Erik Johansson-Evegård
Elias Holmström

TNM033
LiU, 2009-12-04

Data Mining and Visualization

1. Introduction

There is a lot of visualization techniques that analyze data in different ways. Depending on the type of the data set some techniques are more effective than others. In this paper, we look at the survey of visualization tools for data mining that Olivera et al. published [1]. We take a brief look at some techniques commonly used, such as boxplots [2], stem-plots [3], tree-mapping [4], scatter plots [6], RadViz [5], parallel coordinates [6], Chernoff faces [6], Recursive patterns [7], The circle segment technique [7] and graph-based techniques [8].

This paper is divided into three main parts: Visualization techniques for pre-processing, Summary of different techniques and tools and a section Application where we talk about the software WEKA and some real world applications in the visualization field.

2. Visualization Techniques for Pre-processing

2.1 Box-Plots

Boxplots are used for showing groups of numerical data. The lower and upper ends of the boxes represents the 25th and 75th percentile. The line inside the box is the median (50th percentile), and the end of the tails represents the 10th and 90th percentile. In outliers are, in this case, marked by a circle:

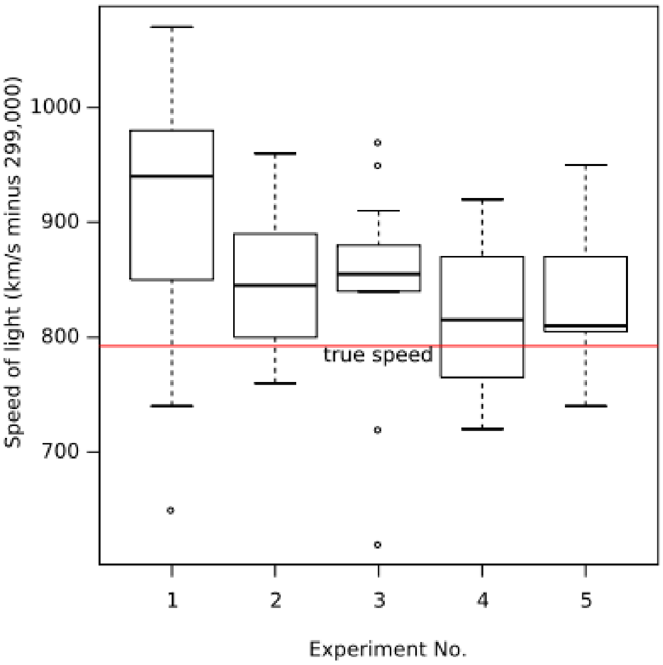


Figure 1: Boxplot [2]

Good to compare distributions between sets of data but only gives a rough overview.

2.2 Stem and Leaf Plot

This method was commonly used in the 1980's [3]. For example, the greatest digit is the stem, and the remaining digits represents the leaves. Consider the data series :

```
44 46 47 49 63 64 66 68 68 72 72 75 76 81 84 88 106.
```

To create a stem and leaf plot, sort the stem and leaf in separate columns:

```
4   | 4 6 7 9
5   |
6   | 3 4 6 8 8
7   | 2 2 5 6
8   | 1 4 8
9   |
10  | 6
```

It is useful to get a quick overview of the density of a data set and to observe outliers. However, when dealing with large sets of data, the plot can be cluttered, and other methods are more suitable, such as a box plot or a histogram.

3. Summary of Different Techniques and Tools

In 2003 Olivera et al. published [1]. A survey of visualization tools for data mining. An overview will be given of the most popular methods in the survey classified by Keims taxonomy [11] . He have written several papers on the topic and chooses to classify the different techniques into six types: Geometric projection, icon-based, pixel-oriented, hierarchical, graph-based and hybrid types. Interaction methods will be briefly discussed and also some more recent techniques will be presented.

3.1 Geometric Projection

Geometric projection techniques are a good choice for finding outliers and correlation between attributes in multivariate data. A geometric projection technique does this by using transformations and projections of the data. When using large data sets a clustering algorithm is usually necessary to apply before the visualization technique to avoid cluttered and unclear data caused by the too much information. Some widely used geometric projection techniques are:

Scatter plots
A scatter plot [6] is one of the most common visualization techniques and can be visualized both in 3D and 2D. The scatter plot visualizes different attributes of the data on the x,y axis for 2D visualizations and also along the z-axis in 3D. Scatter plots are usable to find correlations between attributes in arbitrary small data sets. If the data set gets too big or contains too many attributes the scatter plot gets cluttered and hard to interpret.

