubble	*				*	
Scatter column volume			*	*		
Scatter column volume with total			*	*		
Two axis column line				*	*	
Water fall			*			
Alternative row table			*		*	*
Quartile table			*		*	*
Grouping table			*		*	*

Comparison, distribution, composition, trends, relationships and tables are different data analysis type.

Analysis Type and data selector

This module accepts number of data need to be visualized and data analysis type as the input from the user who wants visualize the data. To use our visualization application, the user should have knowledge about what type of analysis to be done on the data, i.e. Comparison, Distribution, Composition, Trends, and Relationship etc., [6].

The above table [TABLE 1] shows the graphs which are suitable for different data analysis type. The existing flowchart [9], gave us few ideas about data analysis type and graphs associated with is. Using these ideas we came out with a tabular representation of data analysis type.

Graph selector

Choosing a right graph for visualization plays a very important role in knowledge gain process of visualization [6], [10]. Visualizing a dataset without any knowledge of what we want to gain from the process is meaningless. Proposed IBA graph selector algorithm helps the user in choosing the right graph for visualization depending on the number of data and

analysis type selected. The table given below shows the number of dimension, limitation of data value and analysis type for the corresponding graph [TABLE 2].

The graph selector table [TABLE 2] is based on flowchart in which they talk about choosing a right graph depending on analysis type and limitations of graph [9], [6].

Before selecting the graph to visualize, the user should know the following [9].

- Number of dimension to be visualized.
- Data analysis type.

TABLE 2
GRAPH SELECTOR TABLE

Visualization type	Data Dimension	Maximum Data value	Data type
Pie	1	10	Categorical
Bar	1	50	Categorical
Line	1	50	Ordinal , Interval
Stacked Pie	2	10 times 5	Categorical
Stacked Bar	2	50 times 5	Categorical
Stacked Line	2	50 times 10	Ordinal , Interval
Histogram	1	50	Ordinal , Continues
Box Plot	2	10	Continues , Categorical
Scatter Plot	2 or 3	Thousands for each dimension	Continues
Parallel	N	Thousands for each dimensions up to 20D	Any
Link Graph	2 0r 3	100	Any
Map	1	100	Any
Tree map	N	10000	Categorical , Any

- Aim of visualizing the data.

2.4 Proposed Algorithm

Earlier, few papers talk about automatic generation of graph from a dataset in a tabular format and algorithm for grouping dimensions [7].

Input Based Analyzer graph selector algorithm is the proposed algorithm based on graph selector table [TABLE 2] and analysis type table [TABLE 1].This algorithm takes Data and Analysis type [TABLE 1] as input and returns the graph appropriate for the visualizing the data.

As shown the proposed IBA graph selector algorithm starts with a condition. First condition checks if data is null, i.e. if user don't have any data in the database, then algorithm will not return anything. Once the data is not null, the second condition is checked to see what type of analysis the user want to make, if these two conditions are satisfied then the third and final condition is checked to know the number of dimension user wan to visualize. After all these conditions are satisfied, corresponding graphs are returned, from which user can select and visualize their data.

If these conditions are not satisfied then it will not return any graph.

Example:

If user selects analysis type as composition and selects only one dimension, then the user can choose Pie-chart [Fig. 3]. If user selects analysis type as comparison, Bar chart and area chart can be viewed [Fig. 4].

Procedure IBA_Graph_ selector (Var D, DA, DIM: Input);
Input:
D: number of data to be selected
DA: data analysis type to be selected
DIM: number of dimension to be selected

Output:
GRPHS: Required graphs

Begin

Step-1: Check if [D satisfy the number of max or min data]
a) Check if [DA is particular analysis type] then
Check if [DIM satisfy the number of max or min dimension] then
Return GRPHS
b) Otherwise return nothing.

Step-2: Repeat Step-1 Until (D != null)

End;

3 RESULTS AND DISCUSSION

As a first step, the user has to enter valid Identification number.

Once the user is successfully logged in, the user has to enter the IP address of the server for fetching the data [Fig.2]. The application checks for the active ports in the server.

The user needs to select the login details for database provider in which the database is fetched. According to the selected database and table, the columns are fetched from the database.

As the final step, the user has to select the column and the data analysis type to IBA algorithm, and the required graphs are enabled in dashboard.

In Fig.3, the user has selected one dimension and analysis type as composition, thus the IBA graph selector algorithm suggests pie-chart to visualize.

Fig.2. Page for selecting Server, data provider, database and table.

