Visualization of Trees as Highly Compressed Tables with InfoZoom

Michael Spenke[*] and Christian Beilken[†]
FIT – Fraunhofer Institute for Applied Information Technology

Abstract

This paper describes the application of our data analysis tool *InfoZoom* to the tree structured data supplied for the InfoVis 2003 Contest. InfoZoom was not especially designed for the analysis of trees, but is a general tool for visualization and exploration of tabular databases. Nevertheless it is well suited for the analysis and pair wise comparison of trees.

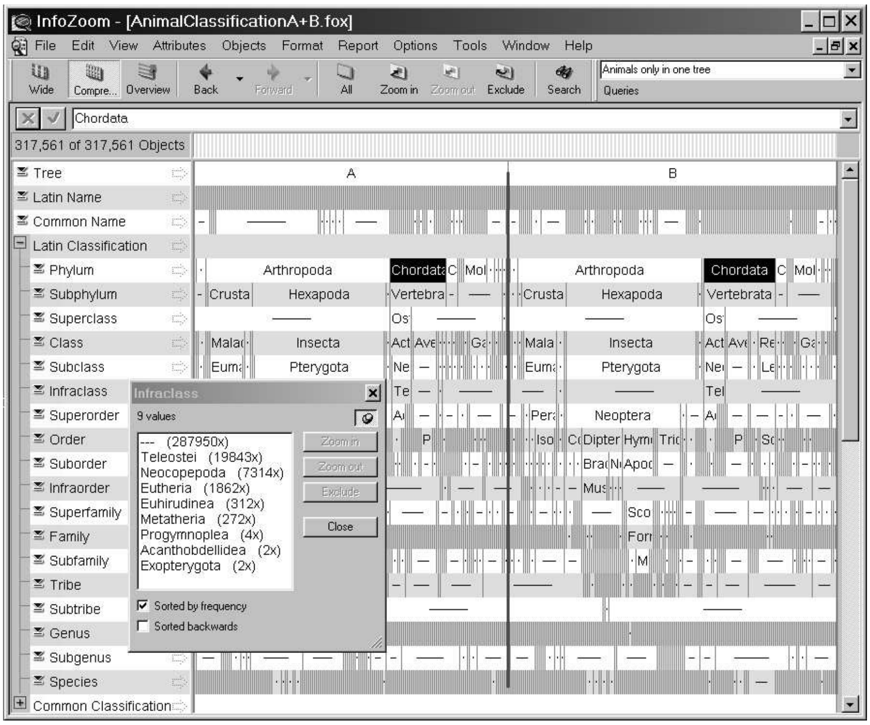
CR Categories: H.5.2 [Information Interfaces and Presentation]: User Interfaces – Graphical User Interfaces (GUI).

Keywords: Information visualization, interactive data exploration, user-interfaces for databases.

1 InfoZoom as a Tree Browser

InfoZoom displays database relations in tables with attributes as rows and objects as columns. Therefore, we had to transform the XML trees to a tabular representation. Each leaf of the tree, i.e. each animal, constitutes a column of the table. The path from a leaf to the root is stored in the attributes (rows) of the table. Since we display both of the trees A and B side by side, our table contains more than 300,000 columns.

The basic concept InfoZoom is to compress even such large tables by reducing the column width until all columns fit on the screen (Figure 1). The column width is here about 0.002 pixels. Special techniques are used to make such highly compressed tables readable. The most important is that neighboring cells with identical values are combined into one larger cell. Because there are 150,000 adjacent columns with the value *A* for the attribute *Tree*, *A* is displayed only once in a large cell. The width of a cell indicates the number of consecutive objects with this value. Therefore, we can conclude from Figure 1, that the trees A and B have roughly the same number of leaves.



