International Journal of Current Engineering and Technology
E-ISSN 2277 – 4106, P-ISSN 2347 - 5161

Available at http://inpressco.com/category/ijcet

Research Article

DIVA: A tool for Data Integration, Visualization and Analysis

Shikha Duggal[Å*], Jyotsna Lilani[Å*], Sudhir Zanje[Å] and B.B. Gite[Å]

[Å]Department of Computer Engineering, Savitribai Phule Pune University, Pune, Maharashtra, India

Accepted 05 Nov 2014, Available online 01 Dec 2014, **Vol.4, No.6 (Dec 2014)**

Abstract

The world is becoming increasingly data-driven and we are in the midst of a data revolution. Data analysis is at the heart of decision-making. There is a need to simplify the process of extracting useful information from data in a simple and intuitive manner. DIVA is a tool for Data Integration, Visualization and Analysis. The tool enables users to visualize data in various forms such as charts and graphs. It supports integration and fusion by performing joins across various tables. It also supports the filtering and aggregation of data. Our aim is to offer a tool that makes data management and analysis easier. Our target audience does not necessarily have any training in using database systems.

Keywords*: Data Integration, Data Visualization, Data Analysis, R programming language, Merging, Fusion, Correlation, Regression, Statistical Analysis.*

1. Introduction

The main purpose of Data Integration, Visualization and Analysis (DIVA) is to provide a single tool to carry out integration, visualization and analysis.

- Integration- Merging and joining of data spread over different tables having common attributes.
- Visualization -Graphical representations of the data stored in tables in the form of bar graphs, histograms, pie-charts, face-cards etc.
- Analysis-Filtering data and Statistical Analysis through correlation and regression analysis which would measure dependency between attributes.

The product is a different take on Google Fusion tables. It is intended for users storing small data-sets in relational database management systems .It is not intended for big data. For example, an organization could keep track of the varied skill sets of their employees for performance analysis, internal job posting, determining the appropriateness of deploying an employee to a particular project depending on his experience in various domains and its correlation etc. Similarly, a teacher could check dependency between marks of different subjects of a student.

Given our target audience, we have converged the objectives that govern the design of the tool. *First*, the tool makes data management and analysis easier. It provides the three aspects of integration, visualization and analysis together in a single tool. This reduces the need to use individual tools for the three different aspects. *Second*, to open the threshold to the statistical power of R. R is the most popular language for statistical analysis and widely used by statisticians and data scientists all over the world.

*Corresponding author: **Shikha Duggal**

2. System Overview

Spreadsheets have been one of the most popular applications for storage and manipulation of data. While spreadsheets are good for storing data in tables, it becomes highly inefficient if more than a few hundred records need to be stored. Spreadsheets provide filtering and analysis as well. However, only one table can be stored in one spreadsheet. When one or more tables need to be merged, merging of tables spread over different spreadsheets becomes a long and tedious process. The main reason is that the data present in different spreadsheets is not present at a central location. Also, visualization and filtering results are shown within the same sheet, making it difficult to simultaneously view the data and its graphical representation.

On the other hand, databases provide a centralized format to store data. All tables can be stored in a single database, as compared to different tables spread over different spreadsheets. Also, tables containing more than a few hundred records can be easily stored and managed. Tables can be queried by writing simple commands in languages such as Structured Query Language (SQL). Merging of tables can be easily performed through SQL joins on tables. Filtering is performed by Select queries. Databases do not inherently provide visualization capabilities. Highly sophisticated visualization tools are developed to do the same (Ying Zhu, 2012).Also, data analysis in SQL is complicated.

R is a free programming language for statistical computing and graphics, widely used by statisticians and data scientists all over the world. It is an open source, free GNU project. R-studio provides an environment to code in R. It allows sophisticated data analysis and visualization.

