BASIC TECHNIQUES OF VISUAL DATA MINING

Ing. Ivan Šmarda, Doctoral Degree Programme (1)
Dept. of Information Systems, FIT, BUT
E-mail: smardaiv@fit.vutbr.cz

Supervised by: Dr. Jaroslav Zendulka

ABSTRACT

This paper treat of basic techniques visual data mining, especially visualization for tabular data types. Exploring and analyzing the vast volume of data becomes increasingly difficult. Information visualization and visual data mining can help to deal with flood of information and their successfully analyzing. Allowing us to get insight into the data sets, draw conclusion, and directly interact with data.

1 INTRODUCTION

Visual data mining is an emerging area in explorative and intelligent data analysis and mining which is based on integration of concept from computer graphics, visualization metaphors and methods, information and scientific data visualization, visual perception, cognitive psychology, diagrammatic reasoning, 3D virtual reality systems and recently, from affective computing and collaborative virtual environments. Visual data mining is collection of interactive reflective methods that support exploration of data sets by dynamically adjusting parameters to see how they affect the information being presented. Visualization methods allow discover overall trends in data set while also affording an opportunity to discover smaller hidden patterns that can often be just important within application.

Visualization has proven to be reliable, easy to learn, and extremely cost effective. Additionally, visualizations provides a natural method for integrating multiple data sets and has been used across a number of disciplines including commercial research, forensic accounting, and throughout the investigative community.

2 BASIC IDEA

For data mining to be effective, it is important to include human in the data exploration process and combine the flexibility, creativity, and general knowledge of the human with the enormous storage capacity and the computational power of today’s computers. Visual data exploration aims at integrating the human in the data exploration process, applying its perceptual abilities to the large data sets available in today’s computer systems. The basic idea of visual data exploration is to present the data in some visual form (2D, 3D pictures of business data sets), allowing the human to get insight into the data, draw conclusions, and directly interact with data. User acts as the data mining or pattern recognition engine. By visually inspecting and interacting with the two or three dimensional visualization, you can identify the interesting (nontrivial, implicit, perhaps previously unknown and potentially useful) information or patterns in the business data sets. The visual data exploration can be seen a hypothesis generation process. The visualizations of data allow the user to gain insight into the data and come up with a new hypothesis. The verification of the hypotheses can also be done via visual data exploration but it may also be accomplished by automatic techniques from statistic or machine learning. Visual data exploration is especially useful when little is known about the data and the exploration goal is vague.

The main advantages visual data exploration:

- Visual data exploration can easily deal with highly inhomogeneous and noisy data
- Visual data exploration is intuitive and requires no understanding of complex mathematical or statistical algorithms or parameters

As a result, visual data exploration usually allows a faster data exploration and often provides better results, especially in cases where automatic algorithms fail. In addition, visual data exploration techniques provide a much higher degree of confidence in the findings of the exploration. This fact leads to a high demand for visual exploration techniques and makes them indispensable in conjunction with automatic exploration techniques (without proof assumed from [1]).

A major problem in the development of visual data mining methods is that in absolute sense there is no one type of visualization model that is better than another. The visualization method of choice will usually be determined by such factors as appropriateness for the application domain, scope of the data sets, how data is mapped onto visualization schemata, how the visualization schemata utilizes cognitive “strengths” of human information processing and how it overcomes its cognitive limitations.

3 METHODOLOGY

This contribution assumes that the reader is familiar with data mining process and has fundamental knowledge about relational databases.

In quick preview, the visual data mining process consist of three main phases:

- Project planning phase
- Data preparation phase
- Data analysis phase

We will aim at third phase – data analysis phase. Third phase is divided to three steps (according to [2]):

- Choose the Visualization or Mining Tool
- Analyze the Visualization or Mining Model
- Verify and Present the Visualization or Mining Model

And at second point of this phase lies (Analyze the Visualization) the major reason of this article. How we can visualize and analyze typical 2D problem of relational data sets in the best way? In next part of this paper we sketch some examples and techniques for better understanding of basic idea of visualization.

4 SKETCH OF VISUAL ANALYSIS

This part focuses on interpreting what the data visualizations are showing you about your prepared data set and what information can be derived to answer business question (like what are the sales trends year-to-date, what are the key business indicator as to why customers are leaving, which customer segment is most likely to buy a new product, …).

The data visualization tool used depends on the nature of the business data sets and underlying structure. Furthermore, the visualization tool or technique depends on the facet of the business question you are investigating. Perhaps the best way to learn how to use data visualization to analyze and evaluate your business data set is through examples. The following is a brief list of some uses for data visualization in analyzing and evaluating your business data sets and gaining insight into your business questions. Assumed to [2] you can use:

- Frequency graphs (histogram and distribution graphs combined) to evaluate key business indicator such as response rate to marketing campaign by comparing them to other columns (data dimensions) in the business data set
- Pareto graphs (histograms and line graphs combined) to evaluate the importance of a column (data dimensions) in the business data set with respect to key business indicators such as response rate, number of failures, or fraudulent claims
- Radar graphs to search and evaluate seasonal trends or other trends in the business data set as a whole
- Line graphs to analyze and search for time-based trends and patterns
- Scatter graphs to evaluate cause-and-effect relationship

Many times when you plot the graphical data table (formatted subset of the business data set used to create graph) using the appropriate graph type, underlying patterns are discovered. For instance, a scatter graph may show a positive, negative, neutral, threshold, or curved relationship between columns. Often you can use this relationship information and other visual discoveries and insights to help analyze and answer the business questions under investigation. In addition, you can use the data visualization as a communication tool to share your discoveries and insights with other data and business analyst, domain experts, and decision makers.

