biovis use-case, visualization ideas

January 9, 2013

1 Use-Cases

When thinking about the functionality of the visualization tool, we decided that it makes the most sense to think in terms of the capabilities required for three primary use cases. These three use cases are the EMR annotator, the NLP scheme reviser, and the EMR analyst, as described below:

- use-case: annotator
  - goals: annotate EMRs with values for relevant variables/indicators.
  - required capabilities: Assist annotator by suggesting values for relevant variables/indicators, and/or directing the annotator's attention to regions of a document which are likely to be relevant. Support document selection based on criteria such as document ID, certainty of NLP predictions, metadata, etc.
  - possible visualizations: when a variable/indicator is selected, predict a value (along with certainty scores) and highlight terms in the EMR which are most strongly associated with the predicted class. When annotating a document, allow the annotator to highlight the region(s) of the document which factored most strongly in their decision in order to provide additional feedback to improve the back-end model.
- use-case: reviser
  - goals:
    - Revise the annotation scheme. This could include addition of new attributes, removal / merging of redundant attributes, updating rules for indicator attributes, etc.

        * Assess the quality of the NLP back-end in terms of model performance, ie, which models perform well / badly? Are the appropriate features being detected? (this second goal may lead to overlap / interaction with the annotator, as the reviser may choose attribute models for improvement via additional annotation)
    - required capabilities: viewing overall frequency counts for variable/indicator categories, viewing correlations between different variables/indicators in order to streamline the annotation scheme by minimizing redundancy, viewing summary information concerning the overall accuracy of NLP predictions, viewing summary information concerning the NLP model's decision-making (ie, top feature weights, etc.)
    - possible visualizations: heatmap representing level of correlation between different variables/indicators and/or specific categories of said variables/indicators; force-directed map representing rule-based and correlation-based relationships between attributes; interactive grid-based or treemap-based visualization where different attributes are projected along the x and y axes on-demand, and documents are grouped accordingly; wordcloud-based visualization to convey the most-important features per attribute.

- use-case: analyst
    - goals: assess overall quality of EMRs or proceedures performed by particular doctors, hospitals, floors of hostpials, etc. Select patients or records matching particular criteria for cohort inclusion in future studies. Optionally, provide feedback conerning how well / poorly NLP predictions matched their expert judgements.
    - required capabilities: aggregate documents according to metadata such as doctor, hospital, etc. Summarize groups of documents and compare different groups along multiple quality dimensions. Provide an informative description of the meaning of each variable/indicator.

2 Visualization Components

Given the use-cases and the capabilities required by each, below are a collection of visualization techniques we've discussed which we judge to have

potential uses for supporting some of the required capabilities. I've included my thoughts (still rather rough) concerning each technique, how it could be applied, and what roles it could fulfill in this system.

- Heatmaps (summary-level)
  - role: Summarize correlations between large numbers of attributes and/or specific values of attributes across all documents.
  - enhancement: By using glyphs within each cell of the heatmap, we can also use the heatmap to illustrate overall distributions of attribute values.
  - weakness: space-inefficient (if each attribute is listed on each axis, half the table will be redundant); this could be mitigated if we provide a mechanism by which the user can manually select which attributes appear on which axes
  - weakness: User may be overwhelmed by amount of data; does not natually orient around uniquely-strong correlations.
  - filtering / zooming method: user manually selecting attributes to incorporate on each axis? Example: initially a large number of attributes on both axes. User deselects attributes with no strong / interesting correlations. After pruning down to a few attributes per axis, user toggles heatmap mode from attribute-only to attribute-broken-down-by-value mode to see greater detail, specifically, correlations between individual attribute values
  - supported selection: attribute and document: clicking on cells selects attributes corresponding to the given row & column, and also selects documents contributing to that cell (ie, documents which share the same value for the pair of attributes). Multi-cell selection should be supported.
  - imagined use-case: (designer) uses heatmap to explore correlations between attributes in annotation scheme
  - imagined use-case: (analyst) selects cells representing interesting attribute / attribute-value combos in order to select documents satisfying these criteria for more detailed review.
  - additional ideas: fisheye-style lens for handling large number of attributes, or dynamically incorporating value-level visualizations on mouse-over of a cell.
- Heatmap (group-level)

  - role: Summarize correlations between small numbers of specific values of attributes across a subset of documents.
  - thought: perhaps we should abandon this idea; most of the same functionality can be incorporated into the summary heatmap; ie, heatmap can be manually "zoomed" based on attribute (de)selection and selected display mode (attribute-level [coarser] versus value-of-attribute-level (finer)). Also, it seems unlikely that the user would need both the group-level and summary-level heatmaps simultaneously.
  - The only unique contribution that this group-level heatmap would provide is the ability to visualize attribute correlations using only a subset of the documents. The summary-level heatmap is incapable of this because heatmap cells themselves provide document selection functionality. (However, if this functionality is particularly important, we could disable document selection via heatmap cell in the summary heatmap?)

