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

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

Research Article

Data Integration, Visualization and Analysis: Combined in a Single Tool

Jyotsna Lilani†*, Shikha Duggal†, Sudhir Zanje† and B.B. Gite†

†Department of Computer Engineering, Savitribai Phule Pune University, Pune, Maharashtra, India

Accepted 15 May 2015, Available online 18 May 2015, **Vol.5, No.3 (June 2015)**

Abstract

Data today is being analyzed thoroughly by all organizations in order to take important business decisions. Though large organizations work with a large amount of data, small organizations work with data that is smaller in size and may have a different format as well. Microsoft Excel is often used by such organizations to store data in tabular format. However it is inefficient in terms of certain operations to be carried out on data such as filtering and good visualization. DIVA overcomes the disadvantages of Excel by converting Excel-based tables into MySQL tables by generating a SQL script and executing it. It allows joins and merges to be easily performed on this converted data. Also, it represents the converted data in various graphical forms and performs statistical analysis on it.

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

1. Introduction

The implementation of Data Integration, Visualization and Analysis (DIVA) includes the implementation of the following three modules:

- Integration Module- Includes Excel to MySQL conversion, generation of SQL script, as well as fusion of converted data.
- Visualization Module–Representation of converted data in the form of bar graphs, pie-charts and face cards, which is carried out in R.
- Analysis Module-Filtering the data on the basis of the range of a numerical attribute and correlation and regression analysis in R.

The product is a result of the integration of the following technologies- Excel, MySQL, Java and R. There are many other alternative technologies which perform similar functions but these technologies fulfill the purpose DIVA was designed for. These reasons are highlighted below.

- Excel: Spreadsheets are the most popular format for data storage used over the world. Our target audience includes small time shops that work with static data in Excel and need to analyze such data efficiently.
- MySQL: This is the world's most widely used open-source database. The MySQL Workbench facilitates easy navigation and provides a user-friendly GUI. Since our target audience does not necessarily have training in database systems, it was essential that the technology was easy to use and understand. Since MySQL is open-source, it saves the users of additional cost.
- Java: It was mainly used to design the GUI of the system. Efficient Integrated Development Environments (IDEs) are available that support Java Swing Applications. They are easy to use due to drag and drop design components.
- R: It is the best language for statistical computing and specifically developed for it. Due to the wide range of functions it allows, the scope of DIVA could be expanded to allow more statistical functions that may be needed in the future.

2. Integration Module

The integration module is mainly concerned with the Excel-MySQL conversion. Also it allows natural joins and cross joins to be performed on two tables.

2.1 Conversion of Excel Sheet to MySQL Table

The columns of an Excel sheet represent different attributes of a table. The name of these attributes is generally present in the first row of the sheet. Also, MySQL may have more than one database schemas associated with it. Each database schema is associated with certain tables. Before conversion, we need to select the database schema we want the generated MySQL table to be stored in. Once the database schema is selected, the Excel sheet needs to be browsed and loaded. On loading, different sheets present within one Excel sheet are displayed and the data within them can be viewed.

*Corresponding author: **Jyotsna Lilani**

All the tables present within the selected database schema are displayed. There is an option to override the name of an existing table or create a new table name. The actual process of conversion involves reading the name of the attributes from the first row of the Excel sheet, identifying its data type and the script generation to create the corresponding table in MySQL. Creation of a table involves the CREATE command. Then the Excel data is read row by row and corresponding INSERT queries are performed to insert the records in MySQL table. A SQL script is generated for the CREATE and INSERT command. This script can be saved for future references. It can be either imported into DIVA itself for execution or the script can be pasted into the MySQL Workbench and executed there. This script is saved with a .sql extension. Fig.1 and Fig.2 highlight the script generation of two tables.

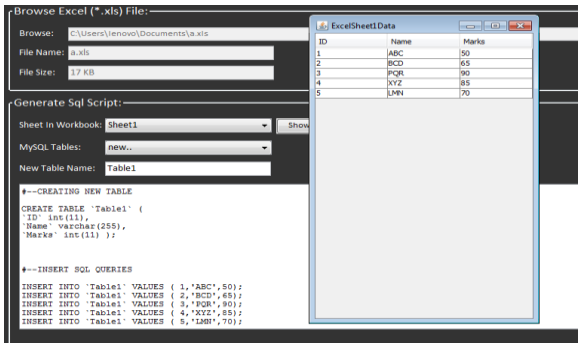

Fig.1 Generation of SQL Script for first table

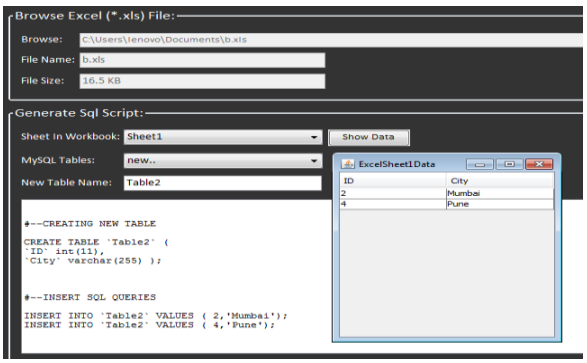

Fig.2 Generation of SQL Script for second table

2.2 Generation of Fusion Table

A fusion table can be generated in two ways. *Firstly,* a natural join can be performed on the two tables. A natural join works on tables that have a common attribute. It displays the records having attributes of both tables for which values of the common attribute match. For example, Table1 highlighted in Fig.1 contains attributes ID, Name and Marks. Table2 highlighted in Fig.2 contains attributes ID and City. A natural join on both tables will produce records containing attributes ID, Name, Marks and City for the values of ID that match, i.e. 2 and 4. This fusion is shown in Fig.3.

