Exploring Meta-Data Associations with Bungee View

Mark Derthick*

Human Computer Interaction Institute, Carnegie Mellon University

ABSTRACT

Bungee View is designed to support non-technical users in familiar document searching and browsing tasks while introducing the unfamiliar task of discovering patterns in the documents' meta-data. Two expert-oriented features were added for the InfoVis contest to support hypothesis testing and discovery of patterns involving multiple attribute values. Most of the contest questions were answered in the negative, or with only weak associations. Stronger associations of potential interest were observed in the course of analysis. The accompanying video documents the analysis, while this document explains the design decisions behind Bungee View.

Keywords: information visualization, exploratory data analysis.
Index Terms: H.5.2 [Information Interfaces and Presentation]: User Interfaces---Graphical user interfaces, Interaction styles, Screen design; H.3.3 [Information Storage and Retrieval]: Information Search and Retrieval---Information filtering, Query formulation; H.2.8 [Database Management]: Database Applications---Data mining, Image databases;

1 BUNGEE VIEW DESIGN

The name "Bungee View" evokes Shneiderman's Information Seeking Mantra [4]: Overview first, zoom and filter, then show details on demand. It also emphasizes that viewing a dataset from a high level and diving into ever smaller subsets is a cycle with many bounces. Bungee View (BV) is an open-source (GPL) Java Web Start application available at BungeeView.com, built with the visualization toolkit Piccolo [1]. While it includes a result list and a summary of the selected document, the InfoVis contest asks about meta-data patterns, and so these features are little used.

The focus is BV's display of associations among meta-data attribute values (called *features*). Associations are deviations from independence, and can be positive or negative. Users bounce through the data by adding and deleting filters on features, which determine an evolving set of selected movies.

The analysis in the video found that R-rated movies are much less likely to have a high box office value than PG or PG-13 movies. To explore this negative association further, in Figure 1 the universe of movies in which to show associations has been restricted to those that grossed at least $70 million (with the Restrict button). Then the rare R-rated movies within that universe were selected. The R bar and label are colored bright green to show that this filter is in effect. (Bright red is used for negative filters.)

Each row of bars is a Mosaic Display [2] for one attribute. In the figure, the Genre attribute is selected, so its Mosaic Display is magnified and labeled. For other attributes, feature names and counts are shown on rollover at the bottom right of the screen.

Bar heights encode the percentage of movies with each feature that are selected, and vary from zero for Animation and Family to 78% for War. That is, 78% of high-grossing War movies were rated R. By definition, a feature is independent of the filters if its selection percentage is the same as that for the whole database. The height of the gray background encodes this expected

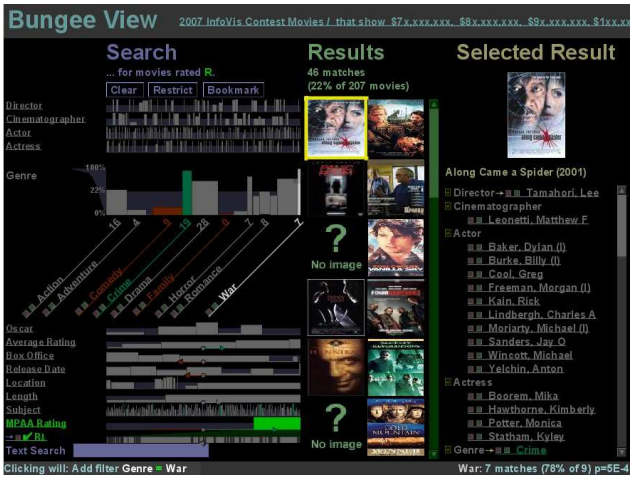

Figure 1. Associations between an R rating and each genre for high-grossing movies.

percentage (22%). To emphasize variation around this value, heights are scaled non-linearly to place it in the middle. Bars higher than this represent positive associations, and vice versa.