- SVD-based scatterplots
  - role: lay out documents based on a range of user-selected criteria; displays the full dataset
  - weakness: axes have no innate meaning; SVD often weights original dimensions unpredictably
  - filtering / zooming method: literal zooming in the space; selection of particular attributes for layout
  - supported selection: document: individual document and group selection
  - imagined use-case: analyst looking for documents matching a fuzzy set of attribute-value criteria would select the relevant attributes, plot the points, and then select groups of documents with roughly similar values in the new space.
  - enhancement idea: rather than showing each individual point in a scatterplot, [2] clusters the points in the scatterplot, displaying the clusters in lieu of the individual points. Each cluster is accompanied by a table containing summary information describing the properties of the cluster. From the perspective of the analyst, perhaps something like this would make sense for our dataset? (Or, perhaps this should this be addressed as an

entirely separate panel / visualization technique from the SVD plot?)

  - enhancement idea: the "gravi++" system discussed in [3] adds special auxiliary nodes to the plot which represent particular properties of interest. Perhaps it would improve the navigability of the SVD plot to add special nodes representing particular attribute values / combinations of values which may be frequent / at particular locations / otherwise important?

- Wordclouds
  - role: abstract, high-level summary of semantics present within a group of documents
  - enhancement: encode additional information via highlighting of terms (as in [1]); for example, perhaps hue and saturation could be used to encode correlations of terms with particular attributes, where hue indicates the most-strongly related attribute, and saturation indicates strength of the correlation?
  - imagined use-case: provides additional coarse-grained information to analyst during the document group selection process

- Force-Directed Plots
  - We briefly discussed using force-directed plots early in the summer, but haven't touched on them recently. One of the problems we kept running into was that I couldn't manage to get them to effectively scale to the size of the full dataset. I still don't think that we can use them for dataset-level summary visualization, but I think they could have a role to play in individual-document/attribute-oriented visualization, ie, given a document or attribute, visualizing the strongest relationships that the given object has to others of its kind.
  - description: nodes as docs / attributes, edges as correlations / similarities; for docs-as-nodes, relatedness can be defined via attribute selections or textual similarity (as was the case for the SVD layout); for attributes-as-nodes, relatedness can be defined in terms of correlations.
  - role: individual-document-centric (or individual-attribute-centric) method for displaying summary information, optionally (probably) for a limited subset of the full dataset / annotation scheme

  - weakness: probably can't visualize entire dataset at once (too much clutter, force-directedness may start to spaz out)
  - imagined use-case: scheme reviser: (attributes-as-nodes) selects an attribute of interest, increases/decreases correlation threshold to dynamically add/remove elements to/from visualization depending whether their correlation with the attribute of interest exceeds current threshold value.
  - imagined use-case: analyst: (documents-as-nodes) given a document that the analyst wants to include (ie, meets certain criteria for study inclusion), sets attributes of interest and increases threshold until a sufficient number of similar documents of interest have been added to the viz.

- Grid-Based Document Plot
  - Projecting different attributes onto the x and y axes, documents (represented as glyphs) are dynamically rearranged to fall within appropriate cells.
  - Can be generalized to a TreeMap (for more than 2 attributes?).
  - role: document grouping according to attribute values
  - imagined use-case: analyst wishes to see relatively how many documents meet a particular set of attribute-value criteria. Analyst enables these attributes causing documents to re-arrange themselves with respect to their values for these attributes. Analyst can then see the relative populations of each of the groups, and can drill-down into specific document groups on-demand.
  - imagined use-case: reviser wishes to view the skew of particular attributes. Reviser enables selected attributes in order to assess relative sizes of each group, and to assess degree to which attributes are correlated.

References

[1] Weiwei Cui, Yingcai Wu, Shixia Liu, Furu Wei, M.X. Zhou, and Huamin Qu. Context preserving dynamic word cloud visualization. *IEEE Trans. on Computer Graphics and Applications*, 2011 2011 2011.

[2] D. Oelke, Ming Hao, C. Rohrdantz, D.A. Keim, U. Dayal, L.-E. Haug, and H. Janetzko. Visual opinion analysis of customer feedback data. In *VAST*, 2009.

[3] A. Rind, T. D Wang, W. Aigner, S. Miksh, K. Wongsuphasawat, Catherine Plaisant, and Ben Shneiderman. Interactive information visualization for exploring and querying electronic health records: A systematic review. 2010.