Infraclass

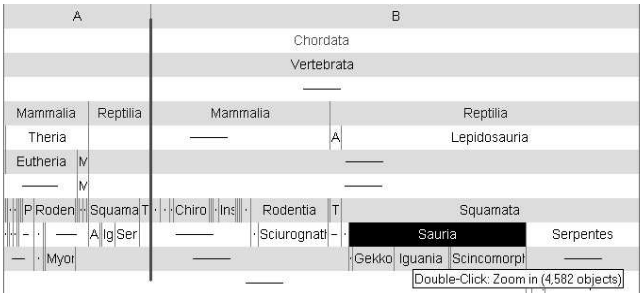
9 values

```
---  (287950x)
Teleostei  (19843x)
Neocopepoda  (7314x)
Eutheria  (1862x)
Euhirudinea  (312x)
Metatheria  (272x)
Progymnoplea  (4x)
Acanthobdellidea  (2x)
Exopterygota  (2x)
```

☑ Sorted by frequency
☐ Sorted backwards

Figure 1. The two animal trees as a table

We can also observe that the *Arthropoda* mainly consist of *Crustacea* and *Hexapoda* and that nearly all *Insecta* are *Pterygota*. The two trees look quite similar at this level of detail. However, the two cells for *Chordata* have different sizes. Obviously, in tree B there are more Chordates than in A. So it is a good idea to take a closer look at the Chordates. Pressing the zoom-in button or double-clicking one of the marked cells leads to an animated zoom on the Chordates: The black cells grow while the other cells in this line slowly disappear. After another zoom on *Mammalia* and *Reptilia* the result shown in Figure 2 is reached.



Double-Click: Zoom in (4,582 objects)

Figure 2. The two trees contain different mammals and reptiles

We can see that there are more mammals and reptiles in tree B. The group of 4,582 *Sauria* is completely missing in A. On the other hand the subclass *Theria* and the infraclass *Eutheria* are missing in B. By further zooms e.g. into the Mammals we can now analyze the differences in more detail.

----

[*] e-mail: michael.spenke@fit.fraunhofer.de
[†] e-mail: christian.beilken@fit.fraunhofer.de

2 Systematic Analysis of the Differences

Like the formula-cells in a spreadsheet program, derived summary attributes (like sum, count, list, average, maximum, etc.) can be defined which are automatically updated by InfoZoom whenever necessary.

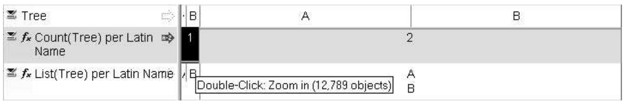

Tree
B
A
B
Count(Tree) per Latin Name
1
2
List(Tree) per Latin Name
B
Double-Click: Zoom in (12,789 objects)
A
B

Figure 3. Which animals can be found in both trees

In Figure 3 we have defined a derived attribute *Count(Tree) per Latin Name*. It shows that most animals can be found in both trees. However, 12,789 animals appear in only one of the trees. We zoom on these by double-clicking the marked cell and get a result similar to that in Figure 2.

Next we want to find all animals which exist in both trees, but with a different classification. Therefore, we define a new derived attribute *Latin Path* as the concatenation of all levels of the classification and the *Latin Name*. We get pathnames like

Annelida/Polychaeta/Palpata/Fauveliopsidae/Flota/Flota flabelligera

Now we can determine where there are two different pathnames for one *Latin Name* (Figure 4).

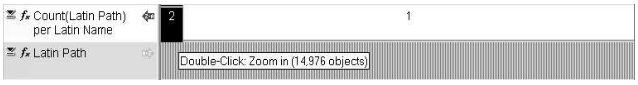

Count(Latin Path) per Latin Name
2
1
Latin Path
Double-Click: Zoom in (14,976 objects)

Figure 4. Which animals are differently classified in each tree

After a zoom on the marked cell only the 7,488 animals with 2 different paths remain visible. The result is shown in Figure 5.

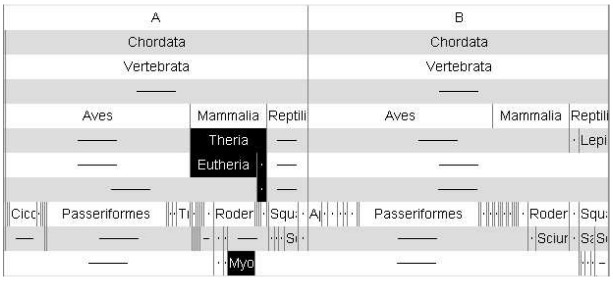


Figure 5. Some sub- and infraclasses are used only in tree A

As we can see, the main reason for different paths is that some subclasses and infraclasses are not used in tree B at all.