Bar widths show the total number of movies having a feature. Thus [distorted] area is proportional to the number of selected movies with the feature. Width controls for the unfiltered distribution, leaving height to show associations.

Experimental subjects using BV sometimes included overly specific features in their hypotheses. For instance, a user might find that Best Pictures tend to have a high box office, when *all* Oscar categories have this property. To reduce this problem, BV now considers only movies having the parent feature in the independence test. The background rectangle now represents the average percentage *for each attribute*. This makes the significance test hierarchical; it tests whether a feature is unusual among its siblings, rather than in the whole database. This doesn't make much difference for an attribute that almost all movies have, like Genre, but it can be important for Oscar or for a nested feature like Location → Canada → Alberta.

BV also minimizes unwarranted conclusions by coloring bars dull green (orange) to show whether a bar's height is significantly higher (lower) than the background, according to a Chi-squared test. Bars that are colored (statistical significance) *and* noticeably higher or lower than the background (practical significance) should be given the most attention.

BV uses Bonferroni correction of the χ^2 significance threshold with the goal of reducing the chances of finding any false positive associations during an analysis to 5%. It assumes a user will check all bars on the screen for color, and that she will view 5 different sets of bars over the course of analysis. For the example in the figure, the resulting threshold is 5% / 1500 bars / 5 sets = $7 \cdot 10^{-6}$. BV settles for a conservative correction procedure in the belief that making any association salient is a bonus compared to other search and browse tools, and that failing to point out less strong but potentially interesting associations is a relatively minor failure. With moderately selective filters, p-values are often very

small (10^{-50} or less), and corrections on the order of 10^4 have negligible effect.

2 Expert Features Added for Contest

For this contest there are specific a priori hypotheses to be tested, so being more precise about significance is important. Therefore the unadjusted p-value for each Chi-squared test is now added to the usual rollover feature summary, allowing an expert to do his own adjustment in his head if desired. In Figure 1, War has the highest bar, but is not colored. Rolling over it shows a p-value of $5 \cdot 10^{-4}$, which is greater than the $7 \cdot 10^{-6}$ computed above. If we hypothesized that Genre is associated with an R rating before looking at any data, the significance threshold could be based on only the 19 Genre bars, giving $0.05 / 19 = 3 \cdot 10^{-3}$ and making the War association significant.

BV only shows associations between the evolving set of filters and individual features. However the contest asks for *teams* of people associated with high-grossing movies. To address this, a command was added to actively search for sets of conjunctive features associated with the current filters. The top 100 of these clusters of features are listed in order of p-value, and can be added as filters just like single features. This sort of data-mining (market-basket analysis) often finds obvious or otherwise useless clusters. BV addresses this problem by only looking for positive associations, and by allowing users to specify features to ignore. It is relatively quick to click on features in uninteresting clusters to remove them, as shown in the video.

3 Related Work

Flamenco [6] supports the same kind of filters on hierarchical attributes as BV, and this capability is now commonly seen on shopping and other web sites. However the reason for filtering is primarily searching and browsing. The number of filtered objects having each feature is shown, but not the baseline frequency needed to judge association.

The Relation Browser (RB) [3] shows both filtered and unfiltered counts, but uses conventional equal-width histograms with a linear scale. It is difficult to compare percentages between bars of different lengths in order to infer associations, and impossible when queries are even moderately restrictive and the filtered bars are less than one pixel high. Both Flamenco and RB show all feature labels all the time, which may require scrolling in Flamenco, or unreadably small fonts in RB.

In the commercial product InfoZoom [5] the simplest mode, Overview, is similar to BV. The most common (in their papers and demos) behavior when filters are applied is to zoom in on the selected objects in order to maximize space for their bars and labels. This loses the baseline distribution, and hence the ability to see associations. BV has less need to expand the selected objects because space is dynamically allocated to a selected attribute, labels don't have to fit inside bars, and a greedy algorithm draws labels with the largest filtered count first, which are expected to be the most relevant. The InfoZoom behavior is available in BV with the Restrict button if you want to ignore associations with the current filters and show those in a restricted universe. InfoZoom supports derived attributes, e.g. percentage of movies with a high box office, which can support visualization of associations as height differences. This is a tradeoff of power versus simplicity, and derived attributes have almost unlimited power.