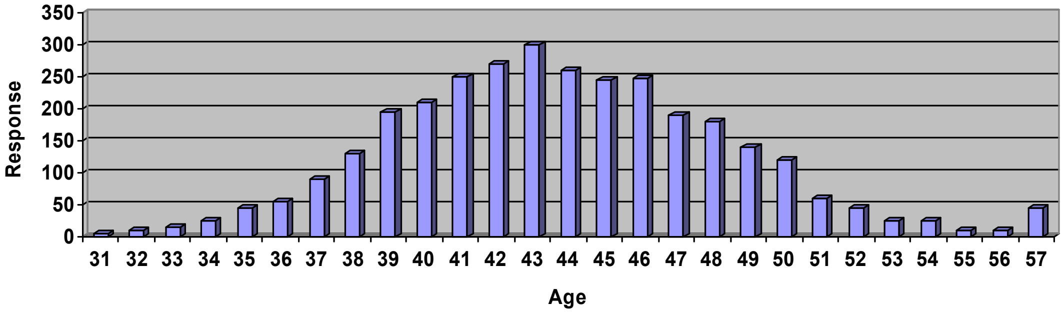
4.1 USING FREQUENCY GRAPHS TO DISCOVER AND EVALUATE KEY BUSINESS INDICATORS

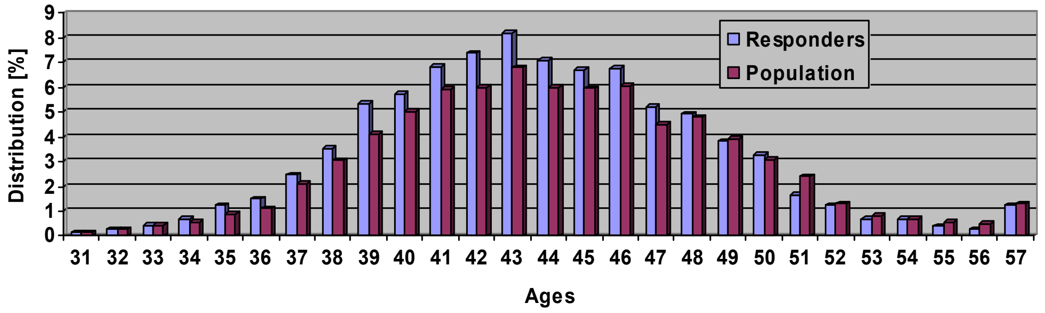
A specialized variation of the column or bar graph is referred to as a frequency graph. Using examples is noted above, but we describe this at frequency graph of responses by customer age. You can use distribution graph to evaluate these business indicator by other data dimensions in the business data set, such as age range (our case), gender, martial status, etc.

Depending on your data visualization tool, you often need to reformat the data in your business data set in order to plot is using a frequency graph. Code below shows a standard SQL statement that construct a graphical data table that you can use for creating a frequency graph.

```
SELECT    age,
          Count(response)
FROM      response_business_set
WHERE     response-1
GROUP BY  age
```

The next figure illustrates how a frequency graph (column graph) can be used to evaluate the distribution of customers who responded to a home equity loan campaign.





Second figure shows a column graph that compares the frequency graph of responders by the age to the frequency graph of the entire customer population. As you can see from second figure, the age trends follow similar patterns. However, up to age 43, the responders tend to have lower percentages than population. Therefore, it appears the younger customers are more interested in the home equity loans than are older customers. The frequency graphs also provide you with descriptive information about the age ranges of your customers and those who did and did not respond to the home equity campaign.

Frequency graphs typically have one of five basic patterns: symmetric, bimodal, skewed, flat, or outlier.

5 CONCLUSION

As we can see, the difficulty of visualization above example is very trivial. And this is one of many sides a strength of visual techniques in visual data mining. There are many other techniques of visualization, but we cannot provide completely brief list of these techniques due range of this contribution. We suggest you to search a read some of articles or books in mentioned in references.

In the future, the visual data mining will involve the tight integration on visualization techniques with traditional techniques from such disciplines as statistics, machine learning, operations research, and simulation. Integration of visualization techniques and these more established methods would combine fast automatic data mining algorithms with intuitive power of the human mind, improving quality and speed of the visual data mining process.

REFERENCES

[1] Keim, D. A.: Information Visualization and Visual Data Mining, IEEE Transactions on visualization and computer graphics, Vol. 7, NO. 1, January – March 2002

[2] Soukup, T., Davidson, I.: Visual Data Mining, Wiley Computer Publishing 2002, ISBN 0-471-14999-3

[3] Muller, W., Schumann, H.: Visual Data Mining, University Rostock, Germany

[4] Simeon, J. S., Noirhomme-Fraiture, M., Bohlen, M.H.: Foreword, Proceedings of the International Workshop on Visual Data Mining, 2001, Freiburg, Germany

[5] Weiss, S. M., Indurkya, N.: Predictive Data Mining – a practical guide, Morgan Kaufmann Publishers, Inc., 1998, San Francisco, California, ISBN 1-55860-403-0

[6] Rud, O. P.: Data Mining, Computer Press, Praha 2001, ISBN 80-7226-577-6

[7] Lacko, L.: Datové sklady, analýza OLAP a dolování dat, Computer Press, Brno 2003, ISBN 80-7226-969-0

[8] Šimůnek, M.: SQL kompletní kapesní průvodce, Grada Publishing, s.r.o., Praha 1999, ISBN 80-7169-692-7