Technical Research Report

A Knowledge Integration Framework for Information Visualization (2004)

by Jinwook Seo, Ben Shneiderman

TR 2005-63

A Knowledge Integration Framework for Information Visualization

Jinwook Seo[1,2] and Ben Shneiderman[1,2,3]

[1]Department of Computer Science, [2]Human-Computer Interaction Laboratory, Institute for Advanced Computer Studies, and [3]Institute for Systems Research
University of Maryland, College Park, MD 20742
{seo, ben}@cs.umd.edu

Abstract. Users can better understand complex data sets by combining insights from multiple coordinated visual displays that include relevant domain knowledge. When dealing with multidimensional data and clustering results, the most familiar displays and comprehensible are 1- and 2-dimensional projections (histograms, and scatterplots). Other easily understood displays of domain knowledge are tabular and hierarchical information for the same or related data sets. The novel parallel coordinates view [6] powered by a direct-manipulation search, offers strong advantages, but requires some training for most users. We provide a review of related work in the area of information visualization, and introduce new tools and interaction examples on how to incorporate users' domain knowledge for understanding clustering results. Our examples present hierarchical clustering of gene expression data, coordinated with a parallel coordinates view and with the gene annotation and gene ontology.

1 Introduction

Modern information-abundant environments provide access to remarkable collections of richly structured databases, digital libraries, and information spaces. Text searching to locate specific pages and starting points for exploration is enormously successful, but this is only the first generation of knowledge discovery tools. Future interfaces that balance data mining algorithms with potent information visualizations will enable users to find meaningful clusters of relevant documents, relevant relationships among dimensions, unusual outliers, and surprising gaps [10].

Existing tools for cluster analysis are already used for multidimensional data in many research areas including financial, economical, sociological, and biological analyses. Finding natural subclasses in a document set not only reveals interesting patterns but also serves as a basis for further analyses. One of the troubles with cluster analysis is that evaluating how interesting a clustering result is to researchers is subjective, application-dependent, and even difficult to measure. This problem generally gets worse as dimensionality and the number of items grows. The remedy is to enable researchers to apply domain knowledge to facilitate insight about the significance of the clustering result. Strategies that enable exploration of clusters will also support sense-making about outliers, gaps, and correlations.

A cluster is a group of data items that are similar to others within the same group and are different from items in other groups. Clustering enables researchers to see overall distribution patterns, and identify interesting unusual patterns, and spot potential outliers. Moreover, clusters can serve as effective inputs to other analysis method such as classification.

Researchers in various areas are still developing their own clustering algorithms even though there are already a large number of general-purpose clustering algorithms in existence. One reason is that it is difficult to understand a clustering algorithm well enough to apply it to their new data set. A more important reason is that it is difficult for researchers to validate or understand the clustering results in relation to their knowledge of the data set. Even the same clustering algorithm might generate a completely different clustering result when the distance/similarity measure changes. A clustering result could make sense to some researchers, but not to others because validity of a clustering result heavily depends on users' interest and is application-dependent. Therefore, researchers' domain knowledge plays a key role in understanding/evaluating the clustering result.

A large number of clustering algorithms have been developed, but only a small number of cluster visualization tools are available to facilitate researchers' understanding of the clustering results. Current visual cluster analysis tools can be improved by allowing researchers to incorporate their domain knowledge into visual displays that are well coordinated with the clustering result view.

This paper describes additions to our interactive visual cluster analysis tool, the Hierarchical Clustering Explorer (HCE) [9]. These additions include 1-D histograms and 2-D scatterplots that are accessed through coordinated views. These views are familiar projections that are more comprehensible than higher dimensional presentations. HCE also implements presentations of external domain knowledge. While HCE users appreciate our flexible histogram and scatterplot views, his paper concentrates on novel presentations for high-dimensional data and for domain knowledge:

- a parallel coordinates view enables researchers to search for profiles similar to a candidate pattern, which is specified by direct-manipulation
- a tabular or hierarchical view enables researchers to explore relationships that may be found in information that is external to the data set.

Visualization techniques can be used to support semi-automatic information extraction and semantic annotation for domain experts. For example, visual analysis by techniques such as dynamic queries has been successfully used in supporting researchers who are interested in analyses of multidimensional data [5][7]. Well-designed visual coordination with researchers' domain knowledge facilitates users' understanding of the analysis result.

This paper briefly explains the interactive exploration of clustering results using our current version, HCE 3.0. Section 3 describes the knowledge integration framework, including the design considerations for direct-manipulation search and dynamic queries. Section 4 presents a tabular view showing gene annotation and the gene ontology browser and section 5 covers some implementation issues.