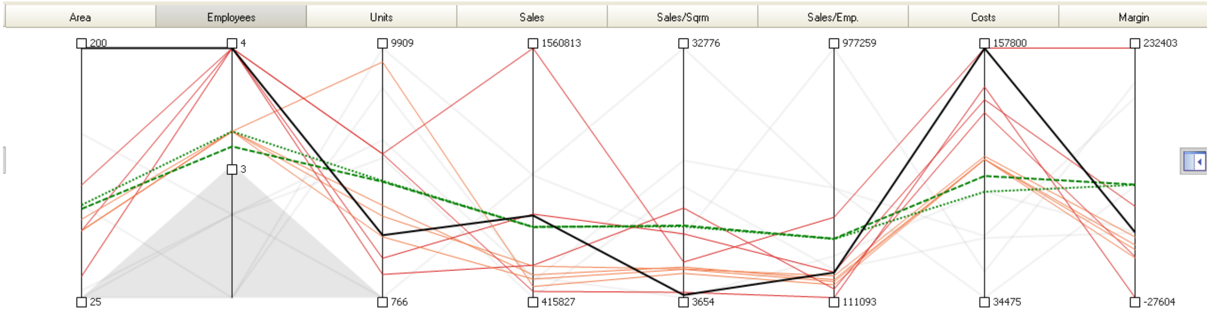


Figure 2: Parallel coordinates [9]

Parallel coordinates
If the data contains many attributes you will need several scatter plots to visualize all attributes and it might become hard to see patterns. A technique often used for multivariate data is parallel coordinates [6]. They work by visualizing each attribute on a vertical axis and connecting each individual data with lines between the axis. Parallel coordinates strength is the possibility to find correlations between a high amount of attributes. The weakness of parallel coordinates is the same as with scatter plots, if the data set is too large it easily gets cluttered.

RadViz
RadViz [5] is another technique for visualizing multivariate data. It maps data to a 2D plane using Hooke's law and a set of anchor points usually specified from the attributes.

3.2 Icon-based

Icon-based techniques visualize data by changing the properties of an icon or glyph according to the data. An early version was Chernoff faces [6] where data is mapped to different face parts as nose, mouth, eyes and more. For example how rich people are can be mapped to the mouth of the Chernoff face. Rich people represented by a happy mouth and and poor people by a sad mouth. Other methods are:

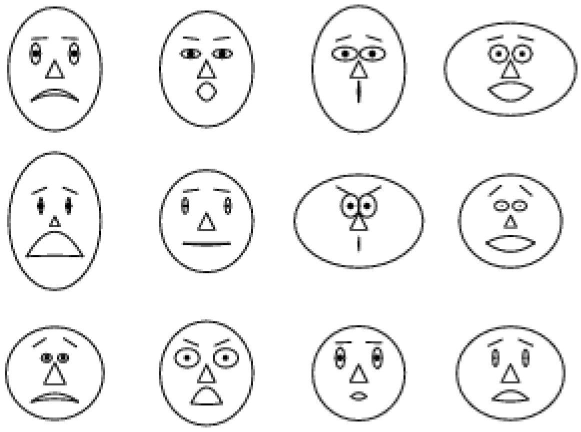


Figure 3: Chernoff faces [10]

Icon-based methods have the strength of being easily interpreted if icons/glyphs are chosen wisely. A poor choice can make it difficult to distinguish differences.

3.3 Pixel-oriented

Pixel-oriented or Pixel-based techniques using a single pixel to represent a data attribute. Depending on the value it is mapped to a color-map to choose the correct color for the pixel.
Pixel-oriented techniques can also be either query dependent or independent. When independent all attributes of the data are drawn on screen independently of the each other. Using a dependent technique you specify a query that decides how pixels should be placed enabling the possibility to find patterns. When using a pixel-oriented method it is possible to visualize very large data sets as each instance only needs a single pixel to be visualized. If not using a good method of drawing pixels you will end up with a visualization that can't be interpreted.