DIVA combines the advantages of SQL and R and overcomes the disadvantages of Excel. It consists of three

Table No. 1

x	y	x^2	y^2	xy
1	3	1	9	3
2	10	4	100	20
3	5	9	25	15
4	1	16	1	4
5	2	25	4	10
6	9	36	81	54
7	4	49	16	28
8	8	64	64	64
9	7	81	49	63
10	6	100	36	60
$\sum x = 55$	$\sum y = 55$	$\sum x^2$=385	$\sum y^2$=385	$\sum xy$=321

modules: Database Operations, Visualization and Statistical Analysis as shown in Fig.1. The GUI of DIVA is designed in Java. All operations related to database are carried out by the Database Operations module. The primary database chosen for this purpose is Oracle and queries are written in SQL. SQL is integrated with Java through Java Database Connectivity (JDBC) driver. Excel spreadsheets may also be imported in Java and can be updated within the Java window itself.

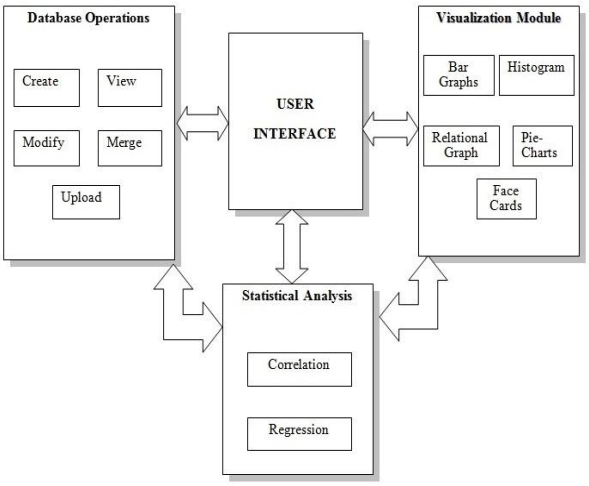

Fig.1 Block Diagram of DIVA

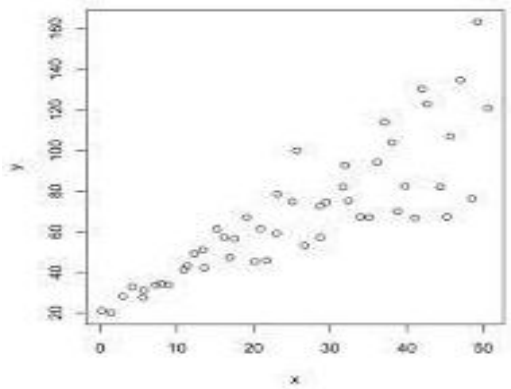

Fig.2 Scatter plot in R-studio

Statistical Analysis and Visualization is carried out in R. Fig.2 and Fig.3 show examples of visualization in R-studio. R-Studio is integrated with Java through the Java-R Interface (JRI).Thus, all visualizations in R can be viewed within the Java window itself. R-Studio can also generate reports on data contained within Comma Separated Value (CSV) files. Statistical analysis is carried out through correlation and regression analysis.

Google Fusion Tables (Hector Gonzalez *et al*, 2010) provides a cloud-based platform for data management and collaboration. However, DIVA is designed to work on static data stored within relational databases and does not work with dynamically generated data. The main disadvantage of using Google Fusion Tables or other cloud-based services is that Small and Medium Scale Enterprises (SMEs) lack trust in cloud computing.

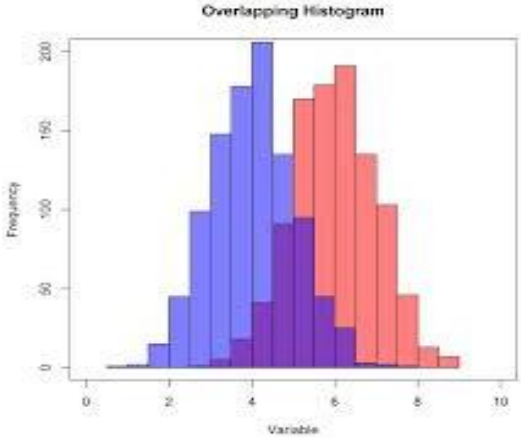

Fig. 3 Histogram in R-studio

3. Correlation

Correlation is a statistical technique that measures the dependency between two variables. The change in one variable may lead to change in the other variable. Increase or decrease in one variable may lead to increase or decrease respectively in the other variable. This is direct or positive correlation. Increase or decrease in one variable may lead to decrease or increase respectively in the other variable. This is indirect or negative correlation.

