ComVisMD - compact visualization of multidimensional data: experimenting with cricket players data

Shridhar B Dandin¹ and Mireille Ducassé²

¹BKBIET-Pilani 333031, Rajasthan, INDIA.
²IRISA-INS Rennes, FRANCE

¹shridhar.dandin@bkbit.ac.in
²mireille.ducasse@irisa.fr

Abstract. Database information is multidimensional and often displayed in tabular format (row/column display). Presented in aggregated form, multidimensional data can be used to analyze the records or objects. Online Analytical database Processing (OLAP) proposes mechanisms to display multidimensional data in aggregated forms. A choropleth map is a thematic map in which areas are colored in proportion to the measurement of a statistical variable being displayed, such as population density. They are used mostly for compact graphical representation of geographical information. We propose a system, ComVisMD inspired by choropleth map and the OLAP cube to visualize multidimensional data in a compact way. ComVisMD displays multidimensional data like OLAP Cube, where we are mapping an attribute \(a\) (first dimension, e.g. year started playing cricket) in vertical direction, object coloring based on \(b\) (second dimension, e.g. batting average), mapping varying-size circles based on attribute \(c\) (third dimension, e.g. highest score), mapping numbers based on attribute \(d\) (fourth dimension, e.g. matches played). We illustrate our approach on cricket players data, namely on two tables Country and Player. They have a large number of rows and columns: 246 rows and 17 columns for players of one country. ComVisMD’s visualization reduces the size of the tabular display by a factor of about 4, allowing users to grasp more information at a time than the bare table display.

1. Introduction

In present age, vast amount of data is available at our disposal by the introduction of various Information Systems or Information Processing Devices. In general, the information in a database is in tabular format, in which values are provided by means of rows (records) and columns (fields). A data visualization system can be developed augmenting the tabular representation of information as a graphical representation of results to give or reveal more meaning from the data stored in a database of an Information System.

Map can also be used as information visualization tool for presenting final results. Mapping data enables to notice new trends. For example, in [1] readers can detect scattergram (a classification method) to map percentage of rural population and percentage of population under the age of 18. A type of map called choropleth map presents geographical data in numerical values using colors. It is popular for cartographic presentations.
OLAP is an acronym for OnLine Analytical Processing, which is a computer-based technique of analyzing data to look for insights. The term cube refers to a multidimensional data set, which is also sometimes called a hyper cube if the number of dimensions is greater than 3. Each cell of the cube holds a number that represents some measure such as sales, profits, expenses, budget, etc.

We propose a compact data visualization system ComVisMD inspired by choropleth map and the OLAP cube. The choropleth maps we are using for visualization have dynamic visualization of multidimensional data. The data for illustrating our mapping are taken from the tables of the database of the sports domain cricket.

2. Choropleth Map
A type of map called choropleth map presents geographical data using colors. A choropleth map is a thematic map in which areas are shaded or patterned in proportion to the measurement of the statistical variable being displayed on the map, such as population density or per-capita income. The data used in choropleth map can be classified as equal intervals or geometric progressions or standard deviation. When using color progressions, generally, five to seven color categories are recommended. The map user should be able to easily identify the implied magnitude of the hue and to match it with the legend.

A choropleth map example, shown in Figure 1, represents population density i.e. persons per square kilometre of India, based on 2001 census, district-wise. Straightway one can see highly populated regions just by observing the map, they are either near sea or hill-regions. As identified in [2] choropleth maps are extensively used for data presentation. Any of the patterns of the map can expose some important feature of the data.

3. Cricket players data
For experimenting in ComVisMD, we are considering data of cricket playing countries and their players. We are considering the countries and players playing only One Day International (ODI) cricket matches. The data of countries and their players (as on March 2017) are taken from Wikipedia web page. Similarly, all other 15 countries data are taken. In terms of relational tables, first relation is: Country (cid, countryName, playingSince, matches, won, lost, tied, noResult, winPer, roundWinPer). The second relation is: Player (Pid, Player, First, Last, Matches, Innings, NO, Runs, HS, Avg, Balls, Maidens, Runsg, Avgb, Wkt, Catches, Stumpings).

---

1. en.wikipedia.org/wiki/OLAP_cube
2. en.wikipedia.org
Here, we consider 16 countries that play One Day Internationals (ODIs) affiliated to International Cricket Board: Afghanistan, Australia, Bangladesh, Canada, England, India, Ireland, Kenya, New Zealand, Pakistan, Scotland, South Africa, Sri Lanka, United Arab Emirates, West Indies and Zimbabwe.

Table 1 shows part of the table of Indian players. It contains 17 columns (fields / attributes) and 216 objects or records. In Table 1 we can just see one-seventh of records or objects on part of screen. In order to view all records or objects we propose compact visualization. Our compact approach through ComVisMD is explained in the next section.

4. ComVisMD

In our system, ComVisMD, taking the inspiration of region and coloring of geographical data of choropleth map, we compactly present the multidimensional data. We designed a special mechanism, where we bring all the records on the part of the display screen and perform the task of analysis. Initially we define the terms used for compact view and then we discuss data scale construction (section 4.2).