Recursive pattern
Recursive patterns [7] makes it possible to group data when visualizing to make it easier to interpret. It works by using several different levels of patterns by specifying height and width of for the different levels. A first level pattern might organize the pixels to make it possible to interpret data over a year while the second makes it possible to interpret over the month.

Circle segments
The circle segment technique [7] divides a circle into k-dimensions for the amount of attributes k in the data set. Within each dimension each attribute is then visualized by coloring a single pixel. Close to the center all attributes are close making it easier to compare their values.

3.4 Hierarchical

Hierarchical techniques visualize data using subspaces created from the data's attributes. Hierarchical techniques are usable when some attributes of the data might be more relevant than other.

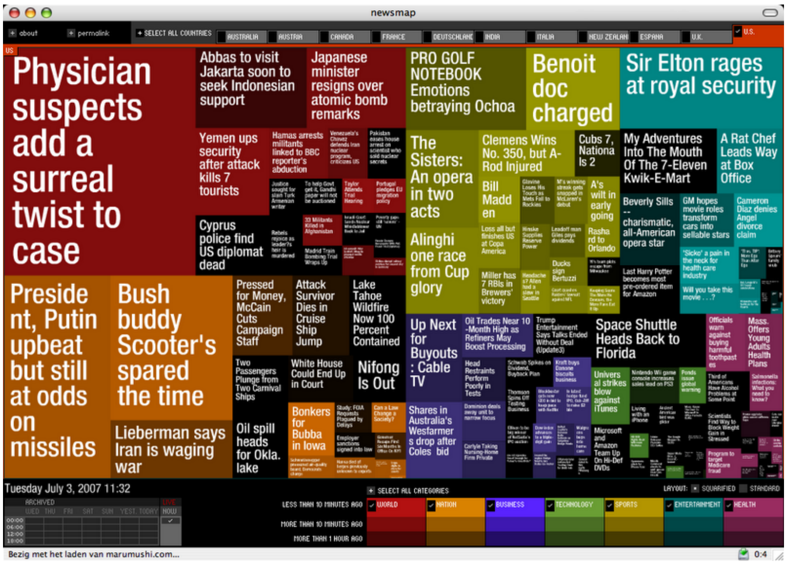

newsmap
about
permalink
SELECT ALL COUNTRIES
AUSTRALIA
AUSTRIA
CANADA
FRANCE
DEUTSCHLAND
INDIA
ITALIA
NEW ZEALAND
ESPAÑA
U.K.
U.S.
Physician suspects add a surreal twist to case
Abbas to visit Jakarta soon to seek Indonesian support
Japanese minister resigns over atomic bomb remarks
PRO GOLF NOTEBOOK Emotions betraying Ochoa
Benoit doc charged
Sir Elton rages at royal security
Yemen ups security after attack kills 7 tourists
Hamas arrests militants linked to BBC reporter's abduction
Cyprus police find US diplomat dead
The Sisters: An opera in two acts
Clemens Wins No. 350, but A-Rod Injured
Cubs 7, Nationa ls 2
My Adventures Into The Mouth Of The 7-Eleven Kwik-E-Mart
A Rat Chef Leads Way at Box Office
Bill Madd en
Alinghi one race from Cup glory
Beverly Sills
charismatic, all-American opera star
Presidênt, Putin upbeat but still at odds on missiles
Bush buddy Scooter's spared the time
Pressed for Money, McCain Cuts Campaign Staff
Attack Survivor Dies in Cruise Ship Jump
Lake Tahoe Wildfire Now 100 Percent Contained
Up Next for Buyouts: Cable TV
Space Shuttle Heads Back to Florida
Two Passengers Plunge from Two Carnival Ships
White House Could End Up in Court
Nifong Is Out
Bonkers for Bubba in Iowa
Shares in Australia's Wesfarmer s drop after Coles bid
Lieberman says Iran is waging war
Oil spill heads for Okla. lake
Tuesday July 3, 2007 11:32
ARCHIVED
SELECT ALL CATEGORIES
LESS THAN 10 MINUTES AGO
MORE THAN 10 MINUTES AGO
MORE THAN 1 HOUR AGO
WORLD
NATION
BUSINESS
TECHNOLOGY
SPORTS
ENTERTAINMENT
HEALTH
LAYOUT: SQUARIFIED STANDARD
Bezig met het laden van marumushi.com...