Using the derived attribute *Count(Phylum) per Latin Name*, we detected that the 17 animals of *Genus Apus*, even belong to two different *Phylums*, namely *Chordata* in A, but *Arthropoda* in B! Also, 3,429 birds are classified in different families.

A full-text search in some or all attributes is also possible. In Figure 6 the result of the search for "horse" in tree A is displayed. Matching values are highlighted at many different levels of the tree. An automatic zoom-operation has already been performed on all animals, where at least one cell is marked. This corresponds to a disjunction like

Common Name in {American horsemussel,...,velvet horse crab}
or Phylum = horsehair worms
or Class = horseshoe crabs
or Suborder = seahorses
or Family in {horse crabs, horsefish, horses, seahorses}
or Genus in {horses, redhorses, seahorses}
or Species in {northern horsemussel, shorthead redhorse}

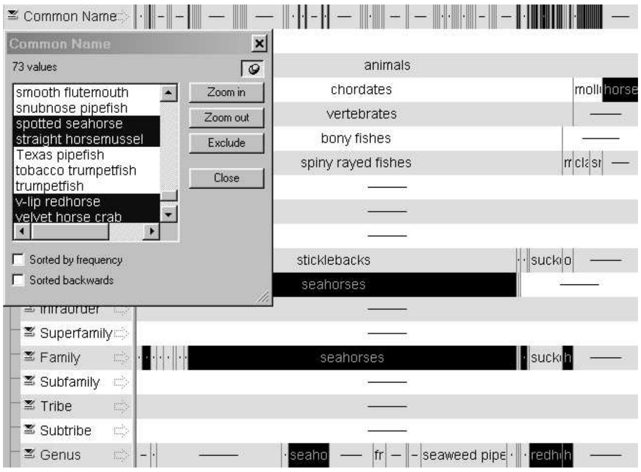
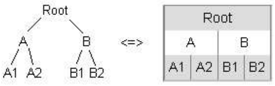


Figure 6. Result of a full-text search for "horse"

3 Conclusion

Even though InfoZoom is a general tool for database analysis, it turned out that it is well suited for the analysis of trees. There is a natural mapping from trees to the stacked table cells of InfoZoom:



The zoom mechanism allows to focus on any subset of the tree. In large trees, cells are often too small to read, but this complies with the zoom metaphor: small objects may be invisible from a distance.

A small weakness of InfoZoom is that it cannot directly import the XML files. We had to write a simple transformation program.

4 Related Work

The TableLens [Rao and Card 1994] is the only approach we know which also uses the basic idea of compressing database tables until they completely fit on the screen. While InfoZoom displays each record in a column, in TableLens each row contains a record. Therefore, the TableLens cannot use the technique of uniting adjacent cells with identical values, which is vital to make textual values readable.

References

RAO, R. AND CARD, S. K. 1994. The Table Lens: Merging Graphical and Symbolic Representations in an Interactive Focus+Context Visualization for Tabular Information. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems* (Boston, MA, Apr 24–28, 1994), pp. 318–322.

SPENKE, M.; BEILKEN, CHR. 2000. InfoZoom - Analysing Formula One racing results with an interactive data mining and visualization tool In: *Data mining II* / Ebecken, N.[Editor], S. 455 – 464.

SPENKE, M. 2001.Visualization and interactive analysis of blood parameters with InfoZoom In: *Artificial Intelligence in Medicine*, Bd. 22, Nr. 2, S.159 – 172.

http://www.humanIT.de – The InfoZoom home page. A free test version of InfoZoom can be obtained.

http://www.fit.fraunhofer.de/~cici/InfoVis2003/Index.htm – This web page contains demo videos and the analysis of the file system data of the InfoVis 2003 Contest.