Secondly, fusion of two tables may involve a cross join. A cross join combines each row from the first table with each row from the second table. This implies that it carries out the Cartesian product of the sets of rows from the joined tables.

3. Visualization Module

The visualization module allows the data to be represented in different graphical forms such as bar graphs, pie-charts, scatter plots etc. Fig.4 shows the bar graph of Marks v/s ID for the table displayed in Fig.1
The table for which the bar graph needs to be plotted needs to be selected first. The values maybe filtered on the basis of a particular attribute. The attributes present in the table are listed. Any two numerical attributes are selected and the resultant bar graph is displayed. The bar graph is also colour-coded, to increase the visual appeal of the graphical representation.

Similarly pie-charts and scatter plots display the values of two numerical attributes. It is essential that the attributes be numerical in nature i.e. they have mathematical values associated with it.

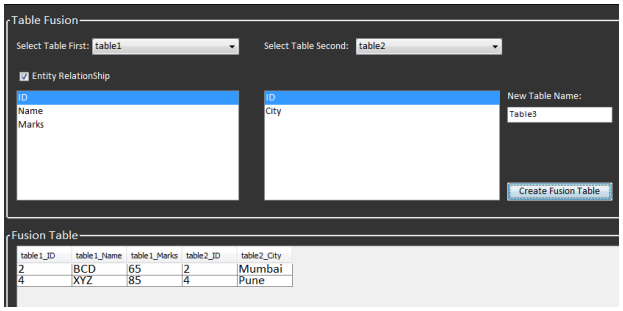

Fig.3 Generation of SQL Script for first table

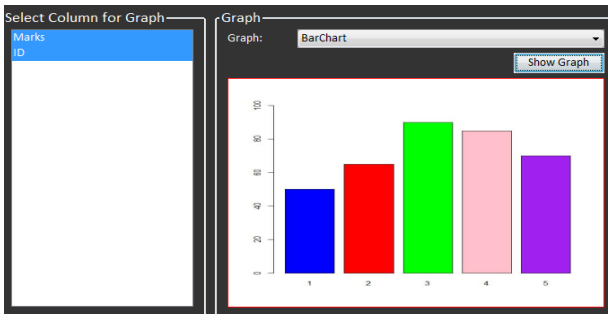

Fig.4 Bar Graph plotted in R

A new concept introduced in DIVA is that of face cards. Face cards allow different attributes of a record to be printed in a sequential manner. Records are mostly arranged horizontally, with each row representing a unique record and each column representing the attribute values. Sometimes there may be a need to view only certain attributes of a record. Also, not all records may be needed to be seen. Face cards allow the user to view certain attributes of certain records in a sequential manner. Thus face cards allow long

horizontal records to be viewed vertically as they would be organised on an identity card containing information for each entity. Also, an option to print the cards is also available.

Fig.5 highlights the use of face cards. The table Table1 contains the attributes ID, Name and Marks. It shows the face cards for the records with ID 2 and 5 with their corresponding marks. The attributes to be displayed need to be selected from the list of attributes and the corresponding records from the list of records needs to be selected.

4. Analysis Module

Analysis module allows filtering of data. Filtering of data implies that certain records may be retrieved from the entire list of records by sorting them on the range of an attribute. For example, filtering of a student data table on the basis of marks between 1 and 100.

It also provides statistical analysis in the form of correlation and regression analysis. Correlation is a measure of interdependency between two numerical attributes. It is measured in the form of a correlation coefficient that has its values between -1 and 1. A positive value indicates direct proportion between attributes whereas a negative value indicates inverse proportion. A value of zero indicates no correlation. For example, we can measure the correlation between the number of years an employee has worked and his salary.

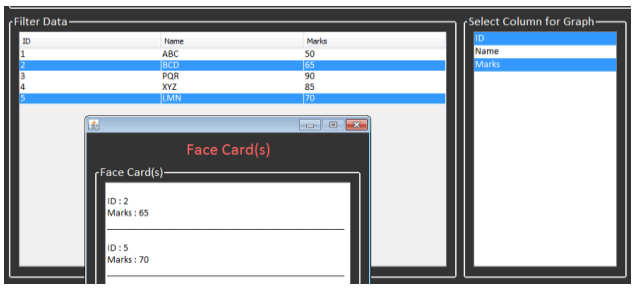

Fig.5 Representation of Face Cards

Regression analysis helps us to predict the value of an attribute for a constant value of another attribute by analyzing the trend between the two attributes in a list of records. For example, we can predict an employee's salary if he has worked for ten years, by analyzing the trend between the two attributes in existing employee records.

Conclusion

DIVA provides a common platform to users to carry out integration, visualisation and analysis of data. It overcomes the disadvantages of spreadsheets and databases and combines their advantages with the computing power of R. Moreover, it is designed keeping simplicity and ease of use in mind. It enables users of Excel to access enhanced functionalities that Excel lacks, without having to store data in a different format.

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.

Shikha Duggal, Jyotsna Lilani, Sudhir Zanje, B.B.Gite,(2014),DIVA: A Tool for Data Integration, Visualization and Analysis, *International Journal of Current Engineering and Technology,*vol. 4,No.6, 3879-3882

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 Visualisation (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*