Figure 4: *Treemap used for displaying news* http://newsmap.jp/

Treemap
Treemaps [4] display hierarchical data using rectangles. Each branch of the tree is assigned a rectangle. Then each sub-branch gets assigned to a rectangle and this continues recursively until a leaf node is found. Depending on choice the rectangle representing the leaf node is colored, sized or both according to chosen attributes. This is a good way to spot relevant data and is good for categorized data, although the result can be cluttered when dealing with large data sets.

3.5 Graph-based

Graph-based techniques visualize large graphs using algorithms and abstraction layers to present a clear overview of the graph. E.g. in a graphical approach presented by Kilian Thiel, Fabian Dill, Tobias Kötter, and Michael R. Berthold published in [8], the extracted terms as well as the periods are represented by vertices's of a graph. The terms occurring in a certain period are connected to the referring vertex. This kind of visualization is less intuitive when searching in large data sets but has the advantage of being able to see relations and large amounts of data in a single graph.

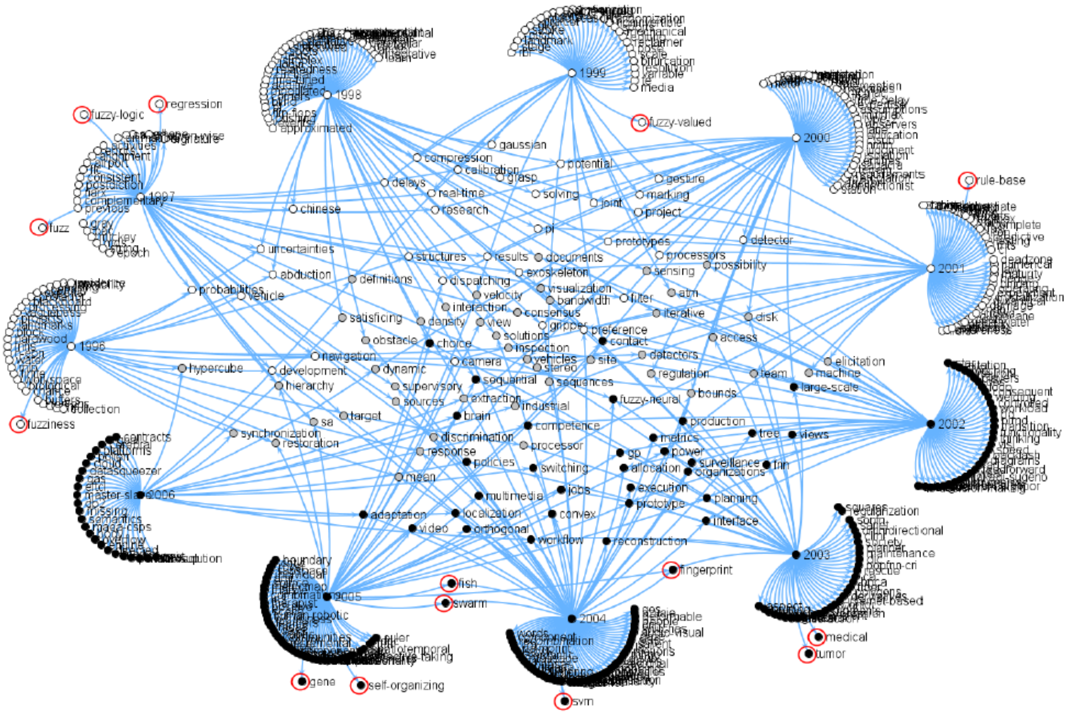


Figure 5: *Visualization based on a network model [8].*

Figure 5 is a good example of a graph-based visualization. Publishing dates of documents, as well as their terms are represented by the vertices's of a graph. Terms related to a specific publishing year are connected to the vertex of the year via an edge. By usage of activation spreading techniques, terms frequently occurring in documents published in particular years can be discovered visually. [8].

