Zoomology: Comparing Two Large Hierarchical Trees

Jin Young Hong, Jonathan D’Andries, Mark Richman, Maryann Westfall
College of Computing
Georgia Institute of Technology

ABSTRACT
Zoomology compares two classification datasets. In our solution the two trees are merged into a single overview, which unfolds top to bottom and left to right. Color represents rank, and the width of a classification node corresponds to the number of its descendants. Matched twin detail windows allow similarities and differences to be compared as the user navigates their hierarchies via a zoomable interface.

THE ZOOMOLOGY APPROACH
Our approach is a hybrid of several known techniques built within an overview and detail framework. The overview provides the “big picture” while the detail view explores substructures and nodes within the hierarchies.

Overview
Zoomology’s overview is intended to show structure, show navigation in context, and indicate regions of difference. This is accomplished in a single view by constructing the union of the two trees. Both trees are mapped to a space-filling, multitree representation of the structure, as proposed by Furnas and Zacks [2]. Because the data sets are so similar—we found them nearly 90% the same—drawing both hierarchies side-by-side unnecessarily repeats most of the structure. Instead, we draw the union of both trees with areas of change marked in white. As shown in figure 1, this makes changed regions immediately identifiable.

Detail
Zoomology’s detail view simulates a top-down navigation of the trees. Figure 2 depicts an immediate comparison of the two data sets as seen from the root node. The yellow circle represents kingdom Animalia, each deep red circle within represents a phylum, and the circles inside each phylum represent its children, color-coded by rank.

In the detail view, spatial encoding distinguishes levels. Root nodes enclose child nodes, which then enclose their own children. As the user zooms into the next level, the current level outgrows the screen, revealing the next. Zooming out shrinks the current level and shows the prior.

To represent twenty different ranks with easily distinguishable colors we assigned the same base color to all levels within a major rank and varied subrankings by differing their tints. As children rarely reside more than two ranks from their parent, few colors are needed in the detail view, avoiding visual clutter.

Navigation in Zoomology is like flying through a tunnel. The user picks a direction and zooms towards more detail in the selected region. As a default, zooming in one tree zooms the other to the corresponding location. However, the windows can be unlinked to explore a single dataset.
Differences between the hierarchies are highlighted by position and border color. To facilitate comparison, enough space is allocated in each detail window to render the union of all nodes visible in both trees. Nodes existing in one tree but absent in the other are drawn with a white border. An empty space in a particular position indicates a node that is absent in that tree but present in the other. In this way, a unique node is highlighted in the correct tree and conspicuously absent in the other.

**The Smart Legend**

Centered between the detail windows is the “smart” legend, which maps all ranks to their associated colors and also records path data. The path marked in white to the left of the legend bar shows the nodes traversed en route to the view in the left detail window (data set A) and the right path to the right window (data set B). Path information helps identify the current level in the tree, the nodes traversed, and the difference between paths in the datasets. In the overview and legend, different shapes are used to mark the navigation paths of the different databases.

**Interaction of Overview and Detail**

Interaction between the overview and detail views enhances Zoomology’s usefulness. As users click on regions of interest in the overview, the detail view smoothly pans and zooms to that area. Context and navigation history, missing from the detail views, are provided both in the legend and the overview. The legend shows the type of ranks traversed, and the overview shows specific location within the hierarchy. This combination of overview, detail, and legend helps overcome the limitations of each view by itself.

**Sample Tasks**

- Explore difference between the hierarchies:
  
  Zoomology promotes ad-hoc exploration of differences between datasets. A white area represents change in the overview. Clicking in the area zooms the detail windows to that location in each tree. White also represents change in the detail windows. A white circle around any node indicates that it either does not exist in the other dataset, or that it exists in another location. Each node encloses its own children and a white border around any of these “grandchild” nodes indicates difference in the same manner. Clicking on the name of any node will locate it in the other tree. The path to each is shown in the overview and the levels of all ancestor nodes in each dataset are marked in the smart legend.

- Find Spirulida in both trees and show its genealogy:
  
  The user selects “Latin Name” and enters “Spirulida” or selects it with the alphaslider. The active detail window zooms to the named node, its path is shown in the overview and the levels of its ancestor nodes are marked in the smart legend. If a white circle surrounds the node, clicking its name will locate it and mark its path in the other dataset.

- Locate all turtles:
  
  The user selects “common name” and enters “turtle.” A window appears showing all nodes with turtle as part of their common name, and the location of each is marked in black on the overview.

**RELATED WORK**

Our framework is similar to Pad++ [1]. Zoomology exploits zooming techniques employed in GVis, a tool for visualizing genome data [3]. The commercial product Grokker [4] uses a similar circular-container zooming methodology, but it is targeted for more general data sets.

**Limitations**

Future work on Zoomology could address some of its current limitations:

- **Intermediate Overview:** Allowing the user to enlarge parts of the tree structure would ease some of the problems we have seen at the overview’s lower levels. It would allow accurate mapping of size and color for lower-level nodes, and allow the user to select one of these for detail view. Areas of difference between trees could be colored to indicate rank and bordered to represent the tree of origin. An intermediate overview would allow the user identify all nodes that share a common name and could aid comparison of subtrees.

- **Qualitative Differences:** There is no way to discern between minor changes such as the insertion of a single node and major revisions such as changes throughout an entire branch. There is no indication of qualitative vs. quantitative change.

- **Multilevel details:** The detail view cannot compare substructures spanning multiple levels.

- **Other Contest Tasks:** Our solution would not apply to a dataset where change is marked by link length. However, it could easily serve as the basis of a system to examine file system differences.

**REFERENCES**


