IVTK : An Information Visualization Toolkit

Brad Lowekamp *Nathan Smith* *Grant Wagner* *Eleanor Chlan*
Penny Rheingans
University of Maryland Baltimore County

Abstract

Visualization of the documents which make up a document corpus can lead to discoveries about the nature of the corpus and its component documents. As part of the CADIP project, we have developed the Information Visualization Toolkit (ivtk) to facilitate the visual exploration of a document corpus. Individual documents are mapped to glyphs, using their meta-data to determine 3D position, color, size, and opacity. A menu-based interface specifies the nature of these mappings, allowing for flexible exploration of the relationships among documents. The display can be viewed interactively through rotation and zoom operations.

1 Basic Approach

Ivtk supports the mapping of documents to glyphs in a flexible and powerful way. Associated with each document are an arbitrary number of numeric meta-data values. These may include document size, date, relevance measures, and matches against various keywords. Meta-data items may be integral to the document or be the result of processing and judgements by the Information Retrieval system. Each meta-data value may be mapped to a different visual attribute. For example, for an exploration of news stories about political candidate views, X, Y, and Z position might show frequency of the terms 'gun control', 'environment', and 'education', while color indicated the date the story appeared. Such a display might facilitate explorations of the way in which the issues addressed by candidates change over time.

Figure 1 shows data where each instance is a person from the 1990 Census. The position of each glyph is determined by the values for years of education (green axis), age (red axis), and hours worked per week (tan axis). Each glyph shows a group of people with the same values for those variables. Each glyph is colored by the average gender for that group of people.

Mappings are specified through menu options in the ivtk interface, allowing the analyst to freely change the correspondence between variables and visual attributes. Normalization of variable range can be used to make data sets where variables gave widely different ranges more comprehensible. Alternatively, data variables can be scaled to arbitrary comparable ranges. Figure 2 shows data where the variables have been normalized to lie in the same range.

Ivtk is implemented using vtk, a freely available graphics, imaging, and visualization toolkit implemented in C++ with bindings to Tcl/tk, Java, and Python. Vtk runs on a wide variety of platforms, including Linux, MacOS, and Windows. Since vtk already includes the core visualization functions we need, the major task has been to design and construct an intuitive menu-based interface to the visualizations.

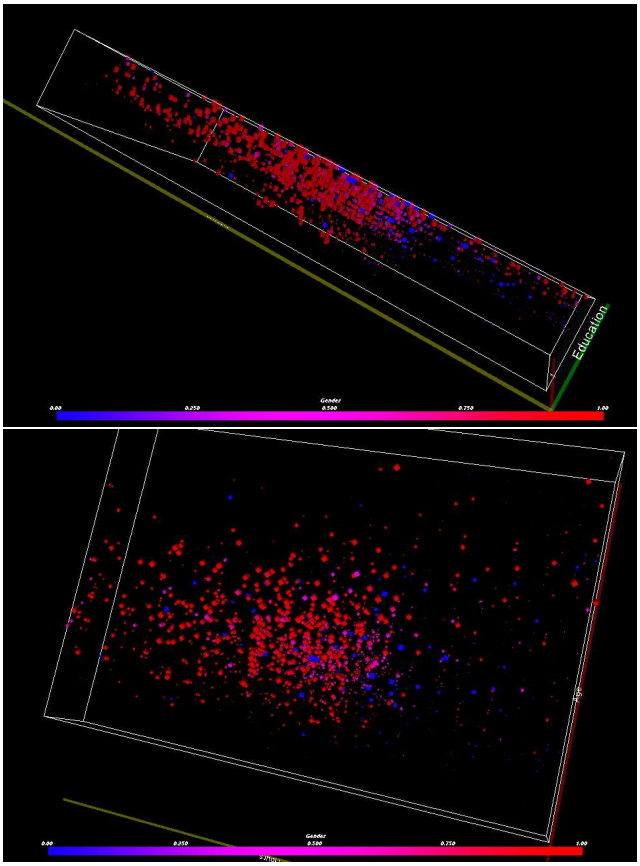

Figure 1 Visualization of income model data; position determined by education (green axis), age (red axis), and hours worked (tan axis); color determined by gender.

2 Visual Emphasis Implications

Although ivtk allows arbitrary mappings from meta-data variables to visual attributes, not all possible mappings are equally effective for every information discovery task. The choice of which visual attribute a variable is mapped to can have a substantial effect on how easy it is to pick up document similarities on that variable. Specifically, visual groupings on position are more apparent than those on color, size, or opacity.

Similarly, the choice of mapping has a strong influence on the attention drawn to individual glyphs based on their variable values. The glyphs likely to draw attention are those that are large, opaque, and brightly colored. Ideally, the data items which are of greatest interest will be mapped into glyphs which have these striking visual characteristics.