3.6 Hybrid

Hybrid methods integrate multiple visualization techniques in several or one window. An important aspect of hybrid methods are linking between windows. Modifying or selecting data in one window shall affect others.

3.7 Interaction

To make better use of many visualization techniques they are combined with different interaction methods. One discussed earlier is linking which enables selection and filtering on one visualization to affect another.

In large visualizations FishEye zooming which is a detail on demand method can be used to give a better view of a special part of the visualization. This technique is especially well suited for large graphs or pixel-oriented techniques. Other methods for detail on demand could be tool-tips for extra information an a selected data or attribute.

4. Applications

4.1 Practical Usage

NCVA at Linköping University [9] are using visualization for several practical tasks. By using hybrid methods combining glyph-based with geometric projection techniques and interactivity to create tools for as separate areas as weather, telecommunication and statistical exploration.
Many of the techniques mentioned are used in all different fields. Especially glyphs are often used to visualize regional data on top of maps, parallel coordinates to find correlations in multivariate data sets and scatter plots for finding them between a limited amount of attributes.

At newsmap.jp Treemaps are used for data mining current news articles. It creates and presents an easily over viewed layout of the biggest headlines of today based categories and countries.

4.2 Visualization Tools Available in WEKA

The main focus of the software is algorithms for data mining tasks, therefore the visualization tools available in WEKA are limited.

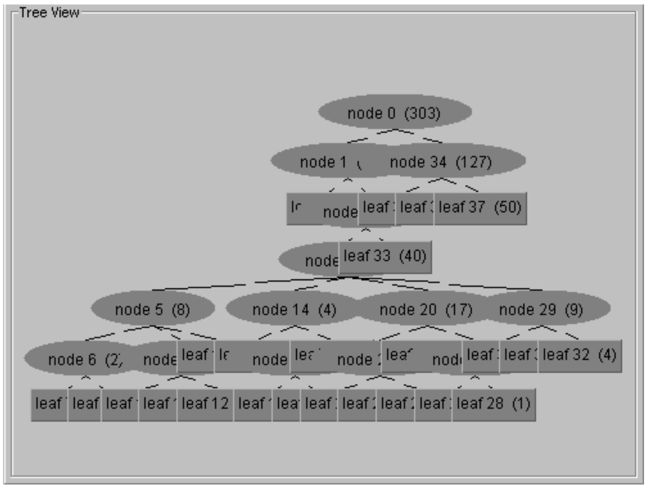


Figure 6: Tree-graph visualization of hierarchical clusters.

In WEKA it is possible to visualize the output of a few hierarchical clustering methods such as Cobweb as tree-graphs. The output is not very intuitive and there is no possibility to rearrange the nodes and leafs to obtain a clearer result.

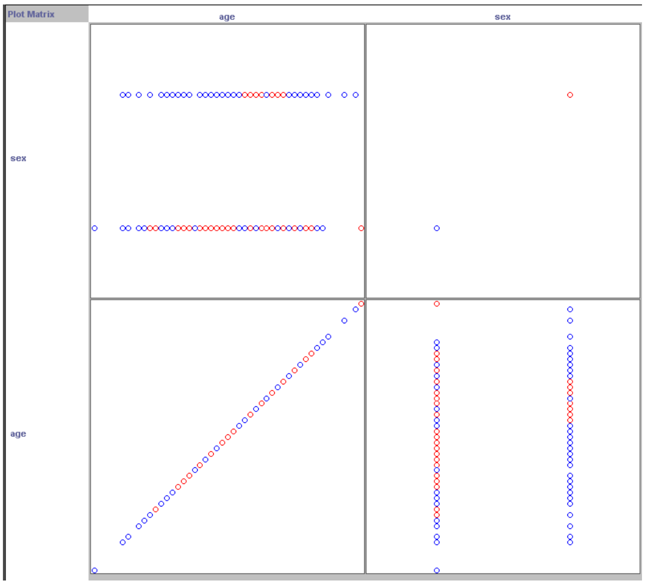
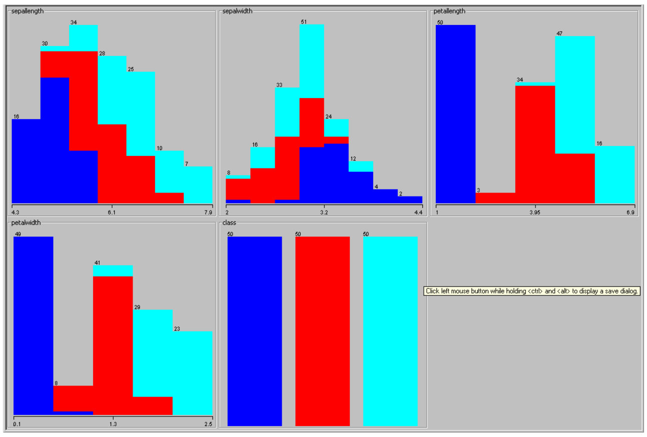