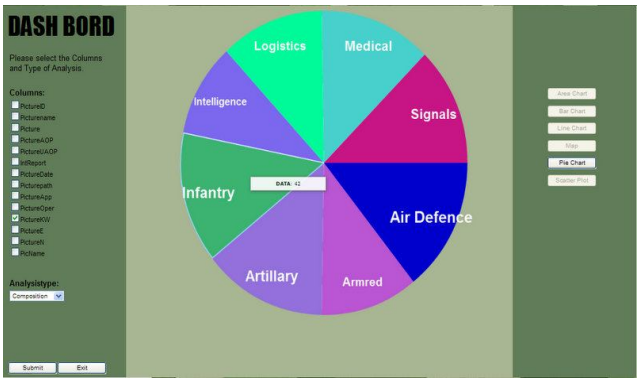

Fig.3. Page for selecting dimension, analysis type and right graph

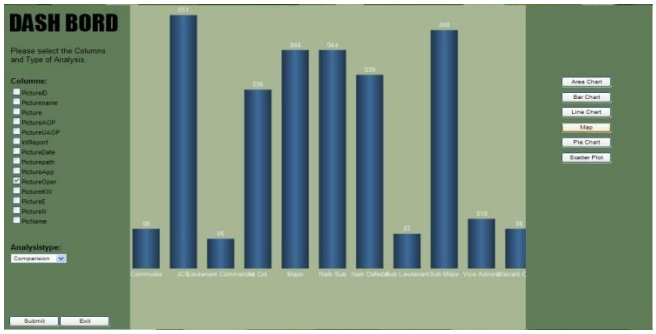

Fig.4. Comparison Bar chart

4 IMPLEMENTATION

The implementation of the application has been done in HTML, JavaScript and PHP. JavaScript Infovis Toolkit (JIT) has been used for visualization. Once the framework was ready and sufficient survey was done on development of application, it took 6 – 7 man weeks' time to build an application by integrating all the modules.

CONCLUSION

The main contribution of this paper is to develop and implement a framework for visualizing big data. We have proposed IBA graph selector algorithm, which guides the user to select the right graph for visualizing the big data.

Framework consists of three modules which mainly focus on connection, mapping and visualization. The main aim of the connection module is to make the application cross platform, i.e. application should work for all the data base providers. Mapping module concentrates on mapping nominal to numeric data types and visualization module implements the proposed IBA graph selector algorithm which helps the user in choosing right graph depending on the analysis type and data input given. By choosing the right graph, user can gain more knowledge as the graph will be more informative.

The mapping module can be included as a future work in the application.

ACKNOWLEDGMENT

We thank Dr.(Fr.)Jossy P George (Director of IT services, Christ University), Prof. Joy Paulose (H.O.D of Computer Science department, Christ University) and Ms. Rajeshwari C.N (Coordinator of MCA and MSc, Christ University), for their constant support during the time of research work.

We would also like to thank Director of Center for Artificial Intelligence and Robotics (CAIR), Bangalore, Ms. Sangeeta Shrivastava (Scientist 'E', CAIR, DRDO), Mr. Imran Syed (Scientist 'B', CAIR, DRDO) and Mr. Alekh Jain (MTech, DAVV, Indore) for their substantial contribution, guidance, encouragement and valuable suggestions during the course of the project work. We thank all members of CV Group (CAIR, DRDO), for providing required infrastructure during our research work.

REFERENCES

[1] G.S.Owen, "Definitions and Rationale for Visualization," ww.siggraph.org/education/materials/HyperVis/visgoals/visgoal2.htm. 1999.

[2] AnilkumarPatro, Matthew O. Ward, andElke A. Rundensteiner,"Seamless Integration of Diverse Data types into Exploratory Visualization Systems," *EUROGRAPHICS 2003 / P. Brunet and D. Fellner.*, vol. 22,pp. 3-7, 2003.

[3] Matthew O. Ward, "XmdvTool: integration of multiple methods ofvisualizing multivariate data," *IEEE*,pp. 326-333 Oct.1994.

[4] Rosario G. E, Rundensteiner E. A, Brown D. C, and Ward M. O, "Mapping NominalValues to Numbers for Effective Visualization," *IEEE, Oct.2003.*

[5] Jiawei Han, MichelineKamber, and JianPei,*Data Mining Concepts and Techniques,* 3rd ed., pp.72-74.

[6] Gina Trapani, *How to choose the Best Chart for your Data,*http://lifehacker.com/5909501/how-to-choose-the-best-chart-for-your-data archive/macros/latex/contrib/supported/IEEEtran/

[7] Fr´Ed´Eric Gilbert, David Auber,"From Database to Graph Visualization,"*IEEE*, Apr.2003.

[8] FaridBaurennani, Ken Q. Pu, and Ying Zhu, "Visualization and Integration of Database using Self-Organizing Map," *IEEE,* Fig. 1. Pre-processing phase,pp. 2, 2009.

[9] Raffael Marty, *Applied Security Visualization,* Addison-wsley publication, pp. 109-115, 2009.

[10] Jean Luc Doumont, Philippe Vandenbroeck," Choosing the Right Graph", *IEEE,* vol 45, pp. 5, Mrc 2002.

[11] NicolasGarcia Belmonte, "JavaScript InfoVis Toolkit," http://philogb.github.com/jit/.2013.

[12] Mikael Jern, "3D Data Visualization on the Web", *IEEE,*pp. 92, 1998.