4.1. Definitions

With reference to Figure 2, we define the different terms used for the compact visualization.

**Map:** Our map is a graphical display to represent all the records (objects) of a database table. All \( n \) number of records of a table \( t \) with reference to selected dimensions can be displayed on the map.

**Row:** A row on the map in horizontal direction displays \( n \) number of records from the table based on classification with respect to values of an attribute \( a \). For example, the second row of map represents 26 players who started playing ODIs in one of the years between 2010 and 2014. The attribute \( a \) is `playingSince`.

**Square:** A square in a row represents a complete record \( rec \) of a database table, displaying a summary computed according to the given attributes, \( b \) and \( c \). For example, the attribute \( b \) is `Runs`, the number of runs scored by the player. The attribute \( c \) is matches, the number of ODI matches played by a player. In each square are displayed,

- **a color:** it is an indicator based on attributes \( b \) and \( c \) to visualize a record \( rec \) according to a 5-value color scale (see section 4.2). It indicates significance of an attribute. For example, batting average based on the attributes, `Runs` and `matches`.

- **a drawn circle:** its size is proportional to the value of an attribute \( d \). It represents magnitude of an attribute. For example `HS`, highest score from all played innings of a player.

- **a number:** the value of an attribute \( e \), for example matches, the number of matches played.

---

3For example, https://en.wikipedia.org/wiki/List_of_India_ODI_cricketers
and https://en.wikipedia.org/wiki/List_of_One_Day_International_cricket_records
For example, in the first square of the seventh row represents the player named Sachin Tendulkar, one of the players from India who has started to play in the year 1989, his batting average is in best slot (red), he has good high score (large circle but not the largest), the number of matches played are 463.

Figure 2 illustrates compact visualization of data from Indian players table. The 216 records of the Player table are first classified on a selected attribute \textit{playingSince} to represent all of them on the map. These squares are distributed in 10 classes (rows), each class is a group of 5 years. If a player started playing ODI in one of these years, the square representing that player (object / entity) is displayed in the respective row.

Blue and light blue color squares signify least important object/s, red and light red color squares signify highly important record/s whereas gray color squares signify average object/s. For example, if the dynamically calculated scale is: 29-59, 16-29, 16, 6-16, 0-6 then 16 represents break point, average or middle value. The upper two ranges are excellent values with respect to the batting average whereas the lower two ranges are least interested values. In this way the objects or records of table are analyzed for significance with respect to an average value.

The size of circle indicates magnitude of the highest score scored by the player. The largest circle indicates largest high score among players of a country whereas the smallest circle indicates lowest high score.

4.2. Dynamic scale construction

A scale is required to classify each record of the table and assign a color. We use divergent color schemes [5] to represent the break point. The break point is the mean value of the data values. With reference to this break point the multidimensional data records of the table can be first classified into three groups: average records, above average records and below average records. Again, based on break point of group of records above and below average are further grouped into two groups resulting in total 5 classes of records.
First, the break point is calculated by finding the average (mean) of n values of the table column e.g. \( \text{Avg} \). Let us call it \( bp \). Second, minimum and maximum values of the column are evaluated. Let us call them \( \text{min} \) and \( \text{max} \) respectively. Third, the average of columns values between \( \text{min} \) and \( bp \) is calculated. Let us call it \( bp_{\text{min}} \). Fourth, the average between the columns values \( bp \) and \( \text{max} \) is calculated. Let us call it \( bp_{\text{max}} \). Now, finally, five ranges of values for the classification are: \( bp_{\text{max}} \) to \( \text{max} \); \( bp \) to \( bp_{\text{max}} \); \( bp \); \( bp_{\text{min}} \) to \( bp \); \( \text{min} \) to \( bp_{\text{min}} \).

The multidimensional data records can be classified on the basis of the dynamic scale constructed. Now using this dynamic scale, the records are clustered into 5 groups, each group represented by means of a chosen color. For example, upper two ranges by red and light red color; lower two ranges by light blue and blue color; the break point value records represented by gray color.

4.3. Discussion

In Figure 2, in total, 216 records / objects of Player table are displayed in the form 216 squares. The left most labels of each row show the period of years like 1970-1974, 1975-1979, and so on up to 2015-2019. This refers to the classification of the players from a country. The players started playing (represented by the field \( \text{playingSince} \) of the Player table) ODI between 1970 and 1974 (inclusive of both the years) is classified under the group, 1970-74. Similarly, all other players are classified and grouped under one of these 10 classes. In the top row represented by class 2015-19, 11 players are grouped, means 11 players started playing ODIs in one of the years of this range. Similarly, 26, 21, 31, 36, 17, 24, 22, 15 and 13 players are grouped under 2010-14, 2005-09, 2000-04, 1995-99, 1990-94, 1985-89, 1980-84, 1975-79 and 1970-74 respectively.