None of these applications show significance, like BV's bar colors, nor do they integrate automated search with user-guided exploration as does the Cluster command.

Mosaic Displays (MDs) [2] do show significance with color. They can show associations among multiple hierarchical attributes, but quickly become difficult to interpret. BV uses multiple MDs each with one n-ary variable (an attribute) and one binary variable (selection). For binary variables the height and color of the positive value uniquely determine those of the negative value, so nothing is drawn to represent the non-selected movies. This makes BV's MDs look more like histograms. This special case is easy to interpret, but cannot show general multivariate associations (though BV does show associations between each feature and the Boolean combination of applied filters). BV also scales heights non-linearly, and adds Bonferroni correction. Attributes are primarily treated as nominal, though multiple selection supports range restrictions as well.

4 Future Work

There is ongoing tension between informing and overwhelming users. In Figure 1, the rollover summary is in the lower right corner and very terse. In order to make a previous version of the video understandable, arrows had to be superimposed to show where to look. The final video uses more verbose popup summaries, which are indeed salient, but can also be annoying.

5 Summary

Bungee View was designed to be simple. Users can't choose what variables to display or how to display them. They can't sort, group, or nest. Yet with the addition of p-values and clustering, an expert user was able to answer all contest questions, while finding additional interesting associations along the way.

The simplified Mosaic Displays encode unconditional attribute distributions with width, conditional distributions with area, and deviations from independence with height. They support alphabetic search by their left to right order, and cardinality search based on width or area. Especially tall or short bars, as well as colored bars, can be picked out quickly even for the unselected attributes. Thus there is room for the 13 movie attributes and thousands of their values on the half of the XGA window devoted to meta-data.

The video answered the contest questions as follows:

- Best Actress is weakly associated with Drama.
- No Oscar category is significantly more or less likely than another to be a box office winner, though there is a strong association with Oscar winners in general.
- The six most bankable people are actresses Cameron Diaz and Halle Berry, cinematographers Oliver Wood and Don Burgess, and actors Clyde Tull and Tom Cruise.
- No multi-person teams are significantly associated with box-office winners due to the large Bonferroni correction.

References

[1] B. B. Bederson, J. Grosjean, and J. Meyer, *Toolkit Design for Interactive Structured Graphics*. IEEE Transactions on Software Engineering, 2004. **30**(8): p. 535-546. http://www.cs.umd.edu/hcil/piccolo/learn/Toolkit_Design_2004.pdf

[2] Michael Friendly, *Visualizing Categorical Data*. 2000: BBU Press.

[3] Gary Marchionini, Carol Hert, Liz Liddy, and Ben Shneiderman. *Extending User Understanding of Federal Statistics in Tables*. in *Conference on Universal Usability*. 2000. Washington, DC: ACM. http://www.ils.unc.edu/~march/CUU/tables.pdf

[4] B. Shneiderman. *The eyes have it: A task by data type taxonomy for information visualizations*. in *Proceedings of the IEEE Symposium on Visual Languages*. 1996: IEEE Computer Society Press.

[5] Michael Spenke and Christian Beilken. *InfoZoom - Analysing Formula One racing results with an interactive data mining and visualisation tool*. in *Second International Conference on Data Mining*. 2000. Cambridge University, United Kingdom. citeseer.ist.psu.edu/spenke00infozoom.html

[6] Ping Yee, Kirsten Swearingen, Kevin Li, and Marti Hearst. *Faceted Metadata for Image Search and Browsing*. in *Proceedings of the SIGCHI conference on Human factors in computing systems*. 2003. Ft. Lauderdale, Florida, USA: ACM Press. http://bailando.sims.berkeley.edu/papers/flamenco-chi03.pdf