Correlation is measured through the correlation coefficient. This can be derived through the following algorithms:

3.1 Karl Pearson's coefficient of correlation

It is based on the concept of co-variance and is given by:

Table No. 2

J1	J2	J3	$D_{12}=R_1-R_2$	$D_{13}=R_1-R_3$	$D_{23}=R_2-R_3$	D^2_{12}	D^2_{13}	D^2_{23}
1	4	6	-3	-5	-2	9	25	4
5	8	7	-3	-2	1	9	4	1
4	7	8	-3	-4	-1	9	16	1
8	6	1	2	7	5	4	49	25
9	5	5	4	5	0	16	16	0
6	9	10	-3	-4	-1	9	16	1
10	10	9	0	1	1	0	1	1
7	3	2	4	5	1	16	25	1
3	2	3	1	0	-1	1	0	1
2	1	4	1	-2	-3	1	4	9
						$\sum D^2_{12}=74$	$\sum D^2_{13}=156$	$\sum D^2_{23}=44$

$$r = \frac{\sum xy - \frac{\sum x \sum y}{n}}{\sqrt{\left(\sum x^2 - \frac{(\sum x)^2}{n}\right)\left(\sum y^2 - \frac{(\sum y)^2}{n}\right)}}$$

Here x and y denote the two variables.
$\sum x$ and $\sum y$ represent the sum of all values of x and y respectively.
$\sum x^2$ and $\sum y^2$ represent the sum of all values of $x^2$and y^2respectively.
n represents number of values of x or y.
For example Table No.1 shows values of x and y and corresponding values are substituted in the formula to get $r=0.224$.

3.2 Spearman's Rank Correlation Coefficient

It is used when ranks need to be assigned to variables instead of values. It is given by:

$$r = 1 - \frac{6\Sigma d^2}{n(n^2-1)}$$

Here d is the difference between corresponding ranks of the two series.

n represents number of values of x or y.

For example, Table No.2 shows the ranking given by three judges to ten contestants participating in a contest. *J1*, *J2* and *J3* represent the rankings given by Judge 1, Judge 2 and Judge 3 respectively. Corresponding differences D_{12}, D_{13} and D_{23} are calculated and then squared for each record. Corresponding values are substituted in the formula to get r_{12}, r_{13} and r_{23} as 0.55, 0.05 and 0.73 respectively. Since r_{23} has maximum value, we conclude that the pair of second and third judges has similar judgment in ranking.

Correlation coefficient can take any value between -1 and 1 including both. A value of -1 implies perfect negative correlation whereas a value of +1 implies perfect positive correlation. A value of 0 implies no correlation. Other values lie between -1 and 0 as well as 0 and +1 and a greater magnitude represents high negative and positive correlation respectively. In the above example, since r_{23} has maximum value, we conclude that the pair of second and third judges has similar judgment in ranking. Fig.4 shows the scatter plots for correlation.