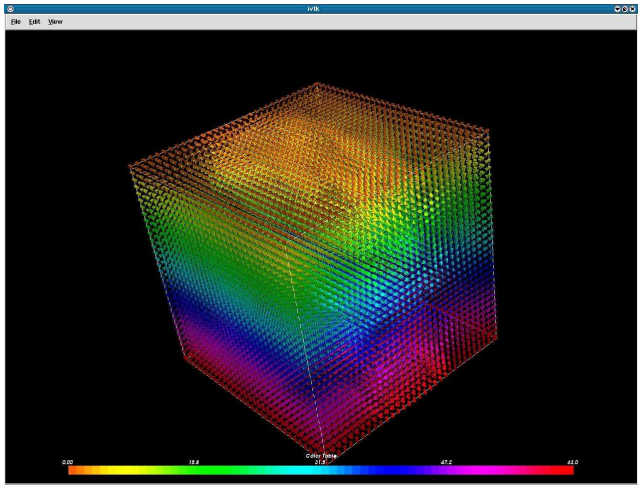

Figure 2 Visualization of range normalized data.

3 Issues of Scale

Ivtk has been used to display hundreds of thousands of glyphs (as in Figure 2), but one-to-one document-to-glyph mappings become unworkable when the number of documents becomes truly large. Accordingly, we have begun investigating ways to gather documents into meaningful clusters and accurately represent the clusters visually. Ivtk provides the option of using k-means clustering to identify document clusters that are then characterized by the cluster mean and variability in each meta-data attribute. The user selects the number of clusters desired and the data is grouped into that number of clusters. These clusters can be represented as glyphs displaying their mean characteristics, as well as new metadata such as number of documents in that cluster.

4 Intrusion Detection Application

One ongoing investigation explores the use of information retrieval and information visualization techniques for the examination of computer and network log files. Using ivtk, we have recreated a visualization (originally constructed using SFA) of telnet sessions for the purpose of detecting attempted intrusions. Telnet packets are grouped together into sessions to create documents. These documents are given relevance scores against queries consisting of three different known attacts. In ivtk, each session is represented as a glyph, with 3D position determined by the relevance of that session against each template attack. Figure 3 shows a group of 60 sessions from the Lincoln Labs IDEVAL data set. The axes correspond to similarites to three template attacks, labeled 'A, 'B', and 'C'. Most sessions cluster near the origin, not matching any attack. A few sessions appear out along a single axis, matching one particular attack. Perhaps the most interesting sessions appear in the middle of the space, matching elements of multiple attacks. These attacks would be missed using standard template matching intrusion detection methods.

5 Future Work

Our next primary emphasis will be on the visualization aspects, specifically how to convey both the central tendency (mean) and variability of multiple attributes for each cluster. Representing central tendency for an attribute is generally straight-forward. Each cluster can become essentially a large document, with the mean of the attributes represented as the same attribute values would be for a single document. For example, the glyph representing the cluster might be positioned according to three different mean relevance scores, colored according to mean date, and sized according to mean document size.

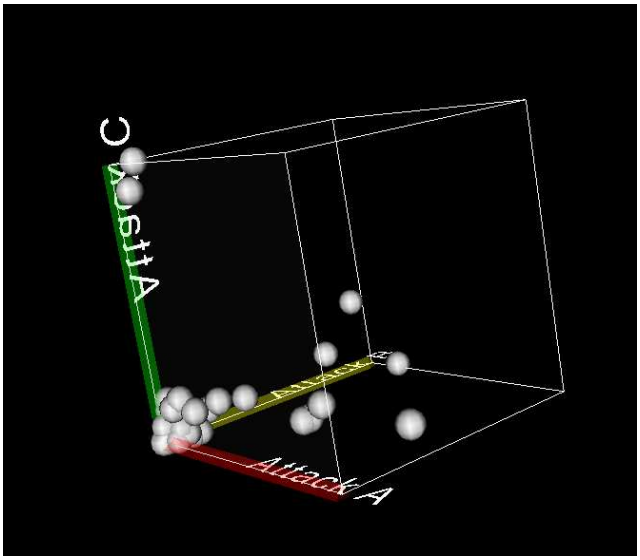

Figure 3 Visualization of telnet sessions matched against attack templates

Showing the variability present in a cluster is a more challenging visualizationtask. We've been designing visualization techniques in the spirit of the box-and-whisker plot for traditional graphs. In a box-and-whisker plot of a single variable mapped to the vertical axis, the data point for a cluster becomes a point or line at the height corresponding to the mean variable value. A box is drawn around this point, with the box top and bottom at the heights corresponding to the first and third quartile (or 5th and 95th percentile, or whatever) values of the cluster. A vertical line (the whisker) is drawn from the top and bottom of the box, extending the the height of the minimum and maximum value present in the cluster. This basic framework might be extended to 3D by showing the cluster mean as a glyph surrounded by a semi-transparent shell indicating the variability of the cluster in the three attributes mapped to position. The mean of the attribute mapped to color might be shown in the color of the glyph used to show mean of the cluster in the attributes mapped to position. The variability might be shown by coloringthe semi-transparent shell with the color resulting from the color spectrum given by the range of values of that attribute. In such a scheme, homogeneous clusters would have vivid colors in their shells, while heterogeneous clusters (in terms of the color attribute) would have less saturated colors in their shells.