Figure 7: Scatter plot visualization of data set.

The main visualization functionality in WEKA is focused around scatter plots. It is possible to compare different attributes. Available settings for appearance of the scatter plot are plot size, point size, jitter and colors.



Click left mouse button while holding <ctrl> and <alt> to display a save dialog.

Figure 8: Bar diagrams of data set during pre-process

During the pre-process step it is possible to visualize attributes as bar diagrams to uncover relations.

5. Conclusion

In this paper we have presented a brief overview of different techniques that can be used for data mining using visualization according to taxonomy by Keim in [1]. Among the techniques that are being used are parallel coordinates and glyphs for multivariate data and scatter plots for a limited amount of attributes. Using visualizations for data mining enables a user to model data on a large scale. With visualization it is possible to develop a model for example weather data within seconds that would be almost impossible with regular data mining methods.

In WEKA the tool used for data mining in this course there are few tools available for visualization. These are tree-graph, scatter plots and bar diagrams.

Different techniques are suited for different areas of usage. Compared to regular data mining visualizations are still a new subject and therefore there are not yet as many practical situations where it is widely used.

Bibliography

[1] Maria Cristina Ferreira de Oliveira, Haim Levkowitz. F*rom Visual Data Exploration to Visual Data Mining: A Survey*. IEEE TRANSACTIONS ON VISUALIZATION AND COMPUTER GRAPHICS, VOL. 9, NO. 3, JULY-SEPTEMBER 2003.

[2] Box plot. (2009, November 30). In *Wikipedia, The Free Encyclopedia*. Retrieved 10:52, December 2, 2009, from http://en.wikipedia.org/w/index.php?title=Box_plot&oldid=328879318

[3] Stemplot. (2009, November 5). In *Wikipedia, The Free Encyclopedia*. Retrieved 11:24, December 2, 2009, from http://en.wikipedia.org/w/index.php?title=Stemplot&oldid=324061080

[4] Treemapping. (2009, December 2). In *Wikipedia, The Free Encyclopedia*. Retrieved 11:29, December 2, 2009, from http://en.wikipedia.org/w/index.php?title=Treemapping&oldid=329207488

[5] Nováková, L. and Štepánková, O. 2006. Multidimensional clusters in RadViz. In *Proceedings of the 6th WSEAS international Conference on Simulation, Modelling and Optimization* (Lisbon, Portugal, September 22 - 24, 2006). A. M. Madureira, Ed. World Scientific and Engineering Academy and Society (WSEAS), Stevens Point, Wisconsin, 470-475.

[6] Pang-Ning Tan, Michael Steinbach and Vipin Kumar, *Introduction to Data Mining* , ISBN 0-321-32136-7. 2005

[7] Daniel A. KEIM, *Pixel-Oriented Visualization Techniques for Exploring Very Large Data Bases*. Journal of Computational and Graphical Statistics. 2006

[8] Kilian Thiel, Fabian Dill, Tobias Kötter, and Michael R. BertholdTowards, *Visual Exploration of Topic Shifts*, 2007

[9] NCVA Linköping University. *GeoAnalytics Visualization*. http://vita.itn.liu.se/pub/jsp/polopoly.jsp?d=13602&a=93795

[10] Gonick, L. and Smith, W. *The Cartoon Guide to Statistics.* New York: Harper Perennial, p. 212, 1993.

[11] Daniel A.Keim. *Information Visualization and Data Mining*. 2002.