The decreasing order of displaying the players (objects) in respective class is based on the number shown on the square. This number represents the number of ODI matches played by the player. For example, in the first row, the numbers on squares starting from the first are: 12, 11, 7, 6, 6, 3, 3, 3, 2, 1 and 1. They represent the matches played by Manish Pandey, Jasprit Bumrah, Hardik Pandya, Barinder Sran, K L Rahul, Rishi Dhawan, Gurkeerat Singh, Yazuvendra Chahal, Karun Nair, Faiz Fazal and Jayant Yadav respectively.

The legends Batting Average (represented by the field, \( \text{Avg} \)) and the High Score (represented by \( HS \)) of the players are shown in Figure 2. In the first legend, the blue colored square represents batting average between 0 and 6. The light blue colored square represents the batting average between 6 and 16. The gray colored square represents the batting average (mean value) 16, light red colored square represents batting average between 16 and 29, and the last dark red colored square represents the batting average between 29 and 59. In the second legend, the circle (displayed on each square) represents highest score of the player. If the size of the circle is small, then the highest score is low and increase in the size of the circle indicates more higher highest score of a player. The largest circle represents the largest highest score of a player or players.

With reference to the Figure 2, 4 players in top row first, third, fifth and tenth (Manish Pandey, Hardik Pandya, K L Rahul and Faiz Fazal respectively) are represented by red color squares. Similarly, 4, 4, 6, 4, 3, 4, 1, 3 and 6 number of squares are in red color on second, third, fourth, fifth, sixth, seventh, eighth, ninth and tenth rows respectively. It indicates that out of 216 players, these 39 players have the batting average between 23 and 59. Analysis is these are better players than the others. Similarly, 51 players have the batting average between 16 and 29, represented by light red color. 8 players have batting average 16 equal to the mean value (break point) represented by gray color. 55 players have the batting average between 6 and 16, represented by light blue color. 63 players have batting average between 0 and 6 represented by blue color, those are least significant objects (players). Observer of the map can focus on only the best players looking at red colored squares very easily among all 216 objects.

With reference to again the Figure 2, the squares having small circles indicate that these players have low highest scores. Similarly, the squares having big circles clearly indicate larger higher scores, and analysis is these players are performing well.

When the user clicks left mouse button keeping mouse pointer on a square, the object details are displayed in a rectangle below the map for final view of selected player, with same background color.
of the player-square. On click of the right button of the mouse, the user goes back to the first map (cross reference), Country map, basically which is the initial point of the system ComVisMD for compact visualization.

In this way, the user of the compact visualization system ComVisMD can analyse the map to produce different meanings. We can pick few interesting players among a large number of players and display further details of only those interesting players to know more about them one-by-one.

5. Related work
In [4] a set of concepts are organized into a multidimensional cube. The cube focuses on a given subset which is defined by a query. An attribute used in a query is a dimension. This approach of dimensions is inspiration for the dimensions of our system, ComVisMD. In the system VisDB [6] users specify queries and an interface permits to use the visualization. Each pixel of the screen is used to visualize results. In ComVisMD, the queries are not specified by the users and to visualize multidimensional data, objects are represented on a map as colored squares. In Visage [7], bar charts are used to display attributes and objects are dragged for visualization. We bring all objects simultaneously and through dynamic scaling squares (representing objects) are colored for analysis. The authors in [3] were inspired by word-cloud, we are inspired by choropleth map. They claim that largest molecule structure catches the eyes of the observer. In ComVisMD, the 5-colored objects are caught by the observer for the analysis.

6. Conclusion
ComVisMD, compact visualization system for multidimensional data, focuses on compact and multidimensional data visualization. It uses dynamic scaling for coloring squares representing objects. The system displays all objects of the table, moving across tables. Through ComVisMD we have contributed 4-dimension visualization and object comparison without query. In the experiment on cricket players’ data the compacting factor was about 4.

References

[1] Leonowicz A 2003 Research on two-variable choropleth maps as a method for portraying geographical relationships Proceedings of the 21st International Cartographic Conference Durban-SA 2290–2296.

[2] Andrienko G, Andrienko N and Savinov A 2001 Choropleth maps: classification revisited ResearchGate.

[3] Ertl P and Rohde B 2012 The molecule cloud - compact visualization of large collections of molecules Journal of Cheminformatics.

[4] Ferré S, Allard P and Ridoux O 2012 Cubes of concepts: multi-dimensional exploration of multi-valued contexts International conference on formal concept analysis-Springer 112–127.

[5] Horrower M and Brewer C A 2003 Colourbrewer.org: An online tool for selecting colour schemes for maps The Cartographic Journal 40 27–37.

[6] Keim D A and Kriegel H P 1994 Visdb: database exploration using multidimensional visualization IEEE Computer Graphics Applications 40–49.

[7] Roth S F, Lucas P, Senn J A, Gomberg C C, Burks M B, Stroffolino P J, Kolojejchick J A and Dunmire C 1996 Visage: a user interface environment for exploring information In proceedings Information Visualization-IEEE 3–12.