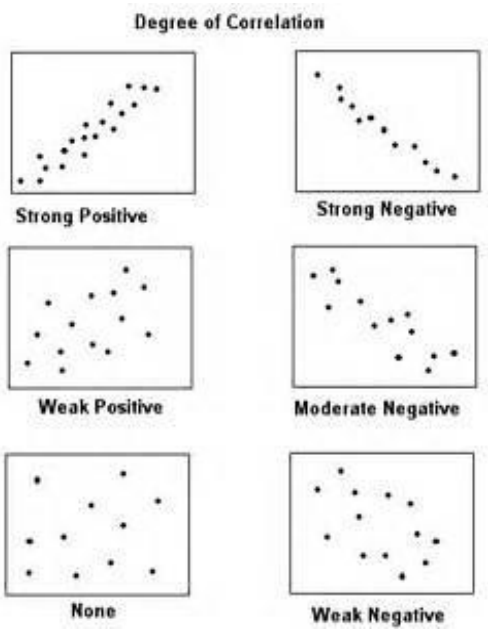

Fig.4 Scatter plots for correlation

4. Regression Analysis

The statistical method which helps us to estimate or predict the unknown value of one variable from the known value of the related variable is called regression. This means that we can calculate the unknown variable from the known variable by analyzing the trend the two variables follow. For two variables, x and y, either y can depend on x or vice versa. Equation (1) shows regression equation of y on x. Equation (2) shows regression equation of x on y.

$$y - y' = \frac{(\sum d_x \cdot d_y)}{\sum (d_x)^2}(x - x') \quad (1)$$

$$x - x' = \frac{(\sum d_x \cdot d_y)}{\sum (d_y)^2}(y - y') \quad (2)$$

Table No.3 shows the years of service of seven employees(x) and their income in multiples of 1000(y). If we need to find out the income of another person who may have worked for 13 years, we can do so by finding the regression line of income on years of service, since the

income needs to be determined depending on the years of service.

Table No. 3

x	y	$d_x=x-8$	$d_y=y-8$	$(d_x)^2$	$d_x . d_y$
11	10	3	2	9	6
7	8	-1	0	1	2
9	6	1	-2	1	-2
5	5	-3	-3	9	9
8	9	0	1	0	0
6	7	-2	-1	4	2
10	11	2	3	4	6
$\sum x=56$	$\sum y= 56$			$\sum (d_x)^2=28$	$\sum dx . dy=21$

Thus by using (1) we get the resultant equation as

$$4y=3x+8 \quad (3)$$

Substituting the value of x as 13, we get y as 11.75. This value in multiples of 1000 is 1175. Hence the income of a person working for 13 years is Rs 1175.

Conclusion

DIVA combines the three aspects of Data Integration, Visualization and Analysis into a single tool. It provides a common platform to work with data in different formats, namely data stored in spreadsheets and databases. It also adds to the ease of interaction as it is centered around the needs of the end user. R programming language provides a great variety of functions that can be used for statistical analysis of the data as well its visualization.

One of the key advantages of using R is that it is a language specifically developed for statistical computing and visualization and one of the most powerful languages within that domain. In future, the tool may be customized to meet the needs of individual organisation and also expanded to enable working with data from different sources. The inception of DIVA as a non cloud-based platform was due to the need to focus on the data, rather than the security aspects or bandwidth requirements that become necessary with cloud-based platforms.

References

Hector Gonzalez, Alan Halevy, Christian S. Jensen, Anno Langen, Jayant Madhavan, Rebecca Shapley, Warren Shen,(2010),Google fusion tables: data management, integration and collaboration in the cloud ,*In: Proceedings of the 1st ACM Symposium on Cloud Computing, SoCC'10,* 175 – 180.

Ying Zhu,(2012), Introducing Google Chart Tools and Google Maps API in Data Visualization Courses, *IEEE Computer Graphics and Application*, vol. 32, 6-9.

J. de Jesus Nascimento da Silva Junior, B.S. Meiguins, N.S Carneiro, A.S.G.Meiguins, R.Y. da Silva Franco, A.G.M Soares, (2012), *16th International Conference on Information Visualization (IV)*, 182 – 187.

L.V.S Lakshmanan, S.N. Subramanian, N. Goyal, R. Krishnamurthy,(1998),On Querying Spreadsheets, *14th International Conference on Data Engineering, Proceedings*, 134 – 141.

T. Pauly, I. Higginbottom, H. Pederson, C. Malzone, J. Corbett, M. Wilson, (2009), Keeping pace with technology through the development of an intuitive data fusion, management, analysis & visualization software solution, *OCEANS 2009 - EUROPE* , vol., no., pp.1,8.

O.P.Malhotra, S.K.Gupta, A.Gangal,(2010), ISC Mathematics, *S. Chand*