**The Similarity Matrix Viewer**

Live version is available here: [http://bento.stanford.edu:3002/co/style](http://bento.stanford.edu:3002/co/style)

---

**The Data**

The data we used is a set of crowdsourced style and domain labels applied to Web pages that was collected for a research project. Participants were asked to apply terms describing the style of the page (i.e. minimal, fun, colorful) and the domain of the site (i.e. business, blog, art). In the resulting dataset we have 10,000 style labels and 5,000 domain labels associated with Web pages via a unique page id. For each page two style labels and one domain label were required. For a subset of 3,000 of the labels, participants labeled the same set of 100 pages, handpicked for diversity, to guarantee overlap among pages. In this assignment, all the results were combined and are being analyzed together. The raw data was spread across a few tables with similar schemas in a SQLite database. Columns available included user id, pageid, domain label, style label (Maxine’s A2).

---

**Motivation**

A few preliminary visualization techniques previously explored on this dataset (Maxine’s A2) resulted in useful, but difficult-to-navigate displays. We found seriated co-occurrence matrices interesting to explore but difficult to interpret. With so many labels, manually identifying the labels that correspond to cells in the matrix is often nontrivial. Moreover, the co-occurrence matrix is far-removed from the data it represents: there is no visual mapping between a style label and the design to which it was applied. These were the primary frustrations that motivated our interactive visualization.

---

**Implementation Techniques**

The two primary implementation techniques we focused on were magnification with brush and link, aimed to facilitate matching cells to design, and gallery exploration to enable mapping labels to designs.

To allow viewers to easily navigate a large co-occurrence matrix with small labels, we provided a magnified view of a small region with large corresponding labels for easy identification of each cell. While in the static version of this visualization, you would either need a small matrix with large labels or static magnified views of each region of the matrix, in this interactive version, users can clearly identify cells and regions that they hover over, as well as see the positioning of that region relative to the overall matrix.

In order to relate the style labels to the underlying designs, we created a gallery feature. Upon clicking any cell, a gallery is populated with pages to which the corresponding labels were applied so the viewer can get an idea for the types of designs people associated with these labels. Since the matrix is symmetric, cells along the diagonal will show any page with the given label.

The other major feature that we added in later stages was cluster selection. This technique would allow users to drag over a region to define a cluster of interest. A list of labels in this cluster would be populated and displayed, and multiple lists could be created.

---

**Storyboards**

This is the initial layout we thought of...

See submission for picture.

---

**Final Interactive Visualization**
Seri (short for seriation) allows viewers to interactively explore seriated co-occurrence matrices of crowdsourced style labels applied to Web pages. Here are some features that Seri provides:

- **Highlighted cell and labels upon hover**: hovering over cells highlights them and their corresponding labels facilitates identification.
- **Magnified zoom**
  - Viewfinder: the upper right corner shows a magnified view of the 5x5 region surrounding the hovered cell.
  - Labels: larger labels corresponding to the magnified region, with the center cell labels accented, allow for easy exploration of a particular region.
- **Gallery thumbnails**: clicking a cell populates a gallery with pages corresponding to the style labels.
  - Hover contributions: hovering over a thumbnail highlights the page’s contributions to the main matrix.
  - Expansion on click: clicking a thumbnail activates a larger view.
- **Cluster selection mode**: turning on the cluster selection button allows you to click and drag over an area of the matrix
  - Persistent cluster selection: selected clusters are tinted in color for subsequent identification
  - Lists: labels included in selected clusters are displayed in lists, with colored bars that map back to the selected cluster
  - Reset: the reset button allows for clearing of selected clusters and corresponding lists
- **Multiple seriation methods**: the dropdown menu underneath the matrix provides six different seriation methods by which to rearrange the matrix, and it is accompanied by a brief description of each method

**Diff Storyboard Final**
Mostly the elements from our storyboard all persisted but were rearranged for better viewing. For example the magnify area was initially at the bottom

**Development Process**
We approximate that the whole process took about 24 hours, ~2 and a half all-nighters. Most of the time was spent together pair-prototyping and programming, with maybe 2-3 hours each spent individually working on different modules. Styling and rearranging the elements of the application took a particularly long time. Implementation was nontrivial, but the main hurdles we came across related to styling rather than functionality, ie. svg elements can’t have borders! In terms of work split evenly among the group members, for all of the major features work was done either together or in modules by both members. Even the parts that were assigned separately all required some action and input from both parties, so it really was an integrated collaboration.

**Features**
- Highlighted cell and labels upon hover (both)
- Magnified zoom viewfinder (Maxine created magnified matrix)
- Magnified zoom labels (both)
- Multiple seriation methods (both implemented the seriation switch, Maxine provided multiple seriations, Cesar created the menu option)
- Gallery thumbnails with expansion on click (Cesar hooked up the gallery to the interface, Maxine added it to the interface)
- Cluster selection mode with list creation (Cesar did cluster selection, Maxine did list creation/removal)
• Lots of styling done by both

**Bundled source code with instructions for building/running**
Live version is available here: [http://bento.stanford.edu:3002/co/style](http://bento.stanford.edu:3002/co/style)

Made using Rails 3.2.1
To run, install rvm, install ruby, install rails
Run bundle install
Run rails server
localhost:3000:co/style

Check critiques of combined previous assignments

**Seriation techniques**

ARSA: Anti-Robinson sseriation by simulated annealing
[http://www.tandfonline.com/doi/abs/10.1080/07408170701411393#preview](http://www.tandfonline.com/doi/abs/10.1080/07408170701411393#preview)
Brusco et al, 2007
Uses a number of optimization criteria to approximate an anti-Robinson structure, i.e., matrix with strictly increasing order as you move away from the diagonal.

TSP: Traveling salesperson problem solver
Takes the connecting circle returned by a TSP and cuts it to create a linear ordering using a dummy point equidistant from each of the other points.

Chen: Rank-two ellipse seriation
Projects points of a rank-2 correlation matrix generated from the supplied distance matrix onto an ellipse, and cuts the ellipse at the top to produce a linear ordering.

MDS: Multidimensional scaling
Given a distance matrix, provides a configuration that best minimizes a function that represents the difference between the matrix value of the rows and their separation in the configuration.

HC: Hierarchical scaling
Uses ordering of a dendogram produced by hierarchical clustering, which merges elements into a hierarchical tree based on values from a distance matrix.

GW, OLO: Hierarchical clustering with optimal reordering
Takes a dendogram produced by hierarchical clustering and reorders clusters at each level such that their neighbors are the clusters with closest distance values